DISS. ETH NO. 30025

TOWARDS PRACTICAL DOMAIN ADAPTATION FOR SCENE UNDERSTANDING

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

Rui Gong
Master of Science in Robotics, Systems and Control (ETH Zurich)

born on 26.08.1995

accepted on the recommendation of
Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Nicu Sebe, co-examiner
Prof. Dr. Ming-Hsuan Yang, co-examiner

2024

ABSTRACT

Scene understanding, which aims to understand visual scenes comprehensively, stands as a pivotal element within the field of computer vision. To empower machine with the human-like scene understanding ability, semantic segmentation emerges as a crucial tool, forming the essence of a broad range of applications, *e.g.*, autonomous driving, robot vision and human-computer interaction. Over the past decade, semantic segmentation models have achieved significant success, propelled by the availability of large-scale datasets and the rapid advancement of deep learning techniques. However, generalization of these models to new and different domains remains limited. Training domain-robust models typically relies on the labor-intensive process of labeling extensive and diverse datasets, resulting in significant costs and hindering the practical deployment of these models in real-world applications.

In such cases, domain adaptation aims at adapting the semantic segmentation model trained on the labeled source domain to the unlabeled target domain, thereby eliminating the need for labeling the target domain. Traditional domain adaptation typically relies on implicit or explicit assumptions, such as assuming a single data distribution for the source or target domain, or maintaining consistent taxonomies between them. However, these assumptions prove impractical in real-world applications. Moreover, prevailing domain adaptation frameworks depend on pseudo-labels assigned to the unlabeled target domain, introducing noise due to domain discrepancies. The presence of low-quality pseudo-labels inevitably impedes the adaptation process. To tackle these challenges, this dissertation introduces a set of domain adaptive semantic segmentation methods that tackle these challenges and close to practical scenarios, ultimately enhancing scene understanding. We propose four main contributions detailed below.

Firstly, we propose a multi-source domain adaptation and label unification (mDALU) problem along with a novel method to address it. In the mDALU setting, there exist multiple source domains and an unlabeled target domain, with only a subset of classes labeled in each source domain. The objective of mDALU is to develop a model encompassing all classes in the target domain. Our approach comprises

a two-stage adaptation process: a partially-supervised adaptation stage and a fully-supervised adaptation stage. In the partially-supervised stage, partial knowledge is transferred from multiple source domains to the target domain and integrated. In the fully-supervised stage, knowledge is transferred within a unified label space following a label completion process involving pseudo-labels.

Secondly, we present a principled meta-learning based approach to tackle open compound domain adaptation (OCDA) problem, wherein the target domain is considered as a compound of multiple unknown sub-domains. Our approach comprises four essential steps: cluster, split, fuse, and update. These steps establish a hyper-network to uncover and integrate the knowledge from the unknown sub-domains in the target domain. Additionally, we incorporate a meta-learning strategy for online model updates during testing, achieved with just a single-gradient step.

Thirdly, we propose a taxonomy adaptive cross-domain semantic segmentation (TACS) problem, addressing both image-level and label-level domain gaps. In particular, the label-level domain gap accommodates inconsistent taxonomies between the source and target domains (e.g., the ”person” class in the source domain being fine-grained as ”rider” and ”pedestrian” in the target domain). To tackle TACS comprehensively, we develop an approach that simultaneously handles image-level and label-level domain adaptation. At the label level, we utilize a bilateral mixed sampling strategy to augment the target domain and employ a relabelling method to harmonize and align the label spaces. To mitigate the image-level domain gap, we propose an uncertainty-rectified contrastive learning method, resulting in more domain-invariant and class-discriminative features.

Lastly, we introduce a framework based on implicit neural representations to enhance domain adaptation performance. In greater detail, the pseudo-label learning mechanism underlies the majority of domain-adaptive semantic segmentation methods. Our proposal involves estimating rectification values for predicted pseudo-labels using implicit neural representations, thereby enhancing the quality of pseudo-labels and facilitating the domain adaptation process.

In a nutshell, we demonstrate that our proposed problems and approaches transcend traditional domain adaptation limitations, enriching practical domain adaptation. This advancement facilitates robust scene understanding and application in real-world scenarios.

ZUSAMMENFASSUNG

Szenenverständnis, das darauf abzielt, visuelle Szenen umfassend zu verstehen, stellt ein entscheidendes Element im Bereich der Computer Vision dar. Um Maschinen mit der fähigkeit des menschenähnlichen Szenenverständnisses zu befähigen, erweist sich die semantische Segmentierung als ein entscheidendes Werkzeug, das den Kern einer Vielzahl von Anwendungen bildet, z. B. autonomes Fahren, Roboter Vision und Mensch-Computer-Interaktion. In den letzten zehn Jahren haben semantische Segmentierungsmodelle durch die Verfügbarkeit von umfangreichen Datensätzen und den raschen Fortschritt von Deep-Learning-Techniken erhebliche Erfolge erzielt. Die Generalisierung dieser Modelle auf neue und unterschiedliche Domänen bleibt jedoch begrenzt. Das Training von domänenrobusten Modellen beruht in der Regel auf dem arbeitsintensiven Prozess des Labelns umfangreicher und vielfältiger Datensätze, was zu erheblichen Kosten führt und die praktische Bereitstellung dieser Modelle in realen Anwendungen behindert.

In solchen Fällen zielt die Domänenanpassung darauf ab, das auf der gelabelten Quelldomäne trainierte semantische Segmentierungsmodell an die ungelabelte Zieldomäne anzupassen, wodurch die Notwendigkeit entfällt, die Zieldomäne zu labeln. Traditionelle Domänenanpassung beruht in der Regel auf impliziten oder expliziten Annahmen, wie etwa der Annahme einer einzigen Datenverteilung für die Quell- oder Zieldomäne oder der Beibehaltung konsistenter Taxonomien zwischen ihnen. Diese Annahmen erweisen sich jedoch als unpraktisch in realen Anwendungen. Darüber hinaus sind gängige Domänenanpassungsrahmen auf Pseudo-Labels angewiesen, die der ungelabelten Zieldomäne zugewiesen sind, was aufgrund von Domänendifferenzen zu Rauschen führt. Die Präsenz von qualitativ minderwertigen Pseudo-Labels behindert zwangsläufig den Anpassungsprozess. Um diese Herausforderungen zu bewältigen, führt diese Dissertation eine Reihe von domänenadaptiven semantischen Segmentierungsmethoden ein, die diese Herausforderungen bewältigen und nah an praktischen Szenarien liegen, um letztendlich das Szenenverständnis zu verbessern. Wir schlagen vier Hauptbeiträge vor.

Erstens schlagen wir ein Multi-Source-Domänenanpassungs- und Labelvereinigungsproblem (mDALU) zusammen mit einer neuartigen Methode zur Bewältigung vor. Im mDALU-Szenario gibt es mehrere Quelldomänen und eine ungelabelte Zieldomäne, wobei in jeder Quelldomäne nur eine Teilmengen von Klassen gelabelt sind. Das Ziel von mDALU ist es, ein Modell zu entwickeln, das alle Klassen in der Zieldomäne umfasst. Unser Ansatz umfasst einen zweistufigen Anpassungsprozess: eine teilweise überwachte Anpassungsstufe und eine vollständig überwachte Anpassungsstufe. In der teilweise überwachten Stufe wird Wissen von mehreren Quelldomänen auf die Zieldomäne übertragen und integriert. In der vollständig überwachten Stufe wird das Wissen in einem vereinheitlichten Labelraum nach einem Labelabschlussprozess unter Verwendung von Pseudo-Labels übertragen.

