frontiers | Frontiers in Big Data

TYPE Methods
PUBLISHED 18 June 2024
DOI 10.3389/fdata.2024.1359317

OPEN ACCESS

EDITED BY
Georgios Leontidis,
University of Aberdeen, United Kingdom

REVIEWED BY
Mamatha Thota,
University of Lincoln, United Kingdom
Xue Li,
The University of Queensland, Australia

*CORRESPONDENCE
Mohammad Rostami
✉ rostamim@usc.edu

RECEIVED 21 December 2023
ACCEPTED 22 May 2024
PUBLISHED 18 June 2024

CITATION
Stan S and Rostami M (2024) Source-free domain adaptation for semantic image segmentation using internal representations. *Front. Big Data* 7:1359317. doi: 10.3389/fdata.2024.1359317

Source-free domain adaptation for semantic image segmentation using internal representations

Serban Stan and Mohammad Rostami*

Department of Computer Science, University of Southern California, Los Angeles, CA, United States

Semantic segmentation models trained on annotated data fail to generalize well when the input data distribution changes over extended time period, leading to requiring re-training to maintain performance. Classic unsupervised domain adaptation (UDA) attempts to address a similar problem when there is target domain with no annotated data points through transferring knowledge from a source domain with annotated data. We develop an online UDA algorithm for semantic segmentation of images that improves model generalization on unannotated domains in scenarios where source data access is restricted during adaptation. We perform model adaptation by minimizing the distributional distance between the source latent features and the target features in a shared embedding space. Our solution promotes a shared domain-agnostic latent feature space between the two domains, which allows for classifier generalization on the target dataset. To alleviate the need of access to source samples during adaptation, we approximate the source latent feature distribution via an appropriate surrogate distribution, in this case a Gaussian mixture model (GMM).

KEYWORDS

domain adaptation, Gaussian mixture model (GMM), optimal transport and Wasserstein distances, sliced Wasserstein distance, image segmentation

1 Introduction

Recent progress in deep learning has led to developing semantic segmentation algorithms that are being adopted in many real-world tasks. Autonomous driving (Zhang et al., 2016; Feng et al., 2020), object tracking (Kalake et al., 2021), or aerial scene parsing (Sun et al., 2021) are just a few examples of these applications. Deep neural networks (DNNs) have proven indispensable for reaching above human performance in semantic segmentation tasks, given the ability of large networks to approximate complex decision functions (He et al., 2015). Training such networks, however, requires access to large continuously annotated datasets. Given that in semantic segmentation each image pixel requires a label, generating new labeled data for semantic segmentation tasks requires significant more overhead compared to regular classification problems.

Domain adaptation (DA) is a sub-field of AI which aims to allow model generalization for input distributions different from those observed in the training dataset (Wang and Deng, 2018). Unsupervised domain adaptation (UDA) addresses this problem for instances where the deployment dataset lacks label information (Wilson and Cook, 2020). This set of approaches is of especial interest for semantic segmentation tasks, where data annotation is expensive and time-consuming. In UDA, a model trained on a fully annotated

source domain needs to generalize on an unannotated *target domain* with a different data distribution. A primary approach for achieving domain generalization is learning a shared embedding feature space between the source and target in which the domains become similar. If domain-agnostic features are learnt, a semantic classifier trained on source data will maintain predictive power on target data. While this high level approach is shared among various UDA frameworks, different methods have been proposed for achieving this goal.

A common line of research achieves domain alignment by the use of adversarial learning (Goodfellow et al., 2020). A feature extractor produces latent feature embeddings for the source and target domains, while a domain discriminator is tasked with differentiating the origin domain of the features. These two networks are trained adversarially, process which leads the feature extractor to learn a domain invariant feature representation upon training completion. There is a large body of UDA work following this methodology (Hoffman et al., 2016, 2018a; Chen et al., 2018a; Hong et al., 2018; Benjdira et al., 2019). A different set of approaches attempts direct distribution alignment between the source and target domains. Distribution alignment can then be achieved by minimizing an appropriate distributional distance metric (Zhang et al., 2017, 2019; Wu et al., 2018; Gabourie et al., 2019; Lee et al., 2019; Yang and Soatto, 2020), such as $l2$-distance, *KL* divergence, or Wasserstein Distance.

While both types of approaches are able to obtain state-of-the-art (SOTA) results on UDA semantic segmentation tasks, most methods assume simultaneous access to both source and target data samples. This benefits model stability during adaptation as source domain access ensures a gradual shift of the decision function. However, in real-world settings, there are many situations where concomitant access to both domains cannot be achieved. For instance, datasets may need to be stored on different servers due to latency constraints (Xia et al., 2021) or data privacy requirements (Li et al., 2020b). To adhere to these settings, UDA has been extended to situations where the source domain data are no longer accessible during adaptation. This class of methods is named *source-free adaptation* and provides a balance between accuracy and privacy (Kim et al., 2021; Kundu et al., 2021b). Compared to regular UDA, source-free UDA is less explored. Our approach addresses source-free adaptation, making it a suitable algorithm for scenarios where data privacy is an issue. Moreover, our proposed method is based on common DNN architectures for semantic segmentation and requires little parameter fine-tuning compared to adversarial approaches.

Contributions: We propose a novel algorithm that performs source-free UDA for semantic segmentation tasks. Our approach eliminates the need for access to source data during the adaptation phase, by approximating the source domain via a internal intermediate distribution. During adaptation, our method aligns the target and intermediate domain via a suitable distance metric to ensure classifier generalization on target features. We demonstrate the performance of our method on two benchmark semantic segmentation tasks, GTA5→CITYSCAPES and SYNTHIA→CITYSCAPES, where the source datasets are composed of computer generated images, and the target datasets are real-world segmented images. We offer theoretical justification for our algorithms' performance, proving our approach minimizes target error under our adaptation framework. We evaluate our approach on well-established semantic segmentation datasets and demonstrate it compares favorably against state-of-the-art (SOTA) UDA semantic segmentation methods [partial results of this study were presented in the AAAI Conference (Stan and Rostami, 2021)].

2 Related work

We provide an overview of semantic segmentation algorithms, as well as describe recent UDA and source-free UDA approaches for this setting.

2.1 Semantic segmentation

Compared to image classification problems, semantic segmentation tasks are more complicated because we require each pixel of an image to receive a label, which is part of a set of semantic categories. As each image dimension may have thousands of pixels, semantic segmentation models require powerful encoder/decoder architectures capable of synthesizing large amounts of image data and encode the spatial relationships between the pixels well. Recent SOTA results for supervised semantic segmentation have thus been obtained by the use deep neural networks (DNNs), and in particular convolutional neural networks (CNNs) (LeCun and Bengio, 1995), which are specialized for image segmentation. While different architecture variants exist (Long et al., 2015; Chen et al., 2018b; Tao et al., 2020; Wang et al., 2020), approaches often rely on embedding images into a latent feature space via a CNN feature extractor, followed by an up-sampling decoder which scales the latent space back to the original input space, where a classifier is trained to offer pixel-wise predictions. The idea is if the extracted features can reconstruct the input image with a relatively high accuracy, then they carry an information content similar to the input distribution, yet in a lower dimensional space. Skip connections between different levels of the network (Ronneberger et al., 2015; Lin et al., 2017), using dilated convolutions (Chen et al., 2017a) or using transformer networks as feature extractors instead of CNNs (Strudel et al., 2021) have been shown to improve supervised baselines.

While improvements in supervised segmentation are mostly tied to architecture choice and parameter fine-tuning, model generalization suffers when changes in the input distribution are made. This phenomenon is commonly referred to as *domain shift* (Sankaranarayanan et al., 2018). Changes in the input distortion translate into shifted extracted features that do not match the internal distribution learned by the DNN. This issue is common in application domains where the same model needs to account for multi-modal data, and the training set lacks a certain mode, e.g., daylight and night-time images (Romera et al., 2019), clear weather and foggy weather (Sakaridis et al., 2018), and medical data obtained from different imaging devices and scanners (Guan and Liu, 2021). Such differences in input data distributions between source and target domains greatly impact the generalization power of learnt models. When domain shift is present, source-only training may be at least three-fold inferior compared to training the same model on the target dataset (Hoffman et al., 2016, 2018a;

Lee et al., 2019). While it is possible to retrain the model to account for distribution shifts, we will require to annotate data again which can be time-consuming and expensive in semantic segmentation tasks. Because label information is expensive to obtain, adding a cost to using such techniques, especially on new datasets. Weakly supervised approaches explore the possibility of having access to limited label information after domain shift to reduce the data annotation requirement (Hung et al., 2018; Wei et al., 2018; Wang et al., 2019). However, data annotation is still necessary.

On the other hand, due to the limited label availability for semantic segmentation tasks, the use of synthetic images and labels has become an attractive alternative for training segmentation models even if domain shift is not a primary concern. The idea is to prepare a synthetic source dataset which can be annotated automatically. Semantic labels are easy to generate for virtual images, and a model trained on such images could then be used on real-world data as a starting point. Overcoming domain shift becomes the primary bottleneck for successfully applying such models to new domains.

2.2 Unsupervised domain adaptation

Unsupervised domain adaptation (UDA) addresses model generalization in scenarios where target data label information is unavailable but there a source domain with annotated data that shares the same labels with the target domain problem. UDA techniques primarily employ a shared feature embedding space between the source and target domain in which the distributions of both domains are aligned. A majority of these methods achieve this goal by either using domain discriminators based on adversarial learning or direct source-target feature alignment based on metric minimization.

2.2.1 Adversarial adaptation for UDA

Techniques based on adversarial learning employ the idea of domain discriminator, used in GANs (Goodfellow et al., 2014), to produce a shared source/target embedding space. A discriminator is tasked with differentiating whether two image encodings originate from the same domain, or one is from the source and one is from the target. A feature encoder aims to fool the discriminator, thus producing source/target latent features more and more similar in nature as training progressed. Over the course of training, this leads the feature extractor producing a shared embedding space for the source and target data.

