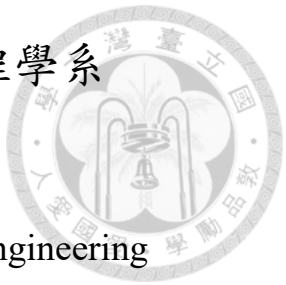


國立臺灣大學電機資訊學院資訊工程學系

學士班學生論文

Department of Computer Science and Information Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Bachelor's Thesis

透過目標整合的表徵學習來根除通用域適應中的隱性不匹配

Uprooting Implicit Misalignment in Universal Domain Adaptation by Target-Integrated Representation Learning

方泓傑

Hung-Chieh Fang

指導教授: 林軒田 博士

Advisor: Hsuan-Tien Lin Ph.D.

中華民國 113 年 8 月

August, 2024

國立臺灣大學學士班學生論文
口試委員會審定書

透過目標整合的表徵學習來根除通用域適應中的隱性不匹配
Uprooting Implicit Misalignment in Universal Domain Adaptation by Target-Integrated Representation Learning

本論文係方泓傑君（學號 B09902106）在國立臺灣大學資訊工程學系完成之學士班學生論文，於民國一百一十三年四月二十九日承下列考試委員審查通過及口試及格，特此證明

口試委員(3 位)：

eSigned by: 林軒田 05/01/2024 @ 00:38 UTC

_____（簽名）

（指導教授）

eSigned by: 李宏毅 04/30/2024 @ 16:43 UTC

eSigned by: 04/30/2024 @ 16:01 UTC

系 主 任 ： 陳祝嵩

摘要

通用域適應旨在解決源域和目標域之間的分佈偏移，且不假設它們的標籤集之間有任何關係所帶來的挑戰。在本文中，我們展示了域適應中的偏差可能出現在兩種情境中。我們將第一種情境稱為隱性偏差，這種偏差源於源偏置的特徵空間。第二種情境，我們稱之為顯性偏差，發生在特徵對齊方法錯誤地對齊了不正確的特徵，進一步加劇了偏差問題。我們進一步展示了通用域適應可能加劇這兩種類型的偏差。以往的研究主要通過設計改進的不確定性測量方法來解決顯性偏差問題，以降低特徵對齊中私有樣本的影響。然而，當隱性偏差較大時，這些方法往往表現不佳。為此，我們提出了一種無源目標融入模塊，該模塊利用自監督學習來緩解隱性偏差。與傳統方法不同，自監督學習促進了更廣泛的指導多樣性，降低了不正確特徵對齊的可能性。我們在 Office-Home 和 Office31 數據集上的不同設置下進行的實驗結果表明，我們的方法在處理大規模隱性偏差問題上更為有效。

關鍵字：域適應、通用域適應、機器學習、自監督學習

Abstract

Universal Domain Adaptation (UniDA) seeks to address the challenges posed by distribution shifts between source and target domains without assuming any relationships between their label sets. In this paper, we demonstrate that misalignment in Domain Adaptation (DA) can occur in two scenarios. The first scenario, which we term as *implicit misalignment*, arises from a source-biased feature space. The second scenario, termed *explicit misalignment*, occurs when feature alignment methods mistakenly align incorrect features, further exacerbating the misalignment issue. We further demonstrate that UniDA can intensify these two types of misalignment. Previous works have primarily addressed the *explicit misalignment* issue by designing improved methods for uncertainty measurement to down-weight the influence of private samples in feature alignment. However, these methods often fall short when implicit misalignment is substantial. To address this, we propose a source-label-free target incorporation module that leverages self-supervised learning to mitigate *implicit misalignment*. Unlike traditional methods, self-supervised

learning promotes a broader diversity of guidance, reducing the likelihood of incorrect feature alignments. Our experimental results on the Office-Home and Office31 dataset with different settings indicate that our approach is more effective at handling problems with large implicit misalignment.

Keywords: Domain Adaptation, Universal Domain Adaptation, Machine Learning, Self-supervised Learning

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Advancements in deep neural networks have heavily relied on the assumption that abundant data follows the same distribution, which make them hard to generalize well on unseen data with different distribution. Unsupervised Domain adaptation (UDA) addresses this issue by transferring knowledge from a source domain with known distributions to a target domain where distributions differ. However most DA methods [11, 17] have a strong assumption that two domains share the same label sets ($C_s = C_t$), limiting the applications in real world scenarios.

To overcome this, more flexible approaches such as Open-set DA ($C_s \subset C_t$) and Partial DA ($C_s \supset C_t$) have been developed [2, 13], addressing scenarios where label sets are not shared completely. The most practical setting to date is Universal DA (UniDA) [20], where there is no assumption about the relationship between the label sets of the source and target domains. In UniDA, the source and target domains may have some common classes, and each domain can has its own private classes. The goal of UniDA is to correctly classify the target-common classes, while detecting target-private classes as an unknown class.