Zweitens präsentieren wir einen fundierten, auf Meta-Lernen basierenden Ansatz zur Bewältigung des Problems der offenen zusammengesetzten Domänenanpassung (OCDA), bei dem die Zieldomäne als Verbindung mehrerer unbekannter Subdomänen betrachtet wird. Unser Ansatz umfasst vier wesentliche Schritte: Clustern, Teilen, Fusionieren und Aktualisieren. Diese Schritte etablieren ein Hyper-Netzwerk, um das Wissen aus den unbekannten Subdomänen in der Zieldomäne aufzudecken und zu integrieren. Darüber hinaus integrieren wir eine Meta-Lernstrategie für Online-Modellaktualisierungen während des Tests, die mit nur einem einzelnen Gradientenschritt erreicht wird.

Drittens schlagen wir ein Problem der taxonomieadaptiven, domänen -übergreifenden semantischen Segmentierung (TACS) vor, das sowohl die Bild- als auch die Label-Ebene der Domänendifferenz bewältigt. Insbesondere berücksichtigt die Label-Ebene der Domänendifferenz inkonsistente Taxonomien zwischen den Quell- und Zieldomänen (z. B. die Klasse ”Person” in der Quelldomäne wird in der Zieldomäne feinkörnig als ”Reiter” und ”Fußgänger” betrachtet). Um TACS umfassend zu bewältigen, entwickeln wir einen Ansatz, der gleichzeitig die Bild- und Label-Ebene der Domänenanpassung handhabt. Auf der Label-Ebene verwenden wir eine bilaterale Mixed-Sampling-Strategie, um die Zieldomäne zu erweitern, und setzen eine Neubelegungsmethode ein, um die Labelräume zu harmonisieren und abzustimmen. Um die Bild-Ebene der Domänendifferenz zu mildern, schlagen wir eine Unsicherheits-berichtigte kontrastive Lernmethode vor, die zu domäneninvarianten und klassenunterscheidenden Merkmalen führt.

Schließlich stellen wir ein auf impliziten neuronalen Repräsentationen basierendes Framework zur Verbesserung der Domänenanpassungsleist -ung vor. Genauer gesagt liegt dem Großteil der domänenadaptiven semantischen Segmentierungsmethoden der Lernmechanismus für Pseudo-Labels zugrunde. Unser Vorschlag beinhaltet die Schätzung von Rektifikationswerten für vorhergesagte Pseudo-Labels unter Verwendung impliziter neuronaler Repräsentationen, wodurch die Qualität der Pseudo-Labels verbessert und der Domänenanpassungsprozess erleichtert wird.

Kurz gesagt zeigen wir, dass unsere vorgeschlagenen Probleme und Ansätze die traditionellen Einschränkungen der Domänenanpassung überwinden und die praktische Domänenanpassung bereichern. Diese Weiterentwicklung erleichtert ein robustes Szenenverständnis und die Anwendung in realen Szenarien.

PUBLICATIONS

The following publications are included in parts or in an extended version in this thesis:

- Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, and Luc Van Gool. „mDALU: Multi-source domain adaptation and label unification with partial datasets.“ In: *ICCV*. 2021

- Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. „Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation.“ In: *CVPR*. 2021

- Rui Gong, Martin Danelljan, Dengxin Dai, Danda Pani Paudel, Ajad Chhatkuli, Fisher Yu, and Luc Van Gool. „TACS: Taxonomy adaptive cross-domain semantic segmentation.“ In: *ECCV*. 2022

- Rui Gong, Qin Wang, Martin Danelljan, Dengxin Dai, and Luc Van Gool. „Continuous Pseudo-Label Rectified Domain Adaptive Semantic Segmentation With Implicit Neural Representations.“ In: *CVPR*. 2023

Furthermore, the following publications were part of my PhD research, are however not covered in this thesis. The topics of these publications are outside of the scope of the material covered here:

- Rui Gong, Wen Li, Yuhua Chen, Dengxin Dai, and Luc Van Gool. „Dlow: Domain flow and applications.“ In: *International Journal of Computer Vision* 129.10 (2021), pp. 2865–2888

- Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, Danda Pani Paudel, and Luc Van Gool. „Analogical image translation for fog generation.“ In: *AAAI Conference on Artificial Intelligence*. 2021

- Edward Gunther, Rui Gong, and Luc Van Gool. „Style Adaptive Semantic Image Editing with Transformers.“ In: *ECCV Workshops*. 2022

- Han Sun, Rui Gong, Konrad Schindler, and Luc Van Gool. „SF-FSDA: Source-Free Few-Shot Domain Adaptive Object Detection with Efficient Labeled Data Factory.“ In: *Conference on Lifelong Learning Agents (CoLLAs)*. 2023
- Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. „Prompting diffusion representations for cross-domain semantic segmentation.“ In: *arXiv preprint arXiv:2307.02138* (2023) (In Submission)

ACKNOWLEDGMENTS

Throughout my journey in pursuing a PhD, I have been fortunate to receive invaluable assistance and support from numerous brilliant individuals. Their contributions have played a pivotal role in shaping this thesis, and here I extend my heartfelt gratitude to each of them for their unwavering support and guidance.

First and foremost, I wish to convey my deepest gratitude to Prof. Dr. Luc Van Gool, my esteemed supervisor, whose unwavering support and guidance have been the cornerstone of my doctoral journey over the past four fulfilling years. Your role as an extraordinary advisor has consistently provided me with a wellspring of knowledge and inspiration, shaping my academic journey. My venture into computer vision commenced during my Master's semester project under your guidance, progressing through my master's thesis, and until now in my Ph.D. thesis. Working in various capacities under your mentorship at the Computer Vision Lab has been both an honor and a privilege.

I extend my sincere appreciation to Prof. Dr. Nicu Sebe and Prof. Dr. Ming-Hsuan Yang for generously agreeing to be my co-examiners. Your dedication to thoroughly reviewing my thesis and providing valuable insights is truly appreciated and holds immense value for me. I feel privileged to have these esteemed scholars in my PhD defense committee.

While at the Computer Vision Lab, I had the honor of collaborating with a team of committed and talented colleagues. My heartfelt gratitude goes to my co-authors, Dengxin Dai, Martin Danelljan, Wen Li, Qin Wang, Yuhua Chen, Danda Pani Paudel, Ajad Chhatkuli, Yawei Li, Han Sun, Edward Gunther, for engaging discussions and fruitful collaborations. I also thank David Brueggemann, Goutam Bhat, Prune Truong, Tim Broedermann, Mengya Liu, Guolei Sun, Jiezhang Cao, Ce Liu, Ardhendu Shekhar Tripathi, Suman Saha, Lukas Hoyer, Jingyun Liang, Lei Sun, Yun Liu, Yulun Zhang, Hao Tang, Christos Sakaridis, Kai Zhang, Zhejun Zhang, Dengping Fan, Wenguan Wang, Hanqing Wang and other colleagues at Computer Vision Lab for the enjoyable moments and friendship.

I would also like to thank Julio Delgado Mangas and Sarah Amsellem for their mentorship and guidance during my internship at Meta. Your support has been invaluable, and I greatly appreciate the insights and knowledge gained under your leadership.