In the context of UDA for semantic segmentation, Luc et al. (2016) employ an image segmentation model and adversarially train a semantic map generator, which uses a label map discriminator to penalize the segmentation network for producing label maps more similar to the generated ones rather than the source ones. Murez et al. (2017) use an intermediate feature distribution that attempts to capture domain-agnostic features of both source and target datasets. To improve the domain-agnostic representation, a discriminator is trained to differentiate whether an encoded image is part of the source or target domain. The encoder networks are then adversarially trained to fool the discriminator, resulting in similar embeddings between source and target samples. Bousmalis et al. (2017) develop a model for pixel-level domain adaptation by creating a pseudo-dataset by making source samples appear as though they are extracted from the target domain. They use generative model to adapt source samples to the style of target samples, and a domain discriminator to differentiate between real and fake target data. Hoffman et al. (2018b) employ the cycle consistency loss proposed in Zhu et al. (2017) to improve the adversarial network adaptation performance. In addition to this, Hoffman et al. (2018b) use GANs to stylistically transfer images between source and target domains, and use a consistency loss to ensure network predictions on the source image will be the same as in the stylistically shifted variant. Saito et al. (2018) use an approach based on a discriminator network without using GANs to attempt to mimic source or target data distributions. They propose the following adversarial learning process on a feature encoder network with two classification heads: (1) they first keep the feature encoder fixed and optimize the classifiers to have their outputs as different as possible, (2) they freeze the classifiers and optimize the feature encoder such that both classifiers will have close outputs. Sankaranarayanan et al. (2018) employ an image translation network that is tasked with translating input images into the target domain feature space. A discriminator is tasked with differentiating source images from target images passed through the network, and a similar procedure is done for target images. A pixel-level cross entropy loss ensures the network is able to perform semantic segmentation. Lee et al. (2019) use a similar idea to Saito et al. (2018) in that a network with two classifiers is used for adaptation. The feature extractor and classification heads are trained in an alternating fashion. The study of Lee et al. (2019) differentiates itself by employing an approximation of optimal transport to compute these discrepancy metrics, leading to improved performance over (Saito et al., 2018).

2.2.2 Adaptation by distribution alignment

Adaptation methods using direct distribution alignment share the same goal as adversarial methods. However, distribution alignment is reached by directly minimizing an appropriate distributional distance metric between the source and target embedding feature distributions.

Wu et al. (2018) propose an image translation network that takes as input source and target images, and outputs source images in the style of the target domain. Their proposed architecture does not use adversarial training, rather is based on the idea that in order for stylistic transfer to be achieved, domain mean and variance should be similar at different levels of abstraction throughout the translation network. They achieve this goal by minimizing ℓ_2-distance in the feature space at various levels of abstraction. Zhang et al. (2017) develop a method for semantic segmentation adaptation by observing that a source trained model should produce the same data statistics on the target domain as present in the annotated source distribution. Examples include label distribution or pixels of a certain class clustering around specific regions in an image. Pseudo-labeling is used to estimate these statistics. To enforce similarity in the output of a source-trained model to the estimated target statistics, KL divergence is

used as a minimization metric. Zhang et al. (2019) employ an adaptation framework based on the idea that source and target latent features should cluster together in similar ways. They use a pseudo-labeling approach to produce initial target labels, followed by minimizing the distance between class specific latent feature centroids between source and target domains. The minimization metric of choice is ℓ_2-distance. For improved performance, they use category anchors to align the adaptation process. Gabourie et al. (2019) propose an adaptation method based on a shared network encoder between source and target domains. Their model is trained by minimizing cross entropy loss for the source samples and is tasked with minimizing the distance between source and target embeddings in the latent feature space. To achieve this goal, Sliced Wasserstein Distance is minimized between the source and target embeddings, leading to improved classifier performance on target samples.

The expectation that continuous access to source data is guaranteed when performing UDA is not always true, especially in the case of privacy sensitive domains. This setting of UDA has been previously explored by methods that do not employ DNNs (Dredze and Crammer, 2008; Jain and Learned-Miller, 2011; Wu, 2016) and has recently become the focus of DNN based algorithms for image classification tasks (Saltori et al., 2020; Kim et al., 2021; Kundu et al., 2021b; Yang et al., 2021). Source-free semantic segmentation has been explored relatively less compared to joint UDA adaptation approach. Kundu et al. (2021a) employ source domain generalization and target pseudo-labeling in the adaptation method. Liu et al. (2021) rely on self supervision and patch level optimization for adaptation. You et al. (2021) allow models trained on synthetic data to generalize on real data by a mixture of positive and negative class inference.

Our adaptation approach shares the idea of direct distribution alignment. As described previously, several choices for latent feature alignment have been previously explored, such as $l2$-distance (Wu et al., 2018), KL divergence (Zhang et al., 2017), or Wasserstein Distance (WD) (Gabourie et al., 2019; Lee et al., 2019). WD has been proven to leverage the geometry of the embedding space better than other distance metrics (Tolstikhin et al., 2017). Empirically, the behavior of using the Wasserstein metric has been observed to benefit the robustness of training deep models, such as in the case of the Wasserstein GAN (Arjovsky et al., 2017), or by improving the relevance of discrepancy measures, as reported by (Lee et al., 2019). One of the limitations of using the WD is the difficulty of optimizing this quantity, as computing the WD distance requires solving a linear program. Therefore, we employ an approximation of this metric, the Sliced Wasserstein Distance (SWD), which maintains the nice metric properties of the WD while allowing for an end-to-end differentiation in the optimization process.

We base our source-free UDA approach on estimating the latent source embeddings via an internal distribution (Rostami, 2019). This approximation relies on the concept that a supervised model trained on K classes will produce a K modal distribution in its latent space. This property of the internal distribution allows us to perform adaptation without direct access to source samples. The idea is to approximate the internal distribution and then sample from the K modal distributional approximation to use them as a surrogate for the source domain distribution. The distribution approximation introduces a small number of parameters into our model. Once we produce a pseudo-dataset from sampling the internal distribution, we align the target feature encodings by minimizing the SWD between the two data distributions. Our theoretical bounds demonstrate our approach leads to minimizing an upperbound for the target domain error.

3 Problem formulation

Let $\mathcal{P}_\mathcal{S}$ be the data distribution corresponding to a source domain, and $\mathcal{P}_\mathcal{T}$ be similarly the data distribution corresponding to a target domain, with $\mathcal{P}_\mathcal{S}$ being potentially different from $\mathcal{P}_\mathcal{T}$. We consider a set of multi-channel images $\boldsymbol{X_S}$ is randomly sampled from $\mathcal{P}_\mathcal{S}$ with corresponding pixel-wise semantic labels $\boldsymbol{Y_S}$. Let $\boldsymbol{X_T}$ be a set of images sampled from $\mathcal{P}_\mathcal{T}$, where we do not have access to the corresponding labels $\boldsymbol{Y_T}$. We consider that both $\boldsymbol{X_S}$ and $\boldsymbol{X_T}$ are represented as images with real pixel values in $\mathbb{R}^{W\times H\times C}$, where W is the image width, H is the image height, and C is the number of channels. The labels $\boldsymbol{Y_S}, \boldsymbol{Y_T}$ share the same input space of label maps in $\mathbb{R}^{W\times H}$ which makes the two domain relevant.

Our goal is to learn the parameters θ of a semantic segmentation model $\phi_\theta(\cdot):\mathbb{R}^{W\times H\times C} \rightarrow \mathbb{R}^{W\times H}$ capable of accurately predicting pixel-level labels for images sampled from the target distribution $\mathcal{P}_\mathcal{T}$. We can formulate this problem as a supervised learning problem, where our goal is to minimize the target domain empirical risk, achieved by $\theta^* = \arg\min_\theta\{\mathbb{E}_{x^t\sim P_\mathcal{T}(X^t)}(\mathcal{L}(f_\theta(x^t), y^t)\}$, where $x^t \in \boldsymbol{X_T}, y^t \in \boldsymbol{Y_T}$. The difficulty of the above optimization stems from the lack of access to the label set $\boldsymbol{Y_T}$. To overcome this challenge, we instead are provided access to the labeled source domain $(\boldsymbol{X_S}, \boldsymbol{Y_S})$, and then sequentially the target domain $\boldsymbol{X_T}$. Many UDA algorithms consider that both domains are accessible simultaneously but the source-free nature of our problem requires that once the target images $\boldsymbol{X_T}$ become available, access to source domain information becomes unavailable. This assumption is a practical assumption because domain shift is often a temporal problem that arises after the initial training phase.

To achieve training a generalizable model for the target domain, we need to first train a model on the provided source dataset and then adapted to generalize well on the target domain. Let N be the size of the source dataset $\boldsymbol{X_S}$, and let (x_i^s, y_i^s) be the image/label pairs from $\boldsymbol{X_S}, \boldsymbol{Y_S}$. Consider K to be the number of semantic classes and $\mathbb{1}_a(b)$ denote the indicator function determining whether a and b are equal. Then, we learn the parameters that minimize empirical risk on the source domain by optimizing the standard cross entropy loss on the labeled dataset:

$$\begin{aligned}\hat{\theta} &= \arg\min_\theta\{\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_{ce}(\phi_\theta(x_i^s), y_i^s)\}\\ \mathcal{L}_{ce}(p, y) &= -\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{k=1}^{K}\mathbb{1}_{y_{wh}}(k)\log(p_{wh}),\end{aligned} \tag{1}$$

The optimization setup in Equation 1 ensures model generalization on inputs sampled from $\mathcal{P}_\mathcal{S}$. In cases where $\mathcal{P}_\mathcal{T}$ differs from $\mathcal{P}_\mathcal{S}$, the model will not generalize on the target

domain, due to domain shift. To account for domain shift, we need to map data points from both domains into an invariant feature space between the two domains, without joint access to $\boldsymbol{X_S}$ and $\boldsymbol{X_T}$. To this end, let $f(\cdot), g(\cdot), h(\cdot)$ be three parameterized sub-networks such that $\phi = f \circ g \circ h$. In this composition, $f : \mathbb{R}^{W\times H\times C} \rightarrow \mathbb{R}^L$ is an encoder sub-network, $g : \mathbb{R}^L \rightarrow \mathbb{R}^{W\times H}$ is an up-scaling decoder, and $\mathbb{R}^{W\times H} \rightarrow \mathbb{R}^{W\times H}$ is a semantic classifier, where L represents the dimension of the latent network embedding. To create a shared source-target embedding space, our goal is that the network $(f \circ g)(\cdot)$ embeds source and target samples in a shared domain-agnostic embedding space. Under this condition, the classifier $h(\cdot)$ trained on source domain samples will be able to generalize on target inputs.