Due to differences in the distribution of data between source and target domains, models trained on source data may exhibit a biased feature space. We refer to this phe-

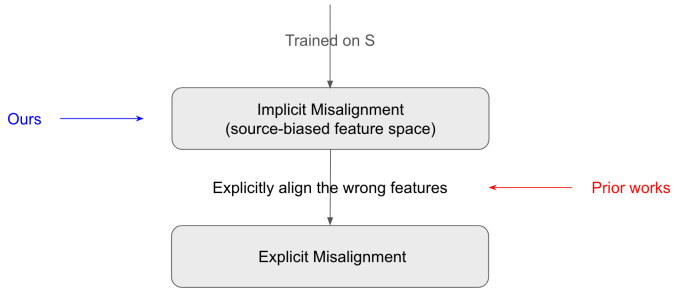


Figure 1.1: Illustration of our difference from prior works.

nomenon as *implicit misalignment*. To bridge the domain gap, numerous methods [5, 6, 10, 15, 20] have been developed to align features of the same class across domains more closely. However, due to unknown information about target labels, features can be wrongly pulled together. We refer to this issue as *explicit misalignment*. In UniDA, both types of misalignment can be intensified. We demonstrate that an increased number of source-private classes can exacerbate implicit misalignment. On the other hand, explicit misalignment is aggravated when applying feature alignment on the misaligned label space.

Existing UniDA methods often address the challenge of misaligned label space with partial alignment. Many prior works [5, 6, 10, 15, 20] utilize score functions (e.g. confidence, entropy) from the source classifier to differentiate between private and common samples. These metrics are then used to down-weight the influence of private samples during feature alignment. However, these approaches mainly mitigate the explicit misalignment problem, which are less capable when faced with substantial implicit misalignment, where the feature space is predominantly skewed towards the source data.

Inspired by this, our research aims to mitigate the implicit misalignment problem by making the feature space less source-biased. The illustration of our differences from prior methods is shown in figure 1.1. A central question we explore is how to effectively

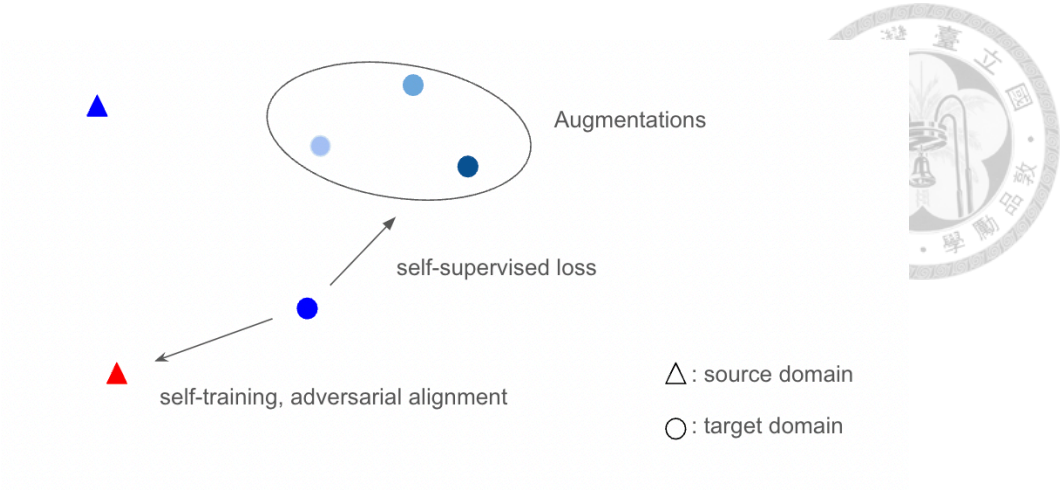


Figure 1.2: Intuition of our work.

integrate target data into the adaptation process. We propose that the integration of target data should meet two key criteria: (1) **Independence from the source data** to prevent source-biased alignment; (2) **Timely integration** to prevent premature entanglement of the feature space. Prior works utilize self-training or adversarial alignment to incorporate the target data, which fail to meet the criteria above. Therefore, we propose a source-label-free target incorporation module that leverages self-supervised learning. The intuition of work is that self-supervised learning pull apart the features with augmented features, which introduce more diversity in feature alignment directions.

In our analysis, we demonstrate that implicit misalignment tends to exacerbate as the number of source-private classes increases. Our experimental results on the Office-Home dataset reveal that our method significantly outperforms the baseline methods by a large margin when the number of source-private classes is large. Conversely, it performs comparably to other methods when the number of source-private classes is small.

The main contributions of this work are as follows:

1. We show that feature misalignment can be categorized into implicit and explicit misalignment. The former results from a source-biased feature space, while the

latter stems from imperfect feature alignment.

2. We propose to mitigate implicit misalignment with source-label-free target incorporation via self-supervised learning. This approach does not depend on source data and can be seamlessly integrated at an early stage.