Finally, I wish to express my sincere appreciation and profound gratitude to my parents and my beloved wife, Kuangqi Yang. Your unwavering understanding, encouragement, and patience have been my pillars of strength throughout my Ph.D. journey.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1

INTRODUCTION

Human scene understanding is a remarkable cognitive ability, encompassing the recognition of objects, the understanding of contextual information, and grasping the overall understanding of a visual scene. To imbue machines with human-like scene understanding capabilities, semantic segmentation stands out as a crucial tool, representing a foundational challenge in computer vision. Semantic segmentation, which aims to assign semantic labels to each pixel in an image, serves as the cornerstone for a diverse array of applications such as autonomous driving, robot vision, human-computer interaction as well as virtual and augmented reality. For instance, through the recognition of semantic information at the pixel level, intelligent robots not only expand their object recognition capabilities but also infer spatial relationships. Such abilities lay the groundwork for robots to develop a comprehensive understanding of the visual world, enabling them to foresee events, make informed decisions, and navigate in complex environments.

In the past decade, given the triumphs of deep learning techniques [68, 165, 40] and large-scale datasets [38, 44, 222, 34], semantic segmentation has made significant strides, exemplified by deep learning based semantic segmentation architectures like FCN [113], DeepLab [20], SegFormer [196] and Mask2Former [33]. Notwithstanding the tremendous success, the training of these deep models primarily relies on extensive labeled datasets on a large scale. However, two main issues, which hinder the practical applications, arise from such a reliance. 1) Manually annotating dense labels for large-scale datasets in real-world scenarios is challenging, incurring significant costs in both time and resources. For example, annotating a single image from Cityscapes [34] requires 1.5 hours, and the process extends to 3.3 hours for more challenging adverse weather conditions image in ACDC [160]. 2) The generalization ability of trained deep models to novel and diverse domains remains constrained, particularly in the presence of domain shift. When the data distribution (*i.e.*, target domain) diverges from the training data (*i.e.*, source domain) distribution, the model's performance typically experiences a significant drop [175, 227, 48, 116]. Consequently, in real-

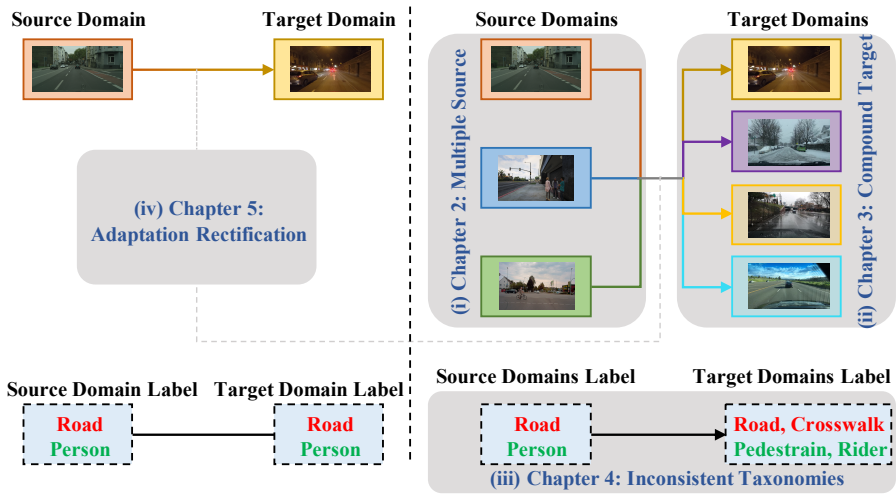
world applications encountering new scenarios and domains, there is a consistent need to collect and label new data for fine-tuning pretrained or re-training deep models. This process poses a considerable challenge due to the labeling effort and associated costs.

To circumvent the issues, domain adaptation emerges as a strategic approach [153, 184, 30, 72, 154]. The purpose of domain adaptation is to enhance the testing performance of deep models in the target domain by training on both labeled source domain and unlabeled target domain data, where the distributions of source and target domain data are different. This process involves adapting the model from the labeled source domain to the unlabeled target domain, presenting an effective mechanism to alleviate the labeling effort. For example, imagine the source domain as street view scene images captured in clear sunny weather, while the target domain consists of images depicting similar street view scene in rainy weather. Adapting from clear sunny weather to rainy weather significantly reduces the effort required for annotating pixel-level semantic labels of rainy weather images, a challenging task even for human annotation due to raindrop blurring and streaks.

A range of research endeavors has demonstrated the benefits of domain adaptation on advancing scene understanding [175, 174, 75]. However, these works typically operate under two primary assumptions, either explicitly or implicitly. 1) Firstly, the setup assumes a single source and a single target domain. 2) Secondly, it presupposes compatibility between the semantic classes in the source domain and those in the target domain; in other words, each class in the source domain can be unambiguously mapped to the corresponding class in the target domain, *i.e.*, consistent taxonomy. This naturally raises the question: *Is the domain adaptation setup comprehensive and practical?* 1) Firstly, in practical scenarios, labeled source domain data may originate from diverse sources, including but not limited to different modalities [50, 12], various scenes [128, 181], and distinct classes [108, 14, 92]. For instance, when training a semantic segmentation model for autonomous driving, one may encounter different public datasets with varying modalities (such as images and LiDAR points) from different cities, all of which collectively contribute to the source domain. Furthermore, when deploying the model in real applications, it must perform under different conditions, including various times of day and weather conditions (such as night, rainy, snowy, and foggy), which are all regarded as the target domain. Therefore, the single-source-single-

target assumption is not practical in such multi-source or multi-target scenarios. 2) Secondly, in many applications, the label spaces of the source and target domains often exhibit inconsistencies, stemming from different scenarios or requirements, inconsistent annotation practices, or the strive towards an increasingly fine-grained taxonomy. For example, in the Cityscapes [34] dataset, the "road" class is further delineated into "road" and "crosswalk" in the Mapillary [128] dataset. This finer distinction aids in achieving a more detailed comprehension of street scenes, particularly beneficial for applications in autonomous driving. When considering Cityscapes as the source domain and Mapillary as the target domain, the assumption of compatible source and target classes is not valid. Consequently, the two typical assumptions of traditional domain adaptation impede their applicability in real-world scenarios. Beyond the impractical assumptions considerations, an additional aspect to address from the approach mechanism side is: *Does the domain adaptation approach effectively and reliably transfer knowledge?* Pseudo-labeling or self-training [227, 226, 212, 174, 75, 76] has recently emerged as a straightforward yet powerful approach for domain adaptation to transfer knowledge. In pseudo-labeling methods, pseudo-labels are initially generated for the unlabeled target domain using the current model. The model is then iteratively fine-tuned with these target pseudo-labels. However, due to domain shift, some pseudo-labels are inevitably incorrect. These low-quality pseudo-labels pose a hindrance to the effective transfer of knowledge and limit the potential applications of domain adaptation.