We can make the shared embedding space domain-agnostic by direct distribution alignment between the embeddings of the two domains at the decoder output. As previously explored in literature (Gabourie et al., 2019), a suitable distributional distance metric $\mathcal{D}(\cdot,\cdot)$ can be minimized between the source and target domain data points at the network $(f \circ g)(\cdot)$ output. However, because the source domains are inaccessible during model adaptation, we cannot compute the distribution distance between the two domains. Hence, directly minimizing $\mathcal{D}(f \circ g(\boldsymbol{X_S}), f \circ g(\boldsymbol{X_T}))$ is not feasible. We need to develop a solution that relaxes the need for access to the source domain samples during adaptation for domain adaptation. Our core idea is to benefit from another distribution that can be served as a surrogate for the source domain distribution. We describe our source domain approximation approach and the choice for $\mathcal{D}(\cdot,\cdot)$ in the next section.

4 Proposed algorithm

We visually describe our method in Figure 1. The first step of our approach is to fully train a segmentation model on the labeled source domain. As training progresses on the source domain, the latent feature space will begin to cluster into K clusters, where each of the clusters encode one of the semantic classes. If we use the output of the softmax layer as our embedding space, the softmax classifier will be able to learn a linear decision function based on the decoder output which leads to high label accuracy at the end of this pre-training stage. After the source-training stage, we approximate the source distribution via a learnt internal distribution. We use this as a surrogate for $f \circ g(\boldsymbol{X_S})$ during adaptation.

As linear separation in the latent space is reached, we can produce an approximation of the source domain distribution in the embedding space and thus relax the need for having access to the source samples during adaptation. We are interested in learning a K-modal approximation to the latent feature distribution at the decoder output, $f \circ g(\mathcal{P}_\mathcal{S})$. Let $p_k, 1 \leq k \leq K$ represent the component of $f \circ g(\mathcal{P}_\mathcal{S})$ corresponding to class k. This characteristic means that we should use a multi-modal distribution for approximating the data distribution in the embedding space. A multi-modal distribution possess distinct "modes," signifying the difference between these modes. These modes represent the most frequently occurring values in the data set. Each mode a local maximum in the distribution. The presence of multiple modes indicates that the data has more than one central tendency or characteristic value. As a result, these distributions can be used to approximate the distribution of the data when we have separable classes that are distinct and different from each other in a feature space.

For approximation purposes, we will a Gaussian mixture model (GMM), with each semantic class approximated by T high Gaussians components, i.e, the GMM would have kT components in total. Our choice for this approximation method stems from the result by Goodfellow et al. (2016) that concludes with sufficient Gaussians, any distribution can be approximated to vanishing error. While the GMM model is traditionally learned in an unsupervised fashion, we can leverage our knowledge of the source labels to partially supervise the process. As we have access to the source domain labeled data, we can directly identify which latent feature vectors correspond to each of the K classes. Once the latent feature vectors are pooled for each class, a T component GMM is learned using expectation maximization. During the learning process, we attempt to avoid inclusion of outlier elements in the GMMs, which may lead to decreased class separability in the latent space. This is a detrimental outcome for us as an increased separability improves the performance of the classification layer. We thus only consider data samples which have high associated classifier confidence. Let τ be this confidence threshold, and let $S_k = \{u_{i,j} | \exists (x^s, y^s), f \circ g(x^s)_{i,j} = u_{i,j}, y^s_{i,j} = k, f \circ g \circ h(x^s)_{i,j,k} > \tau\}$ be the set of source embedding feature vectors at the decoder output which have label k and on which the classifier assigns to k more than τ probability mass. Then, we use expectation maximization to learn $\alpha_{k,t}, \mu_{k,t}, \Sigma_{k,t}, 1 \leq t \leq T$ as the parameters of the T components approximating class k. Thus, for each semantic class k, we model the latent feature distribution p_k as Equation (2):

$$p_k(z) = \sum_{t=1}^{T} \alpha_{k,t} \mathcal{N}(z | \mu_{k,t}, \Sigma_{k,t}) \tag{2}$$

Learning a GMM approximation for each semantic class k alleviates the need for source domain access during adaptation. Once adaptation stage starts, the source domain becomes unavailable, however, we use the learnt GMM approximation distribution as a surrogate. We achieve classifier generalization on the target domain by minimizing a distributional distance metric between then GMM approximation and the target latent embeddings. For this purpose, consider the dataset $\mathcal{Z} = (\boldsymbol{X_Z}, \boldsymbol{Y_Z})$ produced by sampling from the GMM distribution, with $N^z = |\mathcal{Z}|$. Let (x_i^z, y_i^z) be embedding/label pairs from this dataset. We achieve distribution alignment by empirically minimizing an appropriate distributional distance metric D between samples from $\mathcal{Z}$ and from the target embeddings. In addition to distribution alignment, we need to account for shifts in the classifier input space between the internal distribution and the original source embedding distribution. We account for such shifts by fine tuning the classifier on labeled GMM samples. Our adaptation loss can be formalized as:

$$\mathcal{L}_{adapt} = \mathcal{L}_{ce}(h(\boldsymbol{X_Z}), \boldsymbol{Y_Z}) + \lambda \mathcal{L}_D(f \circ g(\boldsymbol{X_T}), \boldsymbol{X_Z}) \tag{3}$$

for an appropriate choice of regularizer λ.

The first loss term in Equation (3) is the cross entropy classifier fine-tuning loss obtained for the GMM samples across the whole

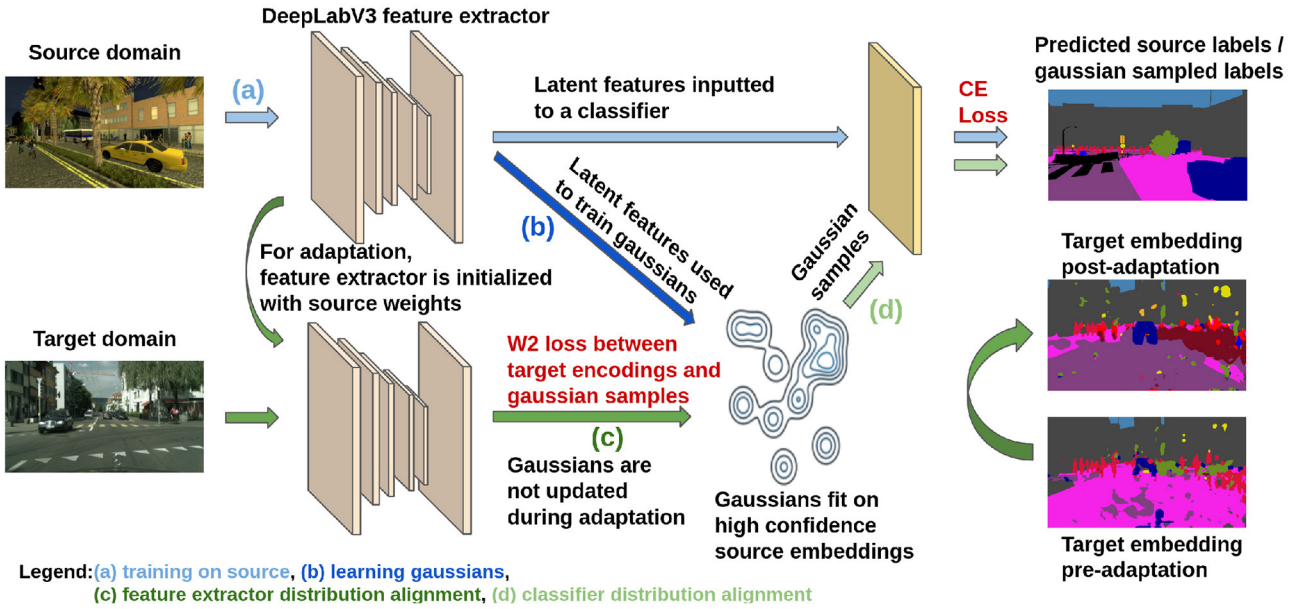

FIGURE 1
Diagram of the proposed model adaptation approach (best seen in color): **(a)** initial model training using the source domain labeled data, **(b)** estimating the internal distribution as a GMM distribution in the embedding space, **(c)** domain alignment is enforced by minimizing the distance between the internal distribution samples and the target unlabeled samples, **(d)** domain adaptation is enforced for the classifier module to fit correspondingly to the GMM distribution.

sampling dataset, i.e. Equation (4),

$$\mathcal{L}_{ce}(h(\boldsymbol{X_Z}), \boldsymbol{Y_Z}) = -\frac{1}{N^z}\sum_{i=1}^{N^z}\frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\sum_{k=1}^{K} \mathbb{1}_{y^z_{i,wh}}(k)\log(h(x^z_i)_{wh}) \tag{4}$$

where (x^z_i, y^z_i) are the i'th GMM data point in the sampling dataset $\mathcal{Z}$. This loss terms helps maintaining model generalizability as we perform distribution alignment.