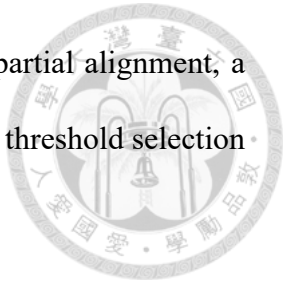
3. We analyze how self-supervised learning can mitigate implicit misalignment.

Chapter 2 Related Work

2.1 Universal Domain Adaptation.

UniDA is a more generalized form of DA that makes no assumptions about the label sets relationship between the source and target domains. Prior works [5, 6, 10, 15, 20] utilized score functions (e.g. confidence score, domain transferability) of the target examples from the trained classifier to obtain uncertainty measurements. These uncertainty measurements are crucial for tasks like partial alignment and private class discovery. In these contexts, examples that have high uncertainty are classified as private classes and are consequently assigned lower weights during the alignment process. In the private class discovery task, however, these methods use a manually-designed threshold to detect private classes, which is not applicable in real world scenario. To address this, several methods have been introduced. For instance, OVANet [16] and MLNet [12] introduce a binary one-versus-all classifier that distinguishes between common and private classes effectively. Meanwhile, UniOT [3] proposes using the statistical mean to dynamically calculate these thresholds, thereby adding a level of automation to the process. Moreover, MATHS [4] introduces an even more robust approach by first employing Hartigan's dip test to confirm the bimodality of the uncertainty scores. If bimodality is confirmed, it then applies a Gaussian mixture model to determine the optimal threshold. Previous research

has primarily concentrated on improving weighting mechanism for partial alignment, a process we refer to as mitigating *explicit misalignment* or refining the threshold selection strategy for enhanced discovery of private classes.



2.2 Self-supervised Learning in Domain Adaptation.

Self-supervised Learning (SSL) has been used in domain adaptation literature [1, 19]. However, prior works have not clearly explained the rationale for utilizing SSL in domain adaptation. STOC [8] has demonstrate that leveraging self-training (ST) on a feature space pretrained with contrastive learning (CL) can yield superior performance compared to using either method in isolation. Their findings indicate that a feature space developed through CL can more effectively capture invariant features, thereby reducing discrepancies. However, applying CL in the pretraining stage requires substantial data and computational cost, which limits its application in small datasets.

In contrast to prior works, we explore the misalignment issues through *implicit* and *explict misalignment*. Our research is the first work to show that *implicit misalignment* can be aggravated in UniDA. We further show that SSL can effectively address this problem.

Chapter 3 Methodology

In Universal Domain Adaptation (UniDA) [20], there is a labeled source domain $D_s = \{(x_i^s, y_i^s)\}_{i=1}^{n_s}$ and an unlabeled target domain $D_t = \{x_i^t\}_{i=1}^{n_t}$ are accessible at training time. The label sets are denoted as C_s for the source domain and C_t for the target domain, respectively. $C = C_s \cap C_t$ represents the common label set shared by both domains. Let $\overline{C}_s = C_s \setminus C$ and $\overline{C}_t = C_t \setminus C$ represent the source-private label set and the target-private label set, that a label only occurs at one of two domains. The goal of UniDA is to classify target examples into $|C| + 1$ classes, where target-private examples are regarded as one unknown class.

In Section 3.1, we first introduce the implicit and the explicit misalignment problems and revisit the limitations of existing UniDA methods. In following sections, we reveal the novel method for UniDA problem, which contains three main components: a *source-free target information integration* module (Section 3.2) and a *partial alignment* module (Section 3.3), for the training; a *private class discovery* module for the inference.

3.1 Misalignment in UniDA

Misalignment in DA. The naive approach is to minimize the classification loss in the source domain L_s, which is defined as follows:

$$L_S(f_\theta, c_\theta) = \frac{1}{|D_s|} \sum_{(x_i^s, y_i^s) \in D_s} \text{CE}(y_i^s, c_\theta(f_\theta(x_i^s))) \tag{3.1}$$

where the parameterized model consists of a feature extractor f_θ and a label classifier c_θ, $\text{CE}(y, x)$ is the cross-entropy loss function. However, due to the differences in source and target distributions, the feature extractor f_θ tends to be biased toward the source data, causing misalignment between the source and target data. We named this kind of misalignment as the *implicit misalignment*.

To address the gap between the source and target domains in DA, techniques like adversarial alignment and self-training is proposed to explicitly align the features of same classes closely. For example, DANN adopts the representation learning techniques to align the two feature spaces [7] via confusing a domain discriminator d_θ:

$$L_{\text{adv}}(f_\theta, d_\theta) = \frac{1}{|D_s|} \sum_{x_i^s \in D_s} \text{CE}(0, d_\theta(f_\theta(x_i^s))) + \frac{1}{|D_t|} \sum_{x_i^t \in D_t} \text{CE}(1, d_\theta(f_\theta(x_i^t))) \tag{3.2}$$

where $d_i = 0$ for $x_i^\bullet \in D_s$ and $d_i = 1$ for $x_i^\bullet \in D_t$. However, due to the lack of information about target labels, the alignment process may wrongly pull the wrong features. We refer to this issue as *explicit misalignment*.