This dissertation is committed to further relaxing the impractical assumptions and overcoming limitations associated with traditional domain adaptation, thereby enhancing scene understanding for real-world applications. In practical scenarios, it is common for the knowledge of deep models to come from various sources, such as different datasets, and these models need to be deployed into diverse scenarios. Additionally, the ground truth class labeling for the same object or pixel may change over time and with evolving requirements. Guided by these observations and insights, the primary methodological motivation behind the proposed approaches in this dissertation is the consideration of practical scenarios, where single-source-single-target and compatibility between source and target classes assumptions does not hold. Additionally, on the approach limitation side, existing domain adaptation frameworks rely on pseudo-labels to transfer knowledge between



(a) Traditional Domain Adaptation (b) Ours

Figure 1.1: Overview: Traditional domain adaptation *vs.* our practical domain adaptation.

different domains. Building on this, the dissertation introduces the pseudo-label rectification method to enhance the quality of pseudo-labels, thereby facilitating more effective domain adaptation. More specifically, this dissertation presents four deep learning based frameworks, striving forward more effective domain adaptation in practical scenarios, *i.e.*, practical domain adaptation. 1) Firstly, we delve into the transfer of knowledge from multiple source domains to the target domain, moving beyond the single-source assumption (see Chapter 2). 2) Secondly, relaxing the single-target assumption, we devise a method to adapt the model to multiple unknown target domains (see Chapter 3). 3) Thirdly, we investigate adapting the model in scenarios where there is incompatibility between source and target domain classes, *i.e.*, inconsistent taxonomy (see Chapter 4). 4) Fourthly, we propose a principled and plug-in module to rectify and enhance knowledge transfer across different domains (see Chapter 5.3.3).

- As our first contribution, we propose a new problem termed multi-source domain adaptation and label unification (mDALU). In this context, multiple source domains coexist with an unlabeled target domain. Within each source domain, only samples (*e.g.*, pixels or LiDAR points) belonging to a subset of classes are

labeled, leaving the rest unlabeled. The objective of mDALU is to derive a scene understanding model that encompasses all classes in the target domain. To address mDALU, we introduce a novel method, comprising a partially-supervised adaptation stage and a fully-supervised adaptation stage. In the former, we facilitate the transfer of partial knowledge from multiple source domains to the target domain, integrating it at the target domain. To avoid negative transfer issues arising from unmatched label spaces, we introduce three novel modules: domain attention, uncertainty maximization, and attention-guided adversarial alignment. In the latter stage, knowledge transfer occurs within the unified label space after a label completion process involving pseudo-labels. Through extensive experiments spanning diverse tasks – 2D semantic image segmentation, and joint 2D-3D semantic segmentation – our method consistently outperforms all competing approaches, demonstrating significant improvements.

- As our second contribution, we propose a principled meta-learning based approach, known as MOCDA, designed for addressing the open compound domain adaptation (OCDA) problem. In OCDA, target domain is modeled as a compound of multiple unknown homogeneous domains, offering the advantage of enhanced generalization to previously unseen domains. Our proposed MOCDA method continuously models the unlabeled target domain through four key steps. First, we *cluster* target domain into multiple sub-target domains by image styles, extracted in an unsupervised manner. Then, different sub-target domains are *split* into independent branches, for which batch normalization parameters are learnt to treat them independently. A meta-learner is thereafter deployed to learn to *fuse* sub-target domain-specific predictions, conditioned upon the style code. Concurrently, we employ model-agnostic meta-learning (MAML) algorithm for online model *update*, thus to further improve generalization. More specifically, four steps involved in our MOCDA are realized as follows. (i) Style codes are extracted from target images and grouped into multiple clusters. (ii) For each cluster, a set of batch normalization (BN) parameters is learned. (iii) Each image can have different domain-specific predictions corresponding to its cluster. The hypernetwork is trained to integrate these predictions. (iv) MAML is employed during the hypertraining process,

providing the model with the capability for online updates in an open domain during the inference stage. We demonstrate the advantages of our approach through extensive experiments on synthetic-to-real knowledge transfer benchmarks. Our method attains state-of-the-art performance in both compound and open domains.

- As our third contribution, we introduce a taxonomy adaptive cross-domain semantic segmentation (TACS) problem, accommodating inconsistent taxonomies between the two domains. I.e., there can exist one-to-many mapping from the source domain to the target domain classes. To address TACS, we propose a novel approach, which addresses both image-level and label-level domain adaptation. At the label level, we employ a bilateral mixed sampling strategy to augment the target domain, and introduce both a stochastic label mapping strategy and a pseudo-label based relabelling method to unify and align the label spaces. To tackle the image-level domain gap, we propose an uncertainty-rectified contrastive learning method, resulting in more domain-invariant and class-discriminative features. We extensively evaluate the efficacy of our framework in various TACS settings, including open taxonomy, coarse-to-fine taxonomy, and implicitly-overlapping taxonomy. Our proposed approach significantly outperforms the previous state-of-the-art, demonstrating superior adaptability to target taxonomies.

- As our fourth contribution, we propose a continuous rectification-aware mixture model (RMM), which rectifies and enhances the knowledge transfer of domain adaptation. While existing domain adaptation approaches have made significant strides leveraging pseudo-labels on unlabeled target-domain images, the presence of low-quality pseudo-labels due to domain discrepancies poses a hindrance to effective adaptation. This underscores the need for accurate and effective methods to estimate the reliability of pseudo-labels, aiming to rectify them. We propose to estimate rectification values for predicted pseudo-labels using implicit neural representations. We view the rectification value as a signal defined across the continuous spatial domain. By taking image coordinates and nearby deep features as inputs, the rectification value at a given coordinate is predicted as an output. This approach

enables high-resolution and detailed estimation of rectification values, crucial for accurate pseudo-label generation, especially at mask boundaries. The rectified pseudo-labels are then incorporated into our RMM, designed to be learned end-to-end, enhancing the adaptation process. We showcase the effectiveness of our approach across various domain adaptation benchmarks, including synthetic-to-real and day-to-night scenarios. Our method consistently outperforms state-of-the-art methods.

2

MULTI-SOURCE DOMAIN ADAPTATION AND LABEL UNIFICATION

This chapter corresponds to our published article:

Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, and Luc Van Gool. „mDALU: Multi-source domain adaptation and label unification with partial datasets.“ In: *ICCV*. 2021

In this chapter, we propose a new problem, multi-source domain adaptation and label unification (mDALU). One challenge of semantic segmentation for scene understanding is to generalize to new domains, to more classes and/or to new modalities. This necessitates methods to combine and reuse existing datasets that may belong to different domains, have partial annotations, and/or have different data modalities. This paper formulates this as mDALU problem, and proposes a novel method for it. Our method consists of a partially-supervised adaptation stage and a fully-supervised adaptation stage. In the former, partial knowledge is transferred from multiple source domains to the target domain and fused therein. Negative transfer between unmatching label spaces is mitigated via three new modules: domain attention, uncertainty maximization and attention-guided adversarial alignment. In the latter, knowledge is transferred in the unified label space after a label completion process with pseudo-labels. Extensive experiments on different tasks - 2D semantic image segmentation, and joint 2D-3D semantic segmentation - show that our method outperforms all competing methods significantly. Besides, it is proven that our method can be extended to other corresponding tasks such as image classification.

2.1 INTRODUCTION

The development of semantic segmentation for scene understanding is carried by two pillars: large-scale data annotation and deep neural networks. With new applications coming out every day, researchers need to constantly develop new methods and create new datasets. While we are able to develop novel neural networks for new tasks, the creation of new datasets can hardly keep up due to its huge cost.