The second loss term in Equation (3), $\mathcal{L}_D$, represents the distributional distance metric between the GMM in the latent space and the target domain data embedding vectors. We choose the Sliced Wasserstein Distance (SWD) as our choice for the distributional distance metric D. In the context of domain adaptation, several distribution alignment metrics have been previously used. Wu et al. (2018) propose an approach where the feature space between source and target images is made similar by directly minimizing the ℓ_2-distance between feature vectors. *KL* divergence has been used in domain adaptation (Yu et al., 2021) to detect noisy samples or target samples from private classes. Wasserstein Distance (WD) has been explored as a distributional distance metric (Gabourie et al., 2019) by directly minimizing the metric on the output feature space for a source and target encoder. While study of appropriate distributional distance metrics is still ongoing, the WD aims to find the optimal way of moving mass between two distributions and thus is tied to the geometry of the data. This has lead to the WD offering improved stability when used in a number of deep learning and domain adaptation tasks (Solomon et al., 2015; Kolouri et al., 2016; Arjovsky et al., 2017; Bhushan Damodaran et al., 2018; Rostami et al., 2019; Li et al., 2020a; Xu et al., 2020; Rostami and Galstyan, 2023). The WD metric (Kolouri et al., 2019) between two distributions P and Q is defined as Equation (5):

$$W_d(P,Q) = \left(\inf_{L\in\mathcal{L}(P,Q)}\int \|x-y\|^d dL(x,y)\right)^{\frac{1}{d}} \tag{5}$$

where $\mathcal{L}(P,Q)$ represents all transportation plans between P and Q, i.e., all joint distributions with marginals P and Q. The WD metric offers a closed-form solution only when P and Q are one-dimensional distributions (Kolouri et al., 2019). For higher dimensions, we need to solve a linear program. While employing the WD has desirable properties, solving a linear program at every optimization step can lead to significant computational costs in the adaptation phase of UDA. To alleviate this issue, we employ the SWD, an alternative for the WD that is fully differentiable, yet has a closed-form formula. Computing the SWD between to high dimensional distributions involves repeatedly projecting them along random on dimensional projection directions, obtaining one-dimensional marginals for which computation of the WD which has a closed-form solution. This process allows for an end-to-end differentiation via gradient based methods, such as Stochastic Gradient Descent (Bottou et al., 2018). Averaging one-dimensional WD over sufficient random projection directions will produce a closed-form approximation to the high dimensional WD objective. We can use the distributional distance term in Equation (3) for two distributions p, q as follows using Equation (6):

$$\mathcal{L}_D(p,q) = \mathcal{SWD}_d(p,q) = \frac{1}{J}\left(\sum_{i=1}^{J}\|\gamma_i p - \gamma_i q\|^d\right)^{\frac{1}{d}} \tag{6}$$

Initial training:
  Input: source domain dataset $\mathcal{D}_S = (\boldsymbol{X}_S, \boldsymbol{Y}_S)$,
    Training on source domain:
      $\hat{\theta}_0 = (\hat{\boldsymbol{w}}_0, \hat{\boldsymbol{v}}_0 \hat{\boldsymbol{u}}_0) = \arg\min_\theta \sum_i \mathcal{L}(f_\theta(\boldsymbol{x}_i^s), \boldsymbol{y}_i^s)$
    Internal distribution estimation:
      Estimate the GMM parameters
Model adaptation:
    Input: target dataset $\mathcal{D}_\mathcal{T} = (\boldsymbol{X}_\mathcal{T})$
    Pseudo-dataset generation:
      $\mathcal{D}_\mathcal{P} = (\mathbf{Z}_\mathcal{P}, \mathbf{Y}_\mathcal{P}) =$
            $([\boldsymbol{z}_1^p, \ldots, \boldsymbol{z}_N^p], [\boldsymbol{y}_1^p, \ldots, \boldsymbol{y}_N^p])$, where:
              $\boldsymbol{z}_i^p \sim \hat{p}_J(\boldsymbol{z}), 1 \leq i \leq N_p$
              $\boldsymbol{y}_i^p = \arg\max_j \{h_{\hat{\boldsymbol{w}}_0}(\boldsymbol{z}_i^p)\}$, $p_{ip} > \tau$
for $itr = 1, \ldots, ITR$ **do**
    draw random batches from $\mathcal{D}_\mathcal{T}$ and $\mathcal{D}_\mathcal{P}$
    Update the model by solving **Equation (3)**
end for

Algorithm 1. MAS^3 (λ, τ)

where $\mathcal{SWD}_d$ represents the d order SWD, J represents the number of random projection to be averaged, and γ_i is one of the J random projections. In our approach, we will choose $\mathcal{SWD}_2$ due to ease of computation and comparable performance to higher order choices of d. Pseudocode for our approach, named Model Adaptation for Source-Free Semantic Segmentation (MAS^3) is provided in Algorithm 1.

5 Theoretical analysis

We prove Algorithm 1 can lead to improving the model generalization on the target domain by minimizing an upperbound for the empirical risk of the model on the target domain. For such a result, we need to tie model generalization on the source domain to the distributional distance between the source and target domains. For this purpose, we use the framework developed by Redko et al. (2017) designed for upper bounding target risk with respect to the distance between the source and target domains in the classic joint UDA process. We rely on the following Theorem 2 from Redko et al. (2017) in our approach:

Theorem 1. *(Redko et al., 2017) For the variables defined under Theorem 2, the following distribution alignment inequality loss holds:*

$$\epsilon_T \leq \epsilon_S + W(\hat{\mu}_S, \hat{\mu}_T) + \sqrt{(2\log(\frac{1}{\xi})/\zeta)}\left(\sqrt{\frac{1}{N^s}} + \sqrt{\frac{1}{N^t}}\right) + e_\mathcal{C}(h^*) \tag{7}$$

The above relation characterizes target error after source training and does not consider our specific scenario of using an intermediate distribution. We adapt this bound for Algorithm 1 to derive the following theorem:

Theorem 2. *Consider the space of all possible hypotheses $\mathcal{H}$ applicable to the proposed segmentation task. Let $\epsilon_S(h), \epsilon_T(h)$ represent the expected source and target risk for hypothesis h, respectively. Let $\hat{\mu}_S, \hat{\mu}_Z, \hat{\mu}_T$ be the empirical mean of the embedding space for the source, intermediate and target datasets respectively. Let $W(\cdot,\cdot)$ represented the Wasserstein distance, and let ξ, ζ be appropriately defined constants. Consider $e_\mathcal{C}(h)$ to be the combined error of a hypothesis h on both the source and target domains, i.e., $\epsilon_S(h) + \epsilon_T(h)$, and let h^* be the minimizer for this function. Then, for a model h, the following results holds:*

$$\begin{aligned}\epsilon_T(h) \leq \epsilon_S(h) + W(\hat{\mu}_S, \hat{\mu}_Z) + W(\hat{\mu}_Z, \hat{\mu}_T) \\ + \sqrt{(2\log(\frac{1}{\xi})/\zeta)}\left(\sqrt{\frac{1}{N^s}} + \sqrt{\frac{1}{N^t}}\right) + e_\mathcal{C}(h^*)\end{aligned} \tag{8}$$

Proof: We expand the second term of Equation (7). Given $W(\cdot,\cdot)$ is a convex optimization problem, we can use the triangle inequality as follows:

$$W(\hat{\mu}_S, \hat{\mu}_T) \leq W(\hat{\mu}_S, \hat{\mu}_Z) + W(\hat{\mu}_Z + \hat{\mu}_T) \tag{9}$$

Combining Equations (8, 7) leads to the result in Theorem 2.

The above results provides a justification Algorithm 1 is able to minimize the right hand side of the Equation (8). The first term is minimized during the initial training phase on the source domain. Note that, as expected, the performance on the target domain cannot be better than the performance on the source domain. We conclude that the model we select should be a good model to learn the source domain. The second term represents the WD distance between the source and sampling dataset. This distance will be small if the GMM approximation of the source domain will be successful. As we explained before, if select a large enough T, we can make this term negligible. The third term is the WD distance e between the sampling dataset and the target domain dataset. This term is directly minimized by the adaptation loss that we use to align the distribution. The term $1 - \tau$ is a constant directly dependent on the confidence threshold τ, which we choose close to 1. The fourth term is directly dependent on the dataset size and becomes small when a large number of samples is present. Finally, the last term is a constant indicating the difficulty of the adaptation problem.

6 Experimental validation

We validate the proposed algorithm using common UDA benchmarks for semantic segmentation. Our implementation code is available as a supplement at https://github.com/serbanstan/mas3-continual.

6.1 Experimental setup

6.1.1 Datasets

We follow the UDA literature to evaluate our approach. We consider three common datasets used in semantic segmentation literature: GTA5 (Richter et al., 2016), SYNTHIA (Ros et al., 2016),

and Cityscapes (Cordts et al., 2016). Both GTA5 and SYNTHIA are datasets consisting of artificially generated street images, with 24,966 and 9,400 instances, respectively. Cityscapes is composed of real-world images recorded in several European cities, consisting of 2,957 training images and 500 test images. We can see that the diversity of sizes of these datasets which demonstrates the challenge of data annotation for semantic segmentation tasks. For all three datasets, images are processed and resized to a standard shape of $512 \times 1,024$.

Following the literature, we consider two adaptation tasks designed to evaluate model adaptation performance when the training set consists of artificial images, and the test set consists of real-world images. For both SYNTHIA→Cityscapes and GTA5→Cityscapes, we evaluate performance under two scenarios, when 13 or 19 semantic classes are available.

6.1.2 Implementation and training details

We use a DeepLabV3 architecture (Chen et al., 2017a) with a VGG16 encoder (Simonyan and Zisserman, 2014) for our CNN architecture. The decoder is followed by a 1×1 convolution softmax classifier. We choose a batch size of 4 images for training and use the Adam optimizer with learning rate of $1e-4$ for gradient propagation. For adaptation, we keep the same optimizer parameters as for training. We choose 100 projections in our SWD computation and set the regularization parameter λ to 0.5. We perform training for $100k$ iterations on the source domain and then fpr adaptation we perform $50k$ iterations.

When approximating the GMM components, we chose the confidence parameter τ to be 0.95. We observe higher values of τ to be correlated with increased performance, as expected from our theorem, and conclude that a τ setting above 0.9 will lead to similar target performance.

We run our experiments on a NVIDIA Titan XP GPU. Given that our method relies on distributional alignment, the label distribution between target batches may vary significantly between different batches. As the batch distribution approaches the target label distribution as the batch size increases, we use the oracle label distribution per batch when sampling from the GMM, which can be avoided if sufficient GPU memory becomes present. Experimental code is provided with the current submission.