Cross-Domain Alignment Score (CDAS). We introduce a new metric, the Cross-Domain Alignment Score (CDAS), to evaluate the degree of misalignment between domains. For each example (x_i^t, y_i^t) in the target domain, we identify the set $N'_{x_i^t}$ of k-nearest

neighbors in the source domain based on a pre-defined distance metric. The score is calculated as follows:

$$\text{CDAS} = \frac{1}{|D_t|} \sum_{(x_i^t, y_i^t) \in D_t} [\![\text{mode}\{y_j^s \mid (x_j^s, y_j^s) \in N'_{x_i^t}\} = y_i^t]\!] \tag{3.3}$$

Misalignment in UniDA. In Figure 3.1, we assess the performance of CDAS using a source classifier at an early stage across various numbers of source-private classes. We can observe that CDAS decreases with an increase in the number of source-private classes, indicating a worsening of implicit misalignment. On the other hand, feature alignment is challenging due to the misaligned label space in UniDA, exacerbating explicit misalignment. The explicit misalignment is addressed by some prior works [5, 6, 10, 15, 20] with partial alignment. UAN [20] initiated an idea of the partial alignment, which aims to align the common label set C between the source and target domains with well-designed weighting function for the source domain $w^s(x)$ and for the target domain $w^t(x)$. The weighting function represents the examples which are probably in the common label set C with higher weights than other examples in adversarial loss, and we can reformulate the loss function in (3.2):

$$L_{\text{adv}}^w(f_\theta, d_\theta) = \frac{1}{|D_s \cup D_t|} \sum_{x_i^\bullet \in D_s \cup D_t} w^\bullet(x_i^\bullet) \cdot \text{CE}(d_i, d_\theta(f_\theta(x_i^\bullet))), \tag{3.4}$$

where $d_i = 0$ for $x_i^\bullet \in D_s$ and $d_i = 1$ for $x_i^\bullet \in D_t$. Partial alignment has inspired several works to improve the design of the weighting function for better estimating examples that belong to the common label set. For example, [10] and CMU [6] proposed to combine different uncertainty measurements to compensate the limitations of single metric.

In the next section, we argue that existing methods fail to address the implicit mis-

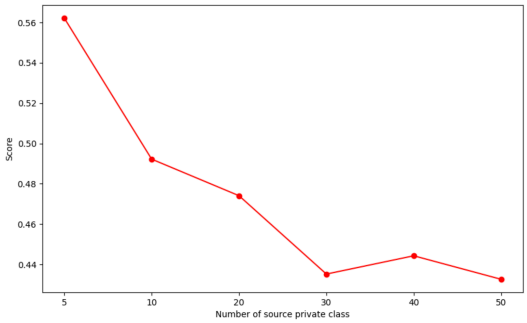


Figure 3.1: The Cross-Domain Alignment Score (CDAS) of a source classifier evaluated at an early stage with different number of source-private classes.

alignment.

3.2 Source-Label-Free Target Information Integration via SSL

To mitigate implicit misalignment, a central question we explore is how to effectively integrate target data into the adaptation process. Our core ideas of integration should meet the following:

1. **Independence from the source data**: The process should not use the source information to select or influence the integration of the target data, ensuring the target data is not influced by source-biased feature spcae.

2. **Timely integration**: Target data should be incorporated early in the training process. Delaying the integration risks entangling the feature space further, making it challenging to disentangle and adapt effectively later on.

Self-training (ST) and adversarial alignment (Adv) are two common techniques used to integrate target information. However, ST relies on the source classifier to generate pseudo-labels for the target examples, that directly uses the source information. On the

other hand, existing methods apply adversarial feature alignment after the feature extractor which has been influenced by the source data. Both methods fail to meet the independence and timely criterion for incorporating target information, that leads to the implicit misalignment.

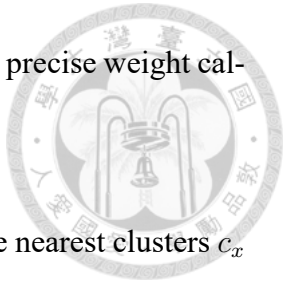
To avoid the implicit misalignment, we introduces a *source-label-free target information incorporation* module via self-supervised learning (SSL) techniques, that extracts representations across the source and target domains without any label information from the source domain. The loss function can be formulated as follows:

$$L_s(f_\theta, c_\theta) + \lambda_{\text{ssl}} \cdot L_{\text{SSL}}(f_\theta) \tag{3.5}$$

where, $L_{\text{SSL}}(f_\theta) = \frac{1}{|D_s \cup D_t|} \sum_{x_i \in D_s \cup D_t} \|f_\theta(\mathcal{T}(x_i)) - f_\theta(\mathcal{T}'(x_i))\|_2^2$ is the self-supervised learning loss for the feature extractor f_θ with random augmentation function $\mathcal{T}$. We provide an intuition of our method in Figure 1.2. Figure 1.2 demonstrates that the integrating target information with self-supervised learning loss would suppress the wrong alignment direction of an example by pulling apart the features with augmented features. In contrast, self-training and adversarial feature alignment may explicitly align features with closely related but incorrect features.