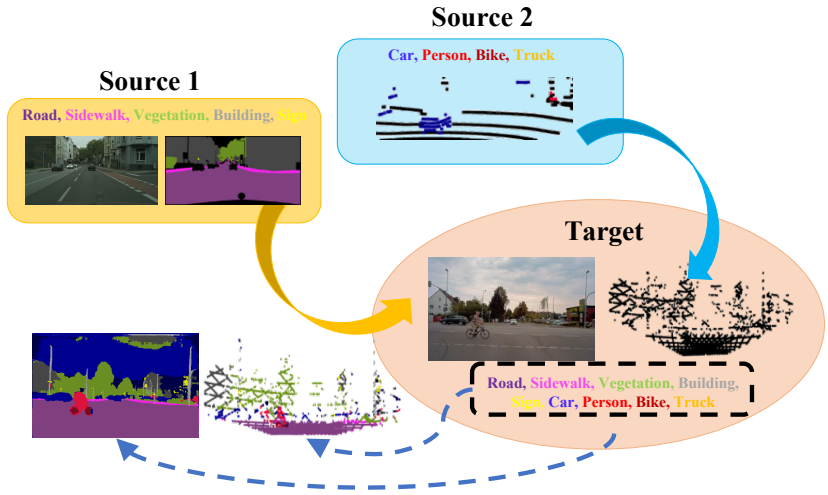


Figure 2.1: mDALU learns a complete-class and complete-modality semantic segmentation model for a new, unlabeled target domain, by using multiple datasets with partial-class annotation and partial data modality as source domains.

In the literature, a diverse set of learning paradigms, such as self-learning [67], semi-supervised learning [74] and transfer learning [30], have been developed to come to the rescue. We enrich this repository by developing a method to combine multiple existing datasets that have been annotated in different domains, for smaller-scale tasks (fewer classes), and/or with fewer data modalities. The importance of the method can be justified by the fact that as time goes, research goals will become more and more ambitious, so semantic segmentation models for more classes, new domains, and/or more data modalities are necessary.

To address this, we propose a multi-source domain adaptation and label unification (mDALU) problem. In this setting, there are multiple source domains and an unlabeled target domain. In each source domain, only samples (images, pixels, or LiDAR points) belonging to a subset of classes are labeled; the rest are unlabeled. The subsets of classes having labels can be different over different source domains, and can have inconsistent taxonomies, *e.g.*, truck is labeled as "truck" in one source domain but labeled as "vehicle" together with other types of vehicles in another. Further, the data modalities in different source domains can also be different, *e.g.*, one contains images and the other contains LiDAR point clouds. The goal is to obtain an semantic segmentation

model for all classes in the target domain. Fig. 2.1 shows an exemplar setting of mDALU. A comparison to other domain adaptation settings, in Table 2.1, shows that mDALU is very flexible.

This goal is challenging. Firstly, there is the notorious issue of negative transfer. While negative transfer is an issue also for standard transfer and multi-task learning, it is especially severe in our mDALU task due to the influence of unlabeled classes. To address this, we propose three novel modules, termed domain attention, uncertainty maximization and attention-guided adversarial alignment, to avoid making confident predictions for unlabeled samples in the source domains, and to enable robust distribution alignment between the source domains and the target domain. The method with the aforementioned modules and attention-guided prediction fusion is able to generate good results in the unified label space and on the target domain. In order to further improve the results, we need to fuse the supervision of all partial datasets to transfer the supervision in the unified label space. To this aim, we propose a pseudo-label based supervision fusion module. In particular, we generate pseudo-labels for the unlabeled samples in the source domains and all samples in the target domain. Standard supervised learning is then performed in the unified label space for the final model.

To showcase the effectiveness of our method, we evaluate it on different tasks: 2D semantic image segmentation, and joint 2D-3D semantic segmentation. Synthetic and real data, and images and LiDAR point clouds are involved. Also, non-overlapping, partially-overlapping and fully-overlapping label spaces, and consistent and inconsistent taxonomies across source domains are covered. Furthermore, it is proven that our method can be seamlessly extended to related tasks such as image classification. Experiments show that our method outperforms all competing methods significantly.

2.2 RELATED WORK

Multi-Source Domain Adaptation. Transfer learning and domain adaptation have been extensively studied in the past years. Several effective strategies have been developed such as minimizing maximum mean discrepancy [178, 114], moment matching [206], adversarial domain confusion [47, 177], entropy regularization [184], and curriculum domain adaptation [36]. While great progress has been achieved, most

Domain Adaptation Setting	Can Handle Multiple Source Domains?	Can Handle Multiple Data Modalities?	Can Handle Different Label Spaces of Source Domains?	Change of Label Space Size from Source to Target Domain	Can Handle Partial Annotations?	Can Handle Inconsistent Taxonomy?
Unsupervised Domain Adaptation [47]	No	No	–	Same Size	No	–
Partial Domain Adaptation [16]	No	No	–	Reduced	No	–
Multi-Source Domain Adaptation [140, 218]	Yes	No	No	Same Size	No	No
Category-Shift Multi-Source Domain Adaptation [198]	Yes	No	Yes	Increased	No	No
Multi-Modal Domain Adaptation [83]	Yes	Yes	No	Same Size	No	No
Multi-Source Open-Set Domain Adaptation [146, 134]	Yes	No	No	Same Size + 1*	Yes	No
Multi-Source Domain Adaptation and Label Unification (mDALU)	**Yes**	**Yes**	**Yes**	**Increased**	**Yes**	**Yes**

Table 2.1: Comparison between our mDALU and other domain adaptation settings (see Sec. 2.2 for details). It is clear that mDALU offers a very flexible and general setting. * "1" means an additional "unknown" class in the target domain.

algorithms focus on the single-source adaptation setting. This limits the methods from being used when data is collected from multiple source domains. That is why multi-source domain adaptation methods are proposed [35, 216, 140, 71, 218]. Yet, these methods all assume the same label space for all domains. Xu *et al.* [198] explores the problem of the category shift among different source domains, and adopts the k-way domain discriminator to reduce the effect of category shift. But the method is mainly proposed for the image classification task, and cannot deal with the problem of partial annotation, inconsistent taxonomies and modal differences among different source domains.

Open-Set/Partial Domain Adaptation. Recent research explores the category "openness" between the source domain and the target domain, which is divided into open-set domain adaptation and partial domain adaptation. Open-set domain adaptation [134, 157, 146] assumes that the target domain includes new classes that are unseen in the source domain, and aims to classify the unseen class samples as "unknown" class in the target domain. Partial domain adaptation [15, 210, 16, 84] aims to transfer knowledge from existing large-scale domains (e.g. 1K classes) to unknown small-scale domains (e.g. 20 classes) for customized applications. Different than both open-set and partial domain adaptation, our label space of the target domain is the union of label spaces of all source domains.

Learning from multiple datasets. Several successful methods [148, 147, 189, 86] have been proposed to learn a single universal network, that can represent different domains with a minimum of domain-specific parameters. But those methods do not consider domain adaptation and label space unification. Recently, Lambert *et al.*[92] presented a composite dataset that unifies different semantic segmentation datasets by reconciling the taxonomies, merging and splitting classes manually. But they do not address the problem of domain adaptation, partial annotation and cross-modal data, and they rely on the manual re-annotation for unification. The object detection method by Zhao *et al.* [219] performs label space unification from multiple datasets with partial annotations, but it does not consider other problems that are considered by our method such as domain discrepancies, inconsistent taxonomies and mismatched data modalities across the datasets.