6.1.3 Baselines for comparison

Source-free model adaptation algorithms for semantic segmentation have been only recently explored. Thus, due to most UDA algorithms being designed for joint training, in addition to source-free approaches we also include both pioneer and recent UDA image segmentation method to be representative of the literature. We have compared our performance against the adversarial learning-based UDA methods: GIO-Ada (Chen et al., 2019), ADVENT (Vu et al., 2019), AdaSegNet (Tsai et al., 2018), TGCF-DA + SE (Choi et al., 2019), PCEDA (Yang et al., 2020), and CyCADA (Hoffman et al., 2018b). We have also included methods that are based on direct distributional matching which are more similar to MAS^3: FCNs in the Wild (Hoffman et al., 2016), CDA (Zhang et al., 2017), DCAN (Wu et al., 2018), SWD (Lee et al., 2019), and Cross-City (Chen et al., 2017b). Source-free methods include GenAdapt (Kundu et al., 2021a) and SFDA (Liu et al., 2021). We also added a joint UDA version of our method, named MAS^3-Joint, in which we used SWD to directly align the distributions. This baseline offers the performance of our algorithm when the source domain samples are directly accessible.

6.2 Comparison results

6.2.1 SYNTHIA→cityscapes task

We provide quantitative and qualitative results for this task in Table 1. We report the performance our method produces on the SYNTHIA→CITYSCAPES adaptation task along with other baselines. Notably, even when confronted with a more challenging learning setting, MAS^3 demonstrates superior performance compared to the majority of classic UDA methods that have access to the source domain data during model adaptation. It is essential to highlight that some recently developed UDA methods leveraging adversarial learning surpass our approach in performance; however, it is worth noting that these methods often incorporate an additional form of regularization, aside from probability matching and are unable to address UDA when the source domain data is missing.

In an overarching evaluation, MAS^3 exhibits commendable performance, particularly when compared to UDA methods that rely on source samples. Furthermore, our method excels in specific crucial categories, such as the accurate detection of traffic lights, where it outperforms its counterparts. These results underscore the robustness and effectiveness of MAS^3 in handling challenging learning scenarios and achieving notable performance, especially in key object categories. We conclude that MAS^3 can be used to address classic UDA setting reasonably well. Interestingly, we observe that MAS^3 does not lead to a good performance. This poor performance can be attributed to the fact that when we use GMM samples, we have far more samples than the original distribution for aligning the two distributions.

6.3 GTA5→cityscapes task

We present the quantitative outcomes for this particular task, detailed in Table 2. It is noteworthy that we observe a more competitive performance in this task, and yet the overall trend in the performance comparison remains similar to Table 1. These findings highlight the versatility of our proposed method, MAS^3. While our primary motivation lies in achieving source-free model adaptation, these results indicate that MAS^3 can effectively function as a joint-training UDA algorithm. We conclude our method manages to achieve state-of-the-art performance even in a setting involving a larger number of semantic classes. This capability underscores the robustness and adaptability of MAS^3 in diverse scenarios, making it a versatile solution that goes beyond its original focus on source-free model adaptation.

TABLE 1 **Model adaptation comparison results for the SYNTHIA→Cityscapes task.**

Method	Adv.	Road	sdwlk	bldng	Light	Sign	vgttn	Sky	Person	Rider	Car	Bus	mcycl	bcycl	mIoU
Source only (VGG16)	N	6.4	17.7	29.7	0.0	7.2	30.3	66.8	51.1	1.5	47.3	3.9	0.1	0.0	20.2
FCNs in the wild	N	11.5	19.6	30.8	0.1	11.7	42.3	68.7	51.2	3.8	54.0	3.2	0.2	0.6	22.9
CDA	N	65.2	26.1	74.9	3.7	3.0	76.1	70.6	47.1	8.2	43.2	20.7	0.7	13.1	34.8
DCAN	N	9.9	30.4	70.8	6.70	23.0	76.9	73.9	41.9	16.7	61.7	11.5	10.3	38.6	36.4
SWD	N	83.3	35.4	82.1	12.2	12.6	83.8	76.5	47.4	12.0	71.5	17.9	1.6	29.7	43.5
Cross-City	Y	62.7	25.6	78.3	1.2	5.4	81.3	81.0	37.4	6.4	63.5	16.1	1.2	4.6	35.7
GIO-Ada	Y	78.3	29.2	76.9	10.8	17.2	81.7	81.9	45.8	15.4	68.0	15.9	7.5	30.4	43.0
ADVENT	Y	67.9	29.4	71.9	0.6	2.6	74.9	74.9	35.4	9.6	67.8	21.4	4.1	15.5	36.6
AdaSegNet	Y	78.9	29.2	75.5	0.1	4.8	72.6	76.7	43.4	8.8	71.1	16.0	3.6	8.4	37.6
TGCF-DA + SE	Y	90.1	48.6	80.7	3.2	14.3	82.1	78.4	54.4	16.4	82.5	12.3	1.7	21.8	46.6
PCEDA	Y	79.7	35.2	78.7	10.0	28.9	79.6	81.2	51.2	25.1	72.2	24.1	16.7	50.4	48.7
SFDA	SF(Y)	81.9	44.9	81.7	3.3	10.7	86.3	89.4	37.9	13.4	80.6	25.6	9.6	31.3	45.89
GenAdapt	SF(Y)	89.9	48.8	80.9	19.5	26.2	83.7	84.9	57.4	17.8	75.6	28.9	4.3	17.2	48.9
MAS^3-Joint	N	66.9	23.7	66.0	4.2	4.3	75.1	60.7	22.3	2.7	17.4	3.5	0.7	0.8	27.4
MAS^3	SF(N)	74.8	51.6	71.5	20.4	32.3	73.0	75.3	48.9	19.7	66.3	25.7	10.1	40.8	47.0

We have used DeepLabV3 (Chen et al., 2017a) as the feature extractor with a VGG16 (Simonyan and Zisserman, 2014) backbone. The first row presents the source-trained model performance prior to adaptation to demonstrate effect of initial knowledge transfer from the source domain.

10.3389/fdata.2024.1359317

TABLE 2 Domain adaptation results for different methods for the GTA5→cityscapes task.

Method	road	sdwk	bldng	Wall	Fence	Pole	Light	Sign	vgttn	trrn	Sky	Person	Rider	Car	Truck	Bus	Train	mcycl	bcycl	mIoU
Source (VGG16)	25.9	10.9	50.5	3.3	12.2	25.4	28.6	13.0	78.3	7.3	63.9	52.1	7.9	66.3	5.2	7.8	0.9	13.7	0.7	24.9
FCNs Wld.	70.4	32.4	62.1	14.9	5.4	10.9	14.2	2.7	79.2	21.3	64.6	44.1	4.2	70.4	8.0	7.3	0.0	3.5	0.0	27.1
CDA	74.9	22.0	71.7	6.0	11.9	8.4	16.3	11.1	75.7	13.3	66.5	38.0	9.3	55.2	18.8	18.9	0.0	16.8	14.6	28.9
DCAN	82.3	26.7	77.4	23.7	20.5	20.4	30.3	15.9	80.9	25.4	69.5	52.6	11.1	79.6	24.9	21.2	1.30	17.0	6.70	36.2
SWD	91.0	35.7	78.0	21.6	21.7	31.8	30.2	25.2	80.2	23.9	74.1	53.1	15.8	79.3	22.1	26.5	1.5	17.2	30.4	39.9
CyCADA	85.2	37.2	76.5	21.8	15.0	23.8	22.9	21.5	80.5	31.3	60.7	50.5	9.0	76.9	17.1	28.2	4.5	9.8	0.0	35.4
ADVENT	86.9	28.7	78.7	28.5	25.2	17.1	20.3	10.9	80.0	26.4	70.2	47.1	8.4	81.5	26.0	17.2	18.9	11.7	1.6	36.1
AdaSegNet	86.5	36.0	79.9	23.4	23.3	23.9	35.2	14.8	83.4	33.3	75.6	58.5	27.6	73.7	32.5	35.4	3.9	30.1	28.1	42.4
TGCF-DA + SE	90.2	51.5	81.1	15.0	10.7	37.5	35.2	28.9	84.1	32.7	75.9	62.7	19.9	82.6	22.9	28.3	0.0	23.0	25.4	42.5
PCEDA	90.2	44.7	82.0	28.4	28.4	24.4	33.7	35.6	83.7	40.5	75.1	54.4	28.2	80.3	23.8	39.4	0.0	22.8	30.8	44.6
SFDA	81.8	35.4	82.3	21.6	20.2	25.3	17.8	4.7	80.7	24.6	80.4	50.5	9.2	78.4	26.3	19.8	11.1	6.7	4.3	35.86
GenAdapt	90.1	44.2	81.7	31.6	19.2	27.5	29.6	26.4	81.3	34.7	82.6	52.5	24.9	83.2	25.3	41.9	8.6	15.7	32.2	43.4
MAS^3-Joint	75.1	41.8	64.9	12.5	8.9	29.7	21.4	9.5	41.7	26.0	25.7	39.2	9.2	42.4	6.9	1.4	0.0	6.0	12.9	25.0
MAS^3	75.8	55.6	72.9	20.9	24.7	20.5	30.5	39.8	80.0	36.9	77.9	51.9	22.4	77.3	26.5	45.2	22.6	18.8	51.7	44.8