3.3 Partial Alignment

Leveraging our balanced feature space, we develop a partial alignment module for $w^\bullet(x_i^\bullet)$ in (3.4) to further promote alignment of common-class features. Diverging from the conventional practices that primarily use uncertainty measurements from classifiers, our method utilizes the relationships in feature distances to compute example-level weights.



Our approach capitalizes on the balanced feature space to enable more precise weight calculation.

Specifically, we compute the distance from each example x to the nearest clusters c_x of the alternate domain. We aggregate these distances for all examples, resulting in a bimodal distribution. This occurs because examples from common classes tend to be closer than those from private classes. We follow prior work [4] to fit the distance from each example to the nearest clusters of the alternate domain with a Gaussian mixture model, formulated as

$$\rho\mathcal{N}(\phi_1, \sigma_1^2) + (1-\rho)\mathcal{N}(\phi_2, \sigma_2^2), \tag{3.6}$$

where $\mathcal{N}(\phi, \sigma)$ is the Gaussian distribution, and $\phi_1 < \phi_2$ represents the mean distance of two groups. Ideally, group of examples in the common sets ϕ_1 are small than group of examples in the target private set ϕ_2.

To improve accuracy, we revised the threshold calculation by $\tau = \frac{\sigma_2\phi_1 + \sigma_1\phi_2}{\sigma_1 + \sigma_2}$. This modification addresses the limitations of the previously suggested threshold $\theta_1 + \sigma_1$ [4], which can't handle diverse UniDA settings.

To compute the weights for partial alignment, we assign a weight of 1 to examples where the distance to nearest clusters $d(x, c_x)$ is below the threshold, and a weight of 0 to all other examples.

$$w^{\bullet}(x_i^{\bullet}) = \begin{cases} 1, & \text{if } d(x, c_x) < \tau \\ 0, & \text{otherwise} \end{cases}$$

Our final loss function can be formulated as follows:

$$\min_{f_\theta, c_\theta, d_\theta} \Big[L_s(f_\theta, c_\theta) + \lambda_{\text{ssl}} \cdot L_{\text{SSL}}(f_\theta) - \lambda_{\text{adv}} \cdot L_{\text{adv}}^{w}(f_\theta, d_\theta) \Big]. \tag{3.7}$$

3.4 Private Class Discovery

To identify private examples in the target domain during inference, we adopt a method similar to the one described in Section 3.3. After computing the distance to nearest source clusters for all target examples, we fit a Gaussian mixture model and determine the threshold τ. Examples are then classified as target-private classes if their distance to the nearest cluster center $d(x, c_x)$ is larger than τ.

Chapter 4 Experiments

We sketch out experimental settings, including datasets, baselines and evaluation metrics in Section 4.1. We then demonstrate experimental results to compare our method with state of the art approaches tailored to various settings in Section 4.2. We further analyze the effectiveness of our proposed method in handling the implicit misalignment in Section 4.3.

4.1 Experimental Settings

Dataset. We report results on two different benchmark. Office31 [14] contains 31 classes and 3 domains: Amazon (A), DSLR (D), Webcam (W). Office-Home [18] is a more difficult dataset with 65 classes and 4 domains: Art (A), Product (Pr), Clipart (Cl), Realworld (Rw).

Setting. Previous studies have explored various domain adaptation (DA) settings and ratios across different datasets. However, these settings have not adequately revealed scenarios where source-private classes predominate across all classes. Therefore, we experiment on OfficeHome with 50 source-private classes, 10 common classes and 5 target-private classes to demonstrate the implicit misalignment brought by large number of source-private classes. Additionally, we conducted experiments on both Office-

Home and Office31 datasets using the standard settings (5 source-private, 10 common, and 50 target-private classes for Office-Home; 10 source-private, 10 common, and 11 target-private classes for Office31) to show the effectiveness in general cases.

Evaluation Metric. We use H-score [6] as our metric, which calculate the harmonic mean of accuracy on common classes and accuracy on unknown classes.

$$\text{H-score} = 2 \cdot \frac{a_c \cdot a_{\bar{c}^t}}{a_c + a_{\bar{c}^t}}$$

Baselines and Comparison with State-of-the-Arts. To conduct experiments in our new setting with extensive source-private classes, we evaluate several baselines for which open-source code is available on the OfficeHome dataset. These including UAN [20], CMU [6], DANCE [15] and UniOT [3]. For Office31, we further include TNT [5] and DCC [9].