2.3 APPROACH

2.3.1 *Problem Statement*

For the problem of mDALU, we are given K source domains $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K$. The K source domains contain the samples from K different distributions $P_{S_1}, P_{S_2}, ..., P_{S_K}$, which are labeled with $C_1, C_2, ..., C_K$ classes, resp. All the source domains can contain both partially labeled and unlabeled samples. The unlabeled samples can belong to the labeled classes of other domains. The label spaces $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K$ can be non-, partially-, or fully-overlapping with each other. Moreover, both consistent and inconsistent taxonomies among $\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K$ are allowed. Then the union of the label spaces $\mathcal{C}_i, i = 1, ..., K$ forms the unified and complete label space $\mathcal{C}_\cup = \mathcal{C}_1 \cup \mathcal{C}_2 \cdots \mathcal{C}_K$, including $C_\cup$ classes. Besides, the unlabeled target domain $\mathcal{T}$ is given, containing samples from the distribution P_T. Denoting the source samples $\mathbf{x}^{s_i} \in \mathcal{S}_i, i = 1, ..., K$ and the target samples $\mathbf{x}^t \in \mathcal{T}$, we have $\mathbf{x}^{s_i} \sim P_{S_i}, \mathbf{x}^t \sim P_T$, $P_{S_1} \neq P_{S_2} \neq ... \neq P_{S_K} \neq P_T$. The mDALU problem aims at training the model on the K source domains $\mathcal{S}_i, i = 1, ..., K$, labeled with C_i classes in each, and the unlabeled target domain $\mathcal{T}$, to improve the performance of the model on the target domain $\mathcal{T}$ in the unified label space $\mathcal{C}_\cup$. We use $\mathbf{y}^{s_i}$ to indicate the ground-truth label map of $\mathbf{x}^{s_i}$. Note that we present most of our approach with the notation of 2D semantic image segmentation. The translation to image classification and 3D point cloud segmentation is straightforward – by replacing pixels with images and by replacing pixels with 3D LiDAR points.

2.3.2 *Our Approach to mDALU problem*

As shown in Fig. 2.2, there are two stages in our approach, the partially-supervised adaptation stage and the fully supervised adaptation stage. In the partially-supervised adaptation stage, the partial supervision is transferred to the target domain from different source domains, respectively. Then in the fully-supervised adaptation stage, the supervision, in complete label space, is fused and self-completed on the unlabeled samples, and jointly transferred in the source domains and target domain. In order to realize adaptation under partial supervision, we propose three modules: DAT, UM and A^3 for the first stage. Then in the second stage, we use PSF and further learning. Below we provide

details of all these components. From Sec. 2.3.2.1 to Sec. 2.3.2.5, we first introduce our method for mDALU under consistent taxonomies. In this part, we first describe a basic version of our method composed of DAT and inference via attention-guided fusion, which will be followed by UM and A^3 to enhance the adaptation ability. Finally, we present PSF. Then in Sec. 2.3.2.6, we extend our proposed method towards mDALU under inconsistent taxonomies.

2.3.2.1 *Partially-Supervised Learning*

Different segmentation networks $G_i, i = 1, ..., K$ are adopted for different source domains $\mathcal{S}_i$. While their annotations cover partial label spaces $\mathcal{C}_i$, we train each network G_i in the unified label space $\mathcal{C}_\cup$ – some classes have no training data – with a standard cross-entropy loss $\mathcal{L}_{\text{psu}}$. The network G_i is composed of a feature extractor E_i and a label predictor B_i, *i.e.*, $G_i = \{E_i, B_i\}$. While we can average the results of these models directly in the target domain for predictions in the unified label space, coined multi-branch (MBR) fusion, this generates poor results. This is because the predictions of each model G_i for its unlabeled classes in $\mathcal{C}_\cup \setminus \mathcal{C}_i$ can be arbitrary numbers that dominate the averages. We thus propose the domain attention (DAT) module, which learns the attention map for G_i to signal on which area its prediction is reliable, for more effective fusion.

The attention map $\mathbf{a}^{s_i}$ in domain $\mathcal{S}_i$ is defined as:

$$\mathbf{a}^{s_i}(h,w) \begin{cases} = 1, \text{if } \mathbf{y}^{s_i}(h,w) \in \mathcal{C}_i \\ = 0, \text{if } \mathbf{y}^{s_i}(h,w) = \texttt{void}, \end{cases} \tag{2.1}$$

where (h, w) are pixel indices and `void` means no label. We train an attention network M_i for each source domain $\mathcal{S}_i$. The attention maps are predicted as $\tilde{\mathbf{a}}^{s_i} = M_i(\mathbf{x}^{s_i})$ and $\tilde{\mathbf{a}}^t_i = M_i(\mathbf{x}^t)$. The attention network M_i is composed of the feature extractor E_i and a new label predictor B^M_i: $M_i = \{E_i, B^M_i\}$. M_i is trained under an MSE loss $\mathcal{L}_{\text{att}}$, together with G_i in a multi-task setting.

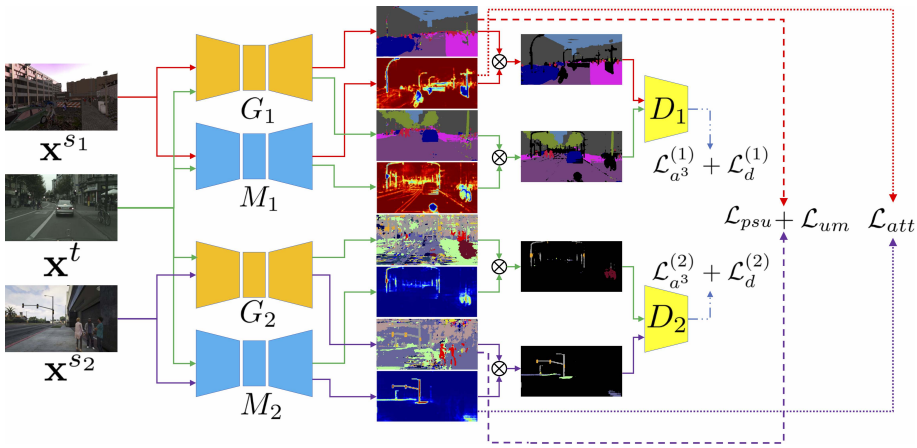
2.3.2.2 *Inference via Attention-Guided Fusion*

We feed an image $\mathbf{x}$ into semantic segmentation networks G_i to generate the corresponding probability maps $\hat{\mathbf{p}}_i \in [0,1]^{H \times W \times C_\cup}$, and into

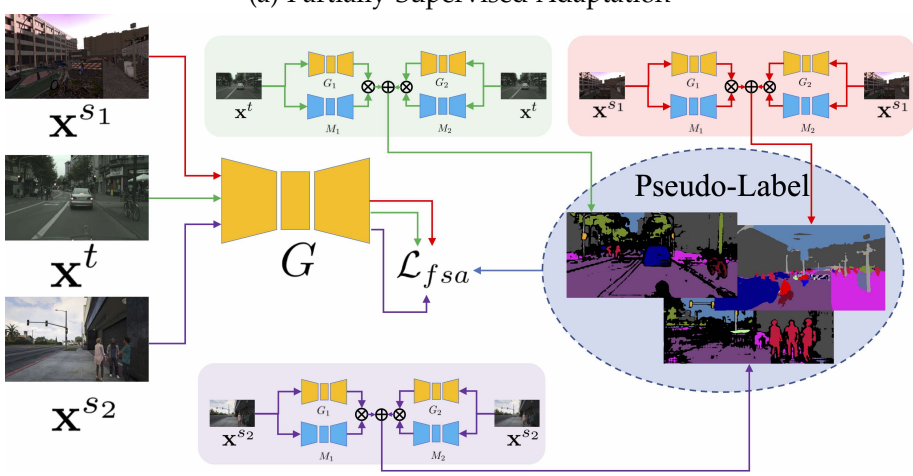
different attention networks M_i to generate attention maps $\hat{\mathbf{a}}_i$. Then we fuse the predictions by averaging $\hat{\mathbf{p}}_i$ weighted by $\hat{\mathbf{a}}_i$:

$$\mathbf{f} = \frac{\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i}{\sum_{j=1}^{C_\cup} (\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i)^{(j)}}, \tag{2.2}$$

where $(\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i)^{(j)}$ yields the probability of the j^{th} class. The predicted class is then obtained via `argmax`.