10.3389/fdata.2024.1359317

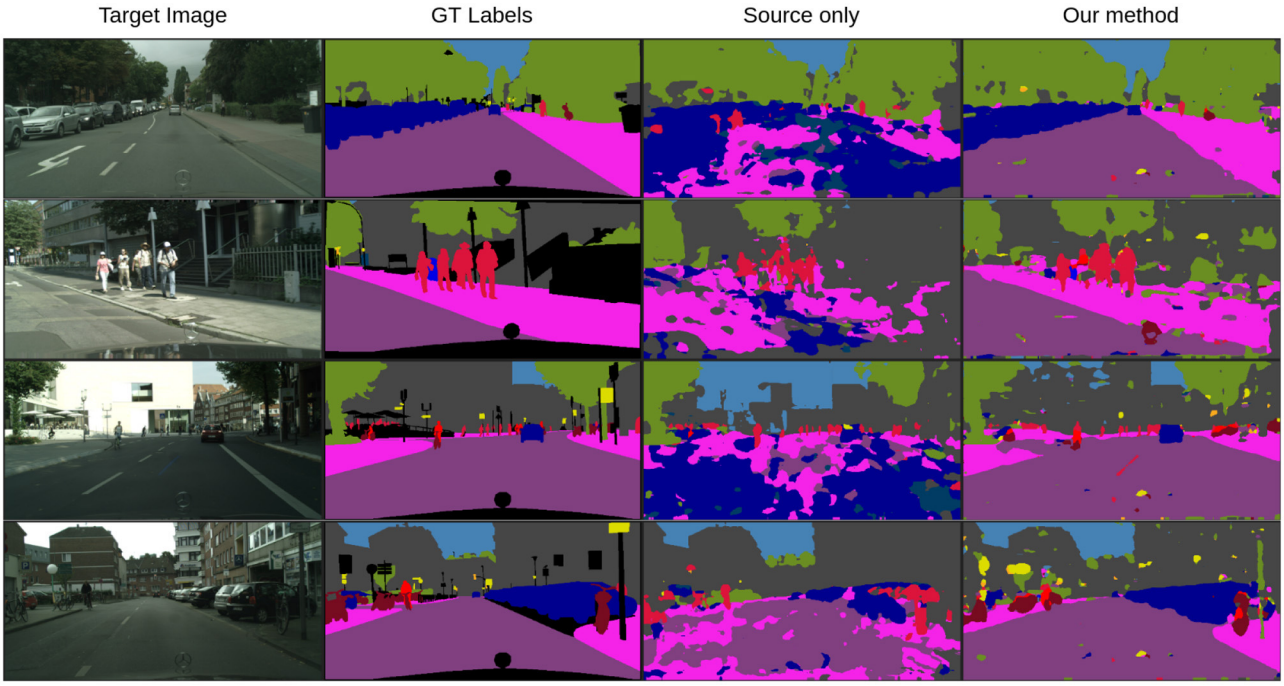

FIGURE 2
Qualitative performance: examples of the segmented frames for SYNTHIA→Cityscapes using the MAS^3 method. **(Left to right)** Real images, manually annotated images, source-trained model predictions, predictions based on our method.

6.4 Analytic experiment

We offer additional experiments to offer a better insight about our algorithm. In Figure 2, we have provided visualizations of representative frames from the Cityscapes dataset for the SYNTHIA→Cityscapes task. These frames are segmented using our model both prior to and after adaptation and are juxtaposed with the corresponding ground-truth manual annotations for each image. Through visual inspection, it becomes evident that our method brings about significant improvements in image segmentation, transitioning from source-only segmentation to post-adaptation segmentation. This improvement is particularly notable in semantic classes such as sidewalk, road, and cars for the model initially trained on GTA5 which are particularly important classes in autonomous driving applications. The visual comparison highlights the considerable enhancement achieved in performance. To further complement these findings, examples of segmented frames for the SYNTHIA→Cityscapes task are included in Figure 3, revealing similar observations. These visualizations collectively underscore the effectiveness of our method in enhancing image segmentation across diverse datasets.

We study the effect our algorithm on data distribution in the embedding space. To validate the alignment achieved by our solution, we employed the UMAP (McInnes et al., 2018) visualization tool to reduce the dimensionality of data representations in the embedding space to two for 2D visualization. Figure 4 visually represents samples from the internal distribution, along with the target domain data both before and after adaptation for the GTA5→Cityscapes task. In this figure, each point corresponds to a single data point, and each color represents a semantic class cluster. Upon comparing Figures 4B, C with Figure 4A, a noticeable observation emerges. The semantic classes in the target domain exhibit greater separation and similarity to the internal distribution after the model adaptation process. This signifies a substantial reduction in domain discrepancy facilitated by *MAS*^{3}, where the source and target domain distributions align indirectly through the intermediate internal distribution in the embedding space, as originally anticipated.

6.5 Sensitivity analysis experiments

An inherent advantage of our algorithm, in contrast to methods relying on adversarial learning, lies in its simplicity and dependence on only a few hyperparameters. We study the sensitivity of performance with respect to these hyperparameters. The primary hyperparameters specific to our algorithm are λ and τ constants. Through experimentation, we have observed that the performance of *MAS*^{3} remains stable with respect to the trade-off parameter λ. This stability is expected as the $\mathcal{L}_{ce}$ loss term remains relatively small from the outset due to prior training on the source domain, and then, optimization mostly reduces the cross-domain alignment loss term. We further delved into the impact of the confidence hyperparameter τ. Figure 5 visually illustrates the fitted Gaussian Mixture Model (GMM) on the source internal distribution for three different values of τ. Notably, when $\tau = 0$, the fitted GMM clusters appear cluttered. However, as we increment the threshold τ and selectively use samples for which the classifier demonstrates confidence, the fitted GMM represents well-separated semantic classes. This increase in interclass clusters in knowledge transfer

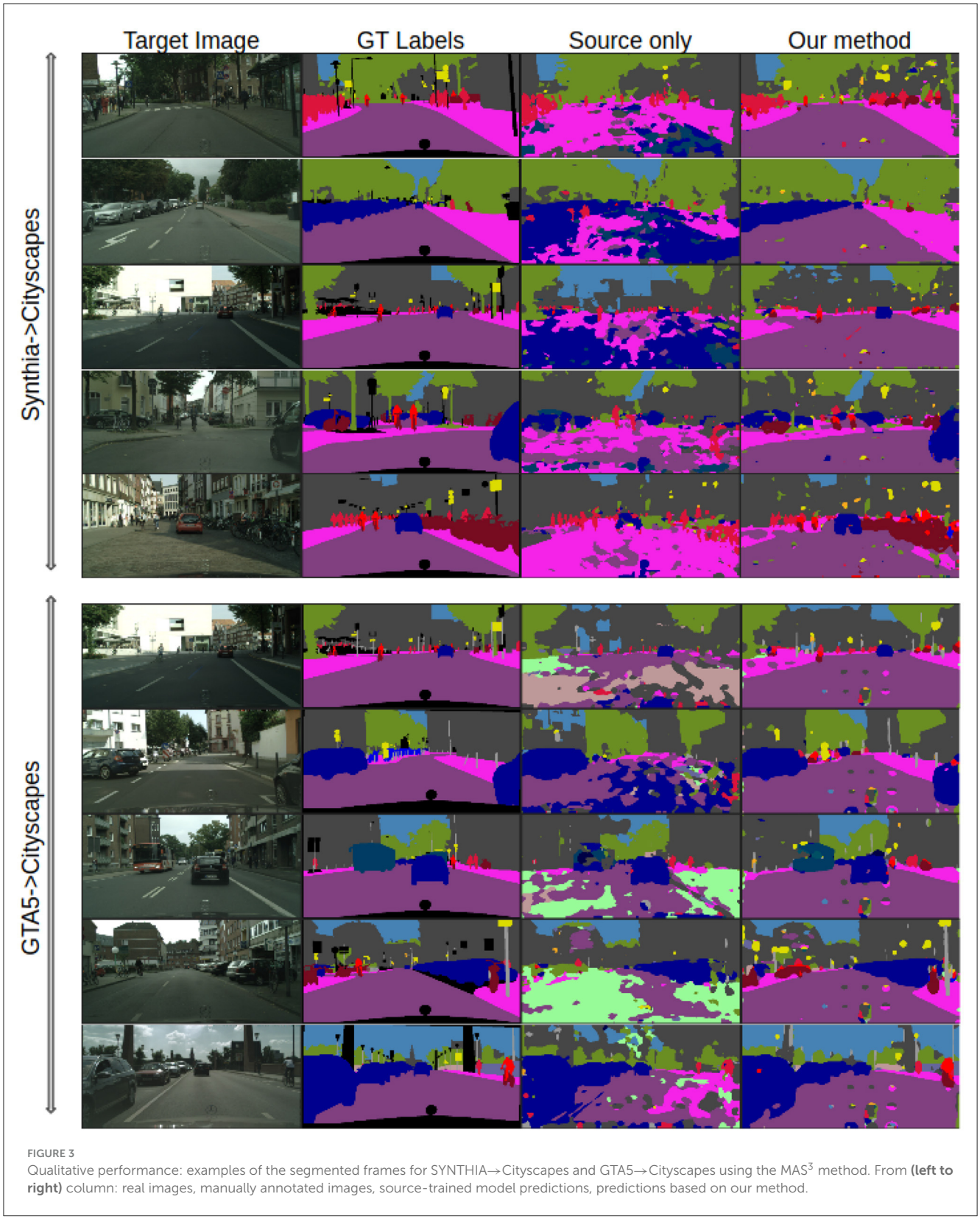

FIGURE 3
Qualitative performance: examples of the segmented frames for SYNTHIA→Cityscapes and GTA5→Cityscapes using the MAS^3 method. From **(left to right)** column: real images, manually annotated images, source-trained model predictions, predictions based on our method.

from the source domain is evident as semantic classes become more distinctly defined. This empirical exploration aligns with our earlier deduction regarding the significance of τ, as outlined in our Theorem, thereby validating the theoretical analysis. The experimental findings underscore the robustness and effectiveness of our method across different hyperparameter configurations.

We also extend our empirical investigation to analyze the balance between the quantity of projections employed for

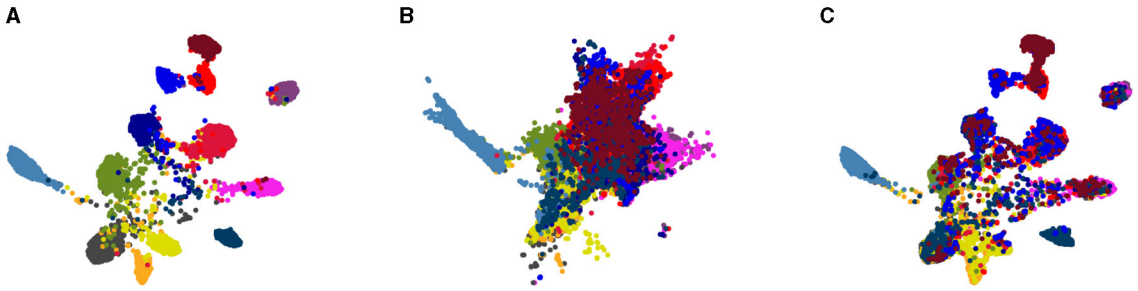

FIGURE 4
Indirect distribution matching in the embedding space: **(A)** drawn samples from the GMM trained on the SYNTHIA distribution, **(B)** representations of the Cityscapes validation samples prior to model adaptation, **(C)** representation of the Cityscapes validation samples after domain alignment.