4.2 Main Results

Table 4.1 presents the results of our method compared to previous works on Office-Home with large source-private classes across 12 domains adaptation scenarios. As shown in Table 4.1, our method significantly outperforms previous approaches in settings with a large number of source-private classes. Specifically, it surpasses UniOT by almost 30% in this setting, demonstrating superior performance in managing implicit misalignment issues.

In Table 4.2 and Table 4.3, we report results on general settings. As indicated in the tables, our method achieves performance comparable to previous works except for

Table 4.1: H-score(%) on **Office-Home** (50/10/5) .

	Office-Home (50/10/5)												
	Ar2Cl	Ar2Pr	Ar2Rw	Cl2Ar	Cl2Pr	Cl2Rw	Pr2Ar	Pr2Cl	Pr2Rw	Rw2Ar	Rw2Cl	Rw2Pr	Avg
UAN [20]	26.8	34.1	24.5	31.3	41.6	35.2	24.3	17.9	34.3	25.1	31.4	30.2	29.7
CMU [6]	46	45.7	51.2	37.4	25.7	43.8	41.6	39.5	59.7	53.1	40.2	58.7	45.2
DANCE [15]	34	55.5	**82.6**	43.4	44.2	60.1	34.4	20.8	61.2	**65.7**	33.6	61.7	44.6
UniOT [3]	27.2	32.3	26.6	28.4	29.9	23.2	31.4	29.3	23	35.9	34.3	35.3	29.7
Ours	**47.1**	**74.4**	76.8	**46.4**	**54.3**	**63.5**	**55.8**	**48.5**	**72.3**	57.7	**46.3**	**68.5**	59.3

UniOT [3]. The reason for this is that the implicit misalignment is inobvious in a small number of source-private classes as discussion in Figure 3.1; therefore, UniOT, which is designed to handle explicit misalignment, performs better in this setting. However, as demonstrated in Table 4.1, UniOT struggles in scenarios involving significant implicit misalignment. This highlights that our method is particularly effective in addressing implicit misalignment issues.

Table 4.2: H-score(%) on **Office-Home** (5/10/50) .

	Office-Home (5/10/50)												
	Ar2Cl	Ar2Pr	Ar2Rw	Cl2Ar	Cl2Pr	Cl2Rw	Pr2Ar	Pr2Cl	Pr2Rw	Rw2Ar	Rw2Cl	Rw2Pr	Avg
UAN [20]	51.6	51.7	54.3	61.7	57.6	61.9	50.4	47.6	61.5	62.9	52.6	65.2	56.6
CMU [6]	56	56.9	59.2	67	64.3	67.8	54.7	51.1	66.4	68.2	57.9	69.7	61.6
DANCE [15]	26.7	11.3	18	33.2	12.5	14.3	41.6	39.9	33.3	16.3	27.1	25.9	25
UniOT [3]	**67.3**	**80.5**	**86**	**73.51**	**77.3**	**84.3**	**75.5**	**63.3**	**86**	**77.8**	**65.4**	81.9	**76.6**
Ours	53.8	75.1	83.9	63.2	67	77.6	72.2	55.9	81.6	74.2	55.9	81.6	70.2

Table 4.3: H-score(%) on **Office** (10/10/11); Comparison with methods that primarily focus on partial alignment .

	Office(10/10/10)						
	A2D	A2W	D2A	D2W	W2A	W2D	Avg
UAN [20]	59.7	58.6	60.1	70.6	60.3	71.4	63.5
CMU [6]	68.1	67.3	71.4	79.3	72.2	80.4	73.1
DANCE [15]	72.6	62.4	63.3	76.3	5 7.4	82.8	66.6
DCC [9]	**88.5**	78.5	70.2	79.3	75.9	88.6	80.2
TNT [5]	85.7	80.4	83.8	92	79.1	91.2	85.4
UniOT [3]	87	**88.5**	**88.4**	**98.8**	**87.6**	96.6	**91.1**
Ours	85.8	83.5	84.7	96.4	84.2	**97.2**	88.6

Overall, our methods effectively address the issue of large implicit misalignment without adversely affecting general misalignment cases. Therefore, our techniques demonstrate overall better performance across a range of universal settings compared to previous approaches.

4.3 Discussion

In this section, we study our proposed method in handling the implicit misalignment. We first conduct an ablation study to investigate the effectiveness of each loss functions in our method. Then we analyze the effectiveness of self-supervised learning in handling implicit misalignment. Finally, we compare the performance of our method with other score functions in calculating partial alignment weights.