(a) Partially-Supervised Adaptation



(b) Fully-Supervised Adaptation

Figure 2.2: **Illustration of our approach to mDALU.** There are 2 stages, (a) partially supervised adaptation and (b) fully-supervised adaptation.

2.3.2.3 *Uncertainty Maximization (UM)*

Due to the lack of ground truth class supervision, while we have the attention-guided fusion, the wrong prediction of unlabeled samples in the source domains can still have negative effects for our cross-domain prediction fusion. In order to further reduce the negative effects of unlabeled samples $\mathbf{x}_u^{s_i}$ in source domains, we propose a module specifically to maximize uncertainties of the predictions on unlabeled samples in those domains. In particular, $G_i(\mathbf{x}_u^{s_i})$ is expected to equally spread the probability mass to all classes, *i.e.*, obeying the uniform categorical distribution $\mathcal{U}(C_\cup)$. The probability density function $q(j)$ of $\mathcal{U}(C_\cup)$ is formulated as $q(j) = \frac{1}{C_\cup}$, where $j = 1, 2, ..., C_\cup$ is to represent different classes. The probability distribution of the network prediction on unlabeled samples $G_i(\mathbf{x}_u^{s_i})$ is denoted as $p(j) = G_i(\mathbf{x}_u^{s_i})^{(j)}$, where $G_i(\mathbf{x}_u^{s_i})^{(j)}$ represents the probability of the j^{th} class. In order to maximize the uncertainty of the prediction on the unlabeled samples, the distribution distance between $p(j)$ and $q(j)$ is expected to be minimized. Following the distribution distance metric in [24], we adopt the Pearson χ^2-divergence for measuring the distribution distance, which is formulated as,

$$D_{\chi^2}(p||q) = \int_j ((\frac{p(j)}{q(j)})^2 - 1)q(j), \tag{2.3}$$

$$D_{\chi^2}(p||q) = C_\cup \sum_{j=1}^{C_\cup} p(j)^2 - 1. \tag{2.4}$$

On the basis of Eq. (2.4), we propose the square loss $\mathcal{L}_{um}$ for minimizing the Pearson χ^2-divergence, *i.e.*, maximizing the uncertainty of the prediction on the unlabeled samples. $\mathcal{L}_{um}$ can be written as

$$\mathcal{L}_{um} = \sum_{j=1}^{C_\cup} (G_i(\mathbf{x}_u^{s_i})^{(j)})^2. \tag{2.5}$$

Through the UM module, we encourage the model to make uniform categorical probability predictions, $\frac{1}{C_\cup}$, for unlabeled samples over the unlabeled classes, to best preserve the uncertainty to let the ground truth supervision of those classes from other source domains make the decision in the further attention-guided fusion and PSF process.

2.3.2.4 *Attention-Guided Adversarial Alignment (A^3)*

It has been proven in the literature that adversarial alignment is effective for domain adaptation. We extend the idea to mDALU. For adversarial alignment, one discriminator D_i is used for each source domain, to align the distribution between the source domain $\mathcal{S}_i$ and the target domain $\mathcal{T}$. In general unsupervised domain adaptation, the discriminator training loss $\mathcal{L}_d$ and the adversarial loss $\mathcal{L}_{adv}$ [175] for the source domain $\mathcal{S}_i$ and the target domain $\mathcal{T}$ is defined as

$$\mathcal{L}_{adv}^{(i)}(\mathbf{x}^t) = -\log(D_i(G_i(\mathbf{x}^t))) \tag{2.6}$$

$$\begin{aligned}\mathcal{L}_{d}^{(i)}(\mathbf{x}_i^{s_i}, \mathbf{x}^t) = &-\log(D_i(G_i(\mathbf{x}^{s_i}))) \\ &-\log(1 - D_i(G_i(\mathbf{x}^t))).\end{aligned} \tag{2.7}$$

Yet, in our mDALU problem, there is no ground truth label guidance available for the unlabeled classes. A direct alignment between the source domain and the target domain will cause negative transfer, *i.e.*, the transfer of incorrect knowledge from the unlabeled parts in the source domains to the target domain. Here, we again use our attention map $\mathbf{a}^{s_i}$ to alleviate this problem by proposing an attention-guided adversarial loss:

$$\mathcal{L}_{a^3}^{(i)}(\mathbf{x}^t) = -\log(D_i(G_i(\mathbf{x}^t) \otimes M_i(\mathbf{x}^t))), \tag{2.8}$$

$$\begin{aligned}\mathcal{L}_{d}^{(i)}(\mathbf{x}_i^{s_i}, \mathbf{x}^t) = &-\log(D_i(G_i(\mathbf{x}^{s_i}) \otimes M_i(\mathbf{x}^{s_i}))) \\ &-\log(1 - D_i(G_i(\mathbf{x}^t) \otimes M_i(\mathbf{x}^t))),\end{aligned} \tag{2.9}$$

where $\otimes$ represents element-wise multiplication.

Then the overall loss for our method at the first stage is:

$$\mathcal{L}_{all} = \mathcal{L}_{psu} + \mathcal{L}_{att} + \mathcal{L}_{um} + \lambda \sum_{i=1}^{K} \mathcal{L}_{a^3}^{(i)}, \tag{2.10}$$

where λ is the hyper-parameter to balance out the attention-guided adversarial loss against other losses. The whole optimization objective for our first partially-supervised domain adaptation stage can be formulated as:

$$\min_{G_i} \max_{D_i} \mathcal{L}_{all}. \tag{2.11}$$

2.3.2.5 *Pseudo-Label Based Supervision Fusion (PSF)*

In the first partially-supervised adaptation stage, knowledge in different label spaces $\mathcal{C}_i$ is transferred from different source domains to the target domain. In the second fully-supervised adaptation stage, we aim at learning and transferring knowledge in the complete and unified label space $\mathcal{C}_\cup$ between all domains jointly. In order to realize that, we complete the label spaces for all the related domains $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K, \mathcal{T}$ with pseudo-labels, *i.e.*, fuse the supervision from different label spaces $\mathcal{C}_i$ to get the complete and unified supervision $\mathcal{C}_\cup$. Here we present our pseudo-label based supervision fusion (PSF) method.