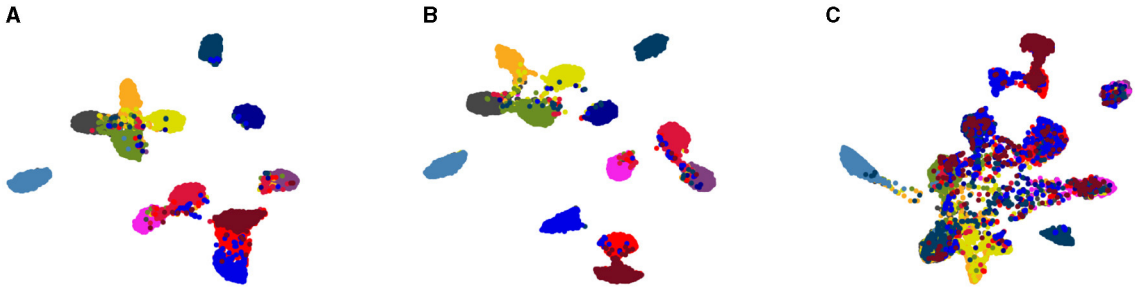

FIGURE 5
Ablation experiment to study effect of τ on the GMM learnt in the embedding space: **(A)** all samples are used; adaptation mIoU = 41.8, **(B)** a portion of samples is used; adaptation mIoU = 42.7, **(C)** samples with high model-confidence are used; adaptation mIoU = 44.7.

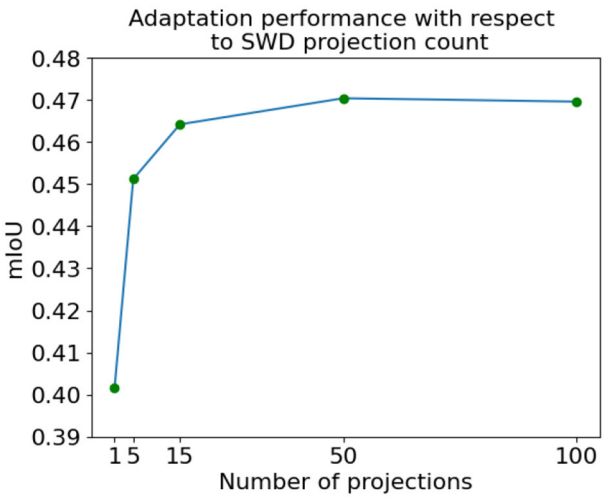

FIGURE 6
Performance accuracy vs. the number of projections used to compute SWD using the SYNTHIA→Cityscapes task.

computing SWD for distribution alignment and the resulting UDA performance on the target domain. The results for this experiment are presented in Figure 6. We observe that the performance remains robust even with a modest number of projections, such as 5. Moreover, empirical evidence indicates that the performance tends to plateau after ∼50 projections. It is noteworthy that the runtime complexity scales linearly concerning the number of projections, as the only operation involving the one-dimensional Wasserstein computations is the subsequent averaging process. In consideration of these findings, the results presented throughout the rest of the study are based on utilizing 100 projections. This choice is made to ensure that we operate within a favorable regime concerning adaptation performance while maintaining a balanced runtime to compute SWD. By doing so, we aim to achieve desirable performance with a decent computational load.

7 Conclusion

We devised an algorithm tailored for adapting an image segmentation model to achieve generalization across new domains, a process facilitated solely through the use of unlabeled data for the target domain during training. At the core of our approach is the utilization of an intermediate multi-modal internal distribution, strategically employed to minimize the distributional cross-domain discrepancy within a shared embedding space. To estimate this internal distribution, we employ a parametric Gaussian Mixture Model (GMM) distribution. Through rigorous experimentation on benchmark tasks, our algorithm has demonstrated its effectiveness, yielding competitive performance that stands out even when compared to existing UDA algorithms rooted in joint-domain model training strategies. The results underscore the robustness

and efficacy of our approach in achieving domain adaptation for image segmentation tasks, particularly in scenarios where only unlabeled data is available for training. Future exploration includes partial domain adaptation settings in which the source and the target domain do not share the same classes.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

SS: Data curation, Investigation, Methodology, Resources, Software, Validation, Visualization, Writing — original draft, Writing — review & editing. MR: Data curation, Investigation, Resources, Supervision, Writing — original draft, Writing — review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein GAN. *arXiv* [Preprint]. arXiv:1701.07875. doi: 10.48550/arXiv.1701.07875

Benjdira, B., Bazi, Y., Koubaa, A., and Ouni, K. (2019). Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sens.* 11:1369. doi: 10.3390/rs11111369

Bhushan Damodaran, B., Kellenberger, B., Flamary, R., Tuia, D., and Courty, N. (2018). "Deepjdot: deep joint distribution optimal transport for unsupervised domain adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Berlin: Springer), 447–463. doi: 10.1007/978-3-030-01225-0_28

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *Siam Rev.* 60, 223–311. doi: 10.1137/16M1080173

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2017). "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Honolulu, HI: IEEE), 3722–3731. doi: 10.1109/CVPR.2017.18

Chen, C., Dou, Q., Chen, H., and Heng, P.-A. (2018a). "Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest X-ray segmentation," in *International workshop on machine learning in medical imaging* (Cham: Springer), 43–151. doi: 10.1007/978-3-030-00919-9_17

Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017a). Rethinking atrous convolution for semantic image segmentation. *arXiv* [Preprint]. arXiv:1706.05587. doi: 10.48550/arXiv.1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018b). "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *ECCV* (Cham: Springer). doi: 10.1007/978-3-030-01234-2_49

Chen, Y., Li, W., Chen, X., and Gool, L. V. (2019). "Learning semantic segmentation from synthetic data: a geometrically guided input-output adaptation approach," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 1841–1850. doi: 10.1109/CVPR.2019.00194

Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Wang, Y.-C. F., Sun, M., et al. (2017b). No more discrimination: cross city adaptation of road scene segmenters. *arXiv* [Preprint]. arXiv:1704.08509. doi: 10.48550/arXiv.1704.08509

Choi, J., Kim, T., and Kim, C. (2019). "Self-ensembling with gan-based data augmentation for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Seoul: IEEE), 6830–6840. doi: 10.1109/ICCV.2019.00693

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Las Vegas, NV: IEEE), 3213–3223. doi: 10.1109/CVPR.2016.350

Dredze, M., and Crammer, K. (2008). "Online methods for multi-domain learning and adaptation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Honolulu: Association for Computational Linguistics), 689–697. doi: 10.3115/1613715.1613801

Feng, D., Haase-Schütz, C., Rosenbaum, L., Hertlein, H., Glaeser, C., Timm, F., et al. (2020). Deep multi-modal object detection and semantic segmentation for autonomous driving: datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* 22, 1341–1360. doi: 10.1109/TITS.2020.2972974

Gabourie, A. J., Rostami, M., Pope, P. E., Kolouri, S., and Kim, K. (2019). "Learning a domain-invariant embedding for unsupervised domain adaptation using class-conditioned distribution alignment," in *2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)* (Monticello, IL: IEEE), 352–359. doi: 10.1109/ALLERTON.2019.8919960

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA: MIT Press. Available online at: http://www.deeplearningbook.org (accessed February 06, 2024).

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Montreal, QC), 2672–2680.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative adversarial networks. *Commun. ACM* 63, 139–144. doi: 10.1145/3422622

Guan, H., and Liu, M. (2021). Domain adaptation for medical image analysis: a survey. *IEEE Trans. Biomed. Eng.* 69, 1173–1185. doi: 10.1109/TBME.2021.3117407

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv* [Preprint]. arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., et al. (2018a). "Cycada: cycle-consistent adversarial domain adaptation," in *International conference on machine learning* (Stockholm: PMLR), 1989–1998.

Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., et al. (2018b). "CyCADA: cycle-consistent adversarial domain adaptation," in *International Conference on Machine Learning* (Stockholm), 1989–1998.

Hoffman, J., Wang, D., Yu, F., and Darrell, T. (2016). FCNS in the wild: pixel-level adversarial and constraint-based adaptation. *arXiv* [Preprint]. arXiv:1612.02649. doi: 10.48550/arXiv.1612.02649

Hong, W., Wang, Z., Yang, M., and Yuan, J. (2018). "Conditional generative adversarial network for structured domain adaptation," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Salt Lake City, UT: IEEE), 1335–1344. doi: 10.1109/CVPR.2018.00145

Hung, W.-C., Tsai, Y.-H., Liou, Y.-T., Lin, Y.-Y., and Yang, M.-H. (2018). Adversarial learning for semi-supervised semantic segmentation. *arXiv* [Preprint]. arXiv:1802.07934. doi: 10.48550/arXiv.1802.07934

Jain, V., and Learned-Miller, E. (2011). "Online domain adaptation of a pre-trained cascade of classifiers," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (Colorado Springs, CO: IEEE), 577–584. doi: 10.1109/CVPR.2011.5995317

Kalake, L., Wan, W., and Hou, L. (2021). Analysis based on recent deep learning approaches applied in real-time multi-object tracking: a review. *IEEE Access* 9, 32650–32671. doi: 10.1109/ACCESS.2021.3060821

Kim, Y., Cho, D., Han, K., Panda, P., and Hong, S. (2021). Domain adaptation without source data. *IEEE Trans. Artif. Intell.* 2, 508–518. doi: 10.1109/TAI.2021.3110179

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. (2019). "Generalized sliced wasserstein distances," in *Advances in neural information processing systems* (Vancouver, CA), 32.