4.3.1 Ablation study of loss functions

In Table 4.4, we examine various combinations of our loss functions. We set the source classifier (SC) alone without any alignment as the baseline, and demonstrate that directly performing partial alignment (SC + Adv) within the source-biased feature space is suboptimal. By contrast, applying self-supervised learning (SC + SSL) helps mitigate the implicit misalignment and achieve improvements. Additionally, we show that combining SSL with partial alignment via adversarial loss (SC + SSL + Adv) can further enhance performance. Therefore, we argue that self-supervised learning is crucial in addressing implicit misalignment, and that partial alignment in a balanced feature space can further improve the performance of our method. In following sections, we will verify the benefits of SSL in handling implicit misalignment and of partial alignment in a balanced feature space.

Table 4.4: Ablation Study on Office-Home. SC: Source Classifier; Adv: Adversarial Network; SSL: Self-supervised Learning.

Method	H-score
SC	63.2
SC + Adv	60
SC + SSL	67.4
SC + SSL + Adv	72.3

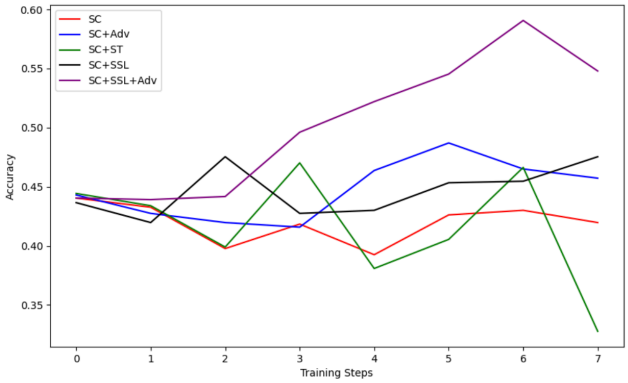


Figure 4.1: The Cross-Domain Alignment Score (CDAS) of different approaches. SC: Source Classifier; Adv: Adversarial Network; ST: Self-Training; SSL: Self-supervised Learning.

4.3.2 The effectiveness of SSL in handling implicit misalignment

As shown in Figure 4.1, it is evident that SC+SSL+Adv outperforms other methods that do not address implicit misalignment via observing the Cross-Domain Alignment Score (CDAS, 3.3) in training. The performance using self-training technique (SC+ST) is inferior to using SC alone, suggesting that aligning with a source-biased feature space may lead to severe negative transfer. Notably, the adversarial partial alignment used here incorporates ground truth information, i.e., we identify which examples belong to a common class. Surprisingly, SC+SSL performs comparably to SC+Adv (gold), which explicitly aligns the correct features. This suggests that SSL may serve as a form of noise-free partial alignment.

4.3.3 The effectiveness of calculating partial alignment weights with feature distances in a balanced feature space

Comparison with other score functions. We evaluate with AUROC score to verify discriminative capability of the score function in distinguishing between common and private classes. Previous works [5, 6, 10, 15, 20] mainly employ score functions from classifiers to calculate partial alignment weights for individual examples. In Figure 4.2 and Figure 4.3, we present a comparison of the performance achieved using feature distances against other prior methods. As illustrated in Figure 4.2, the score function derived from the source classifier fails to accurately differentiate between common and private examples within the source data, while the scores calculating with feature distances can separate well on both domains.

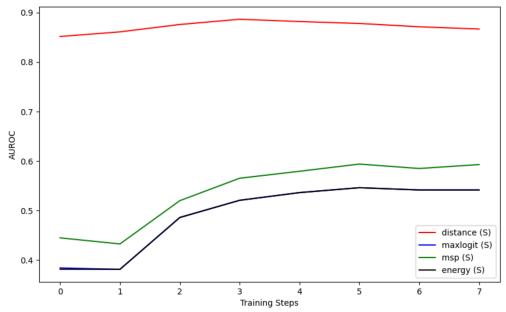


Figure 4.2: Comparison of AUC on source data with other score functions

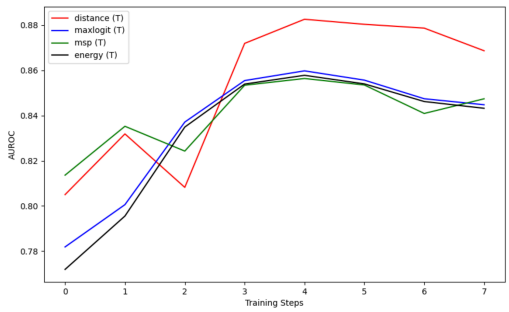


Figure 4.3: Comparison of AUC on target data with other score functions

Comparison with source-biased feature space. We also compare the feature space trained with self-supervised learning to the source-biased feature space. In Figure 4.4, we can observe that our method can better distinguish between common and private examples in both domains.