Kolouri, S., Zou, Y., and Rohde, G. K. (2016). "Sliced wasserstein kernels for probability distributions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 5258–5267. doi: 10.1109/CVPR.2016.568

Kundu, J. N., Kulkarni, A., Singh, A., Jampani, V., and Babu, R. V. (2021a). "Generalize then adapt: source-free domain adaptive semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 7046–7056. doi: 10.1109/ICCV48922.2021.00696

Kundu, J. N., Venkat, N. V., Rahul, M. V., and Babu, R. V. (2021b). "Universal source-free domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Montreal, QC).

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed. M. A. Arbib (Cambridge, MA: MIT Press), 3361.

Lee, C.-Y., Batra, T., Baig, M. H., and Ulbricht, D. (2019). "Sliced wasserstein discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 10285–10295. doi: 10.1109/CVPR.2019.01053

Li, M., Zhai, Y.-M., Luo, Y.-W., Ge, P.-F., and Ren, C.-X. (2020a). "Enhanced transport distance for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 13936–13944. doi: 10.1109/CVPR42600.2020.01395

Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., Duncan, J. S., et al. (2020b). Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: abide results. *Med. Image Anal.* 65:101765. doi: 10.1016/j.media.2020.101765

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., et al. (2017). "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI: IEEE). doi: 10.1109/CVPR.2017.106

Liu, Y., Zhang, W., and Wang, J. (2021). "Source-free domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville), 1215–1224. doi: 10.1109/CVPR46437.2021.00127

Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Boston, MA: IEEE), 3431–3440. doi: 10.1109/CVPR.2015.7298965

Luc, P., Couprie, C., Chintala, S., and Verbeek, J. (2016). "Semantic segmentation using adversarial networks," in *NIPS Workshop on Adversarial Training* (Barcelona).

McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018). UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* 3:861. doi: 10.21105/joss.00861

Murez, Z., Kolouri, S., Kriegman, D. J., Ramamoorthi, R., and Kim, K. (2017). "Image to image translation for domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEE). doi: 10.1109/CVPR.2018.00473

Redko, I., Habrard, A., and Sebban, M. (2017). "Theoretical analysis of domain adaptation with optimal transport," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Cham: Springer), 737–753. doi: 10.1007/978-3-319-71246-8_45

Richter, S. R., Vineet, V., Roth, S., and Koltun, V. (2016). "Playing for data: ground truth from computer games," in *European conference on computer vision* (Cham: Springer), 102–118. doi: 10.1007/978-3-319-46475-6_7

Romera, E., Bergasa, L. M., Yang, K., Alvarez, J. M., and Barea, R. (2019). "Bridging the day and night domain gap for semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)* (IEEE), 1312–1318. doi: 10.1109/IVS.2019.8813888

Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2015. MICCAI 2015* (Cham: Springer). doi: 10.1007/978-3-319-24574-4_28

Ros, G., Sellart, L., Materzynska, J., Vazquez, D., and Lopez, A. M. (2016). "The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition* (Las Vegas, NV: IEEE), 3234–3243. doi: 10.1109/CVPR.2016.352

Rostami, M. (2019). *Learning Transferable Knowledge Through Embedding Spaces* [PhD thesis]. University of Pennsylvania, Philadelphia, PA.

Rostami, M., and Galstyan, A. (2023). Overcoming concept shift in domain-aware settings through consolidated internal distributions. *Proc. AAAI Conf. Artif. Intell.* 37, 9623–9631. doi: 10.1609/aaai.v37i8.26151

Rostami, M., Kolouri, S., Eaton, E., and Kim, K. (2019). Deep transfer learning for few-shot sar image classification. *Remote Sens.* 11, 1374. doi: 10.3390/rs11111374

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018). "Maximum classifier discrepancy for unsupervised domain adaptation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3723–3732. doi: 10.1109/CVPR.2018.00392

Sakaridis, C., Dai, D., Hecker, S., and Van Gool, L. (2018). "Model adaptation with synthetic and real data for semantic dense foggy scene understanding," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: IEEE), 687–704. doi: 10.1007/978-3-030-01261-8_42

Saltori, C., Lathuilière, S., Sebe, N., Ricci, E., and Galasso, F. (2020). "SF-UDA 3D: source-free unsupervised domain adaptation for lidar-based 3D object detection," in *2020 International Conference on 3D Vision (3DV)* (Fukuoka: IEEE), 771–780. doi: 10.1109/3DV50981.2020.00087

Sankaranarayanan, S., Balaji, Y., Jain, A., Nam Lim, S., and Chellappa, R. (2018). "Learning from synthetic data: addressing domain shift for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 3752–3761. doi: 10.1109/CVPR.2018.00395

Simonyan, K., and Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition. arXiv* [Preprint]. arXiv:1409.1556. doi: 10.48550/arXiv.1409.1556

Solomon, J., De Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., et al. (2015). Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Trans. Graph.* 34:66. doi: 10.1145/2766963

Stan, S., and Rostami, M. (2021). Unsupervised model adaptation for continual semantic segmentation. *Proc. AAAI Conf. Artif. Intell.* 35, 2593–2601. doi: 10.1609/aaai.v35i3.16362

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. (2021). "Segmenter: transformer for semantic segmentation," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE). doi: 10.1109/ICCV48922.2021.00717

Sun, L., Wang, J., Yang, K., Wu, K., Zhou, X., Wang, K., et al. (2021). "Aerial-pass: panoramic annular scene segmentation in drone videos," in *2021 European Conference on Mobile Robots (ECMR)* (Bonn: IEEE), 1–6. doi: 10.1109/ECMR50962.2021.9568802

Tao, A., Sapra, K., and Catanzaro, B. (2020). Hierarchical multi-scale attention for semantic segmentation. *Arxiv*.

Tolstikhin, I., Bousquet, O., Gelly, S., and Schoelkopf, B. (2017). Wasserstein auto-encoders. *arXiv* [Preprint]. arXiv:1711.01558. doi: 10.48550/arXiv.1711.01558

Tsai, Y.-H., Hung, W.-C., Schulter, S., Sohn, K., Yang, M.-H., Chandraker, M., et al. (2018). "Learning to adapt structured output space for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT: IEEE), 7472–7481. doi: 10.1109/CVPR.2018.00780

Vu, T.-H., Jain, H., Bucher, M., Cord, M., and Pérez, P. (2019). "Advent: adversarial entropy minimization for domain adaptation in semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Long Beach, CA: IEEE), 2517–2526. doi: 10.1109/CVPR.2019.00262

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.-C., et al. (2020). "Axial-deeplab: stand-alone axial-attention for panoptic segmentation," in *European Conference on Computer Vision (ECCV)* (Cham: Springer). doi: 10.1007/978-3-030-58548-8_7

Wang, M., and Deng, W. (2018). Deep visual domain adaptation: a survey. *Neurocomputing* 312, 135–153. doi: 10.1016/j.neucom.2018.05.083

Wang, Q., Gao, J., and Li, X. (2019). Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Trans. Image Process.* 28, 4376–4386. doi: 10.1109/TIP.2019.2910667

Wei, Y., Xiao, H., Shi, H., Jie, Z., Feng, J., Huang, T. S., et al. (2018). "Revisiting dilated convolution: a simple approach for weakly- and semi-supervised semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Salt Lake City, UT: IEEE). doi: 10.1109/CVPR.2018.00759

Wilson, G., and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Trans. Intell. Syst. Technol.* 11, 1–46. doi: 10.1145/3400066

Wu, D. (2016). Online and offline domain adaptation for reducing bci calibration effort. *EEE Trans. Hum. Mach. Syst.* 47, 550–563. doi: 10.1109/THMS.2016.2608931

Wu, Z., Han, X., Lin, Y.-L., Gokhan Uzunbas, M., Goldstein, T., Nam Lim, S., et al. (2018). "Dcan: dual channel-wise alignment networks for unsupervised scene adaptation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer), 518–534. doi: 10.1007/978-3-030-01228-1_32

Xia, W., Wen, W., Wong, K.-K., Quek, T. Q., Zhang, J., Zhu, H., et al. (2021). Federated-learning-based client scheduling for low-latency wireless communications. *IEEE Wirel. Commun.* 28, 32–38. doi: 10.1109/MWC.001.2000252

Xu, R., Liu, P., Wang, L., Chen, C., and Wang, J. (2020). "Reliable weighted optimal transport for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 4394–4403. doi: 10.1109/CVPR42600.2020.00445

Yang, S., Wang, Y., van de Weijer, J., Herranz, L., and Jui, S. (2021). "Generalized source-free domain adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Montreal, QC: IEEE), 8978–8987. doi: 10.1109/ICCV48922.2021.00885

Yang, Y., Lao, D., Sundaramoorthi, G., and Soatto, S. (2020). "Phase consistent ecological domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 9011–9020. doi: 10.1109/CVPR42600.2020.00903

Yang, Y., and Soatto, S. (2020). "FDA: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Seattle, WA: IEEE), 4085–4095. doi: 10.1109/CVPR42600.2020.00414

You, F., Li, J., Zhu, L., Chen, Z., and Huang, Z. (2021). "Domain adaptive semantic segmentation without source data," in *Proceedings of the 29th ACM International Conference on Multimedia, MM '21* (New York, NY: Association for Computing Machinery), 3293–3302. doi: 10.1145/3474085.3475482

Yu, Q., Hashimoto, A., and Ushiku, Y. (2021). "Divergence optimization for noisy universal domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 2515–2524. doi: 10.1109/CVPR46437.2021.00254

Zhang, Q., Zhang, J., Liu, W., and Tao, D. (2019). "Category anchor-guided unsupervised domain adaptation for semantic segmentation," in *Advances in Neural Information Processing Systems* (Nashville), 435–445.

Zhang, Y., David, P., and Gong, B. (2017). "Curriculum domain adaptation for semantic segmentation of urban scenes," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 2020–2030. doi: 10.1109/ICCV.2017.223

Zhang, Z., Fidler, S., and Urtasun, R. (2016). "Instance-level segmentation for autonomous driving with deep densely connected MRFS," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 669–677. doi: 10.1109/CVPR.2016.79

Zhu, J., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV* (Venice: IEEE), 2223–2232. doi: 10.1109/ICCV.2017.244