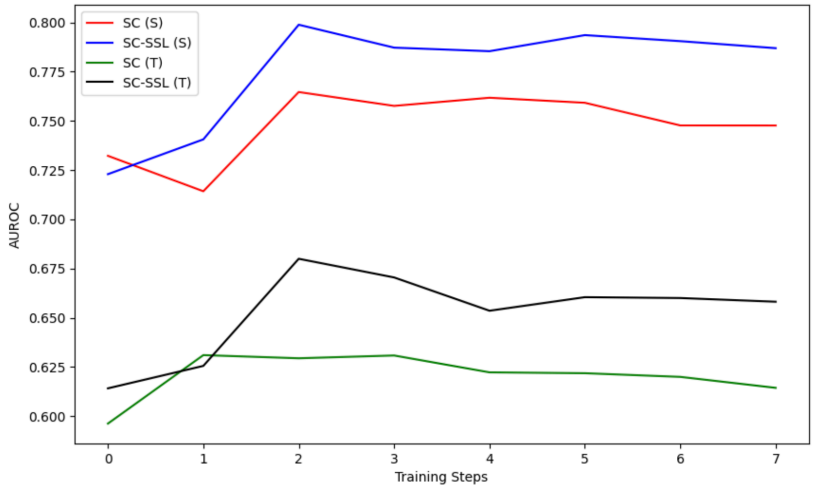


Figure 4.4: Comparison of AUC calculated with feature distance

Chapter 5 Conclusion

In this paper, we identify the *implicit* and *explicit misalignment* problems underlying in domain adaptation. Implicit misalignment arises from a feature space biased towards the source domain, whereas explicit misalignment occurs due to imperfect alignment of features. We explore how UniDA exacerbates these misalignment. While prior works have primarily focused on mitigating the explicit misalignment, we seek to mitigate implicit misalignment by debiasing the feature space. We incorporate self-supervised learning as auxiliary loss, which does no depend on source data and can be integrated at an early stage. Our experimental results show significant improvements over baseline methods in scenarios with severe implicit misalignment. Furthermore, we analyze the effectiveness of self-supervised learning in mitigating implicit misalignment.

References

[1] I. Achituve, H. Maron, and G. Chechik. Self-supervised learning for domain adaptation on point clouds. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 123–133, 2021.

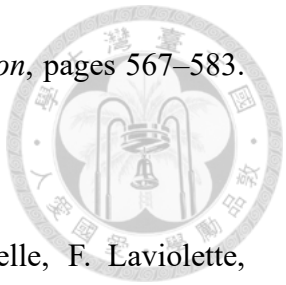
[2] Z. Cao, L. Ma, M. Long, and J. Wang. Partial adversarial domain adaptation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 135–150, 2018.

[3] W. Chang, Y. Shi, H. D. Tuan, and J. Wang. Unified optimal transport framework for universal domain adaptation. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[4] L. Chen, Q. Du, Y. Lou, J. He, T. Bai, and M. Deng. Mutual nearest neighbor contrast and hybrid prototype self-training for universal domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6248–6257, Jun. 2022.

[5] L. Chen, Y. Lou, J. He, T. Bai, and M. Deng. Evidential neighborhood contrastive learning for universal domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6258–6267, Jun. 2022.

[6] B. Fu, Z. Cao, M. Long, and J. Wang. Learning to detect open classes for universal



domain adaptation. In *European Conference on Computer Vision*, pages 567–583. Springer, 2020.

[7] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. March, and V. Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

[8] S. Garg, A. Setlur, Z. Lipton, S. Balakrishnan, V. Smith, and A. Raghunathan. Complementary benefits of contrastive learning and self-training under distribution shift. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] G. Li, G. Kang, Y. Zhu, Y. Wei, and Y. Yang. Domain consensus clustering for universal domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[10] O. Lifshitz and L. Wolf. Sample selection for universal domain adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8592–8600, May 2021.

[11] M. Long, Z. Cao, J. Wang, and M. I. Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018.

[12] Y. Lu, M. Shen, A. J. Ma, X. Xie, and J.-H. Lai. Mlnet: Mutual learning network with neighborhood invariance for universal domain adaptation. In *AAAI*, 2024.

[13] P. Panareda Busto and J. Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017.

[14] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to

new domains. In *European conference on computer vision*, pages 213–226. Springer, 2010.

[15] K. Saito, D. Kim, S. Sclaroff, and K. Saenko. Universal domain adaptation through self supervision. *Advances in neural information processing systems*, 33:16282–16292, 2020.

[16] K. Saito and K. Saenko. Ovanet: One-vs-all network for universal domain adaptation. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 9000–9009, 2021.

[17] A. Sicilia, X. Zhao, and S. J. Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve. *Machine Learning*, 112(7):2685–2721, 2023.

[18] H. Venkateswara, J. Eusebio, S. Chakraborty, and S. Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017.

[19] J. Xu, L. Xiao, and A. M. López. Self-supervised domain adaptation for computer vision tasks. *IEEE Access*, 7:156694–156706, 2019.

[20] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan. Universal domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.