In order to complete the label space in the source domain $\mathcal{S}_i$, we feed each of the source image samples $\mathbf{x}^{s_i}$ into every semantic model $G_k, k = 1, ..., K$, to generate 'partial' semantic probability maps $\hat{\mathbf{p}}_k^{s_i} \in [0,1]^{H \times W \times C_\cup}$ and to every attention network $M_k, k = 1, ..., K$ for the attention map $\hat{\mathbf{a}}_k^{s_i} \in [0,1]^{H \times W}$. The fused prediction $\mathbf{f}^{s_i}$ is obtained via Eq.(2.2). We denote the predicted label map as $\bar{\mathbf{y}}^{s_i}$, generated by using an `argmax` operation over $\mathbf{f}^{s_i}$. The 'pseudo-label' map $\hat{\mathbf{y}}^{s_i}$ for the source domain $\mathcal{S}_i$ is defined as:

$$\hat{\mathbf{y}}^{s_i}(h,w) = \begin{cases} \mathbf{y}^{s_i}(h,w), \text{if } \mathbf{y}^{s_i}(h,w) \neq \texttt{void} \\ \bar{\mathbf{y}}^{s_i}(h,w) \text{if } \mathbf{y}^{s_i}(h,w) = \texttt{void} \\ \quad \text{and } \mathbf{f}^{s_i}(h,w,\bar{\mathbf{y}}^{s_i}(h,w)) > \delta \\ \texttt{void}, \text{ otherwise} \end{cases} \tag{2.12}$$

where δ is a threshold determining whether to select the predicted pseudo-label.

On the target domain $\mathcal{T}$, since no ground truth labels are available, we obtain pseudo labels directly from the predicted label map $\bar{\mathbf{y}}^t$ (obtained from $\mathbf{f}^t$ via an `argmax`):

$$\hat{\mathbf{y}}^t(h,w) = \bar{\mathbf{y}}^t(h,w) \text{ if } \mathbf{f}^t(h,w,\bar{\mathbf{y}}^t(h,w)) > \delta. \tag{2.13}$$

By using the generated fused pseudo-label $\hat{\mathbf{y}}^{s_i}, \hat{\mathbf{y}}^t, i = 1, ..., K$, we complete the label space from $\mathcal{C}_i$ to $\mathcal{C}_\cup$ for the source domain $\mathcal{S}_i$, and from $\emptyset$ to $\mathcal{C}_\cup$ for the target domain $\mathcal{T}$. We then train the network G for all the related domains $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_K, \mathcal{T}$ with all the datasets in the unified label space. In total, the loss $\mathcal{L}_{fsa}$ for our second 'fully-supervised' adaptation stage is:

$$\mathcal{L}_{fsa} = \sum_{i=1}^{K} \mathcal{L}_{ce}^{s_i} + \mathcal{L}_{ce}^{t}, \tag{2.14}$$

where $\mathcal{L}_{ce}$ is the standard cross-entropy loss.

2.3.2.6 *Inconsistent Taxonomies*

The above method is able to deal with the mDALU problem under consistent taxonomies, *i.e.*, the different classes in all source domains are exclusive with each other. Yet, there might be inconsistent taxonomies between different source domains, causing a performance drop for the inconsistent taxonomies classes. Here, we introduce the extension of our above method, to handle the inconsistent taxonomies problem. Denoting the classes in the label spaces $\mathcal{C}_i$ as $\mathbf{c}_i^o$, we have $\mathcal{C}_i = \{\mathbf{c}_i^o, o = 1, 2, ..., C_i\}$. Then the inconsistent taxonomies among different source domains can be defined as, $\exists \mathbf{c}_p^q \in \mathcal{C}_p, \mathbf{c}_m^n \in \mathcal{C}_m, p, m = 1, ..., K, p \neq m, q = 1, ..., C_p, n = 1, ..., C_m$, we have $\mathbf{c}_p^q \neq \mathbf{c}_m^n$, and $\mathbf{c}_p^q \cap \mathbf{c}_m^n \neq \emptyset$. The inconsistent taxonomies classes between different source domains $\mathcal{S}_p$ and $\mathcal{S}_m$ are denoted as $\mathbf{c}_p^q \in \mathcal{C}_p$ and $\mathbf{c}_m^n \in \mathcal{C}_m$. For example, the truck is labeled as "truck" class $\mathbf{c}_p^q$ in one dataset $\mathcal{S}_p$, while it is labeled as "vehicle" class $\mathbf{c}_m^n$ together with other vehicles in another dataset $\mathcal{S}_m$. Another typical example is motorcycles being labeled as "cycle" class $\mathbf{c}_p^q$ together with other cycles in one dataset $\mathcal{S}_p$, but being labeled as "vehicle" class $\mathbf{c}_m^n$ together with other vehicles in another dataset $\mathcal{S}_m$. In the unified label space of the target domain, the conflict part $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ is assigned to either $\mathbf{c}_p^q$ or $\mathbf{c}_m^n$ exclusively. Without loss of generality and for reasons of clarity, it is assumed that the $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ is assigned to $\mathbf{c}_p^q$. Then in order to solve the conflict of $\mathbf{c}_p^q$ and $\mathbf{c}_m^n$, in the attention-guided fusion, we introduce the additional class-wise weight map $\mathbf{w}_i \in \mathbb{R}^{H \times W \times C_\cup}$, and Eq. (2.2) is extended to Eq. (2.16),

$$\mathbf{w}_i(h, w, j) = \begin{cases} = v, \text{if argmax}\, \hat{\mathbf{p}}_i(h, w) = q', \text{and } i = p, \\ \text{and } \text{argmax}\, \hat{\mathbf{p}}_m(h, w) = n', \text{and } j = q' \\ = 1, \text{otherwise} \end{cases} \tag{2.15}$$

$$\mathbf{f} = \frac{\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i \otimes \mathbf{w}_i}{\sum_{j=1}^{C_\cup} (\sum_{i=1}^{K} \hat{\mathbf{a}}_i \otimes \hat{\mathbf{p}}_i \otimes \mathbf{w}_i)^{(j)}}, \tag{2.16}$$

where $v > 1$ in Eq. (2.15) is a hyper-parameter, set to 5.0. v is used to increase the weight of class $\mathbf{c}_p^q$ of the corresponding prediction $\hat{\mathbf{p}}_p$ in Eq. (2.16), to convert $\mathbf{c}_p^q \cap \mathbf{c}_m^n$ to $\mathbf{c}_p^q$ in the prediction fusion. q', n' are the class indices of $\mathbf{c}_p^q$ and $\mathbf{c}_m^n$ in the unified label space $\mathcal{C}_\cup$. Correspondingly, under inconsistent taxonomies, besides the unlabeled samples in the

source domains being completed with the predicted pseudo-label as in Eq. (2.12), the conflict part $\mathbf{c}_p^q \cap \mathbf{c}_m^n$, which is labeled as $\mathbf{c}_m^n$ originally in $\mathcal{S}_m$, is relabeled with the predicted pseudo-label $\bar{\mathbf{y}}^{s_i}(h,w)$, *i.e.*,

$$\begin{aligned}\hat{\mathbf{y}}^{s_m}(h,w) = q', \text{ if } \mathbf{f}^{s_m}(h,w,q) > \delta \\ \text{and } \bar{\mathbf{y}}^{s_m}(h,w) = q' \text{ and } \mathbf{y}^{s_m}(h,w) = n'.\end{aligned} \tag{2.17}$$

2.4 EXPERIMENTS

We evaluate the effectiveness of our method mDALU under different settings. We build benchmarks for image classification, 2D semantic image segmentation, and 2D-3D cross-modal semantic segmentation.

Parameters. In the image classification experiment, the hyperparameter λ in Eq. (2.10) is set as 1.0, and δ in Eq. (2.12) and Eq. (2.13) is set as 0.5. The images are resized to 32×32. We use the the Adam optimizer [88] with $\beta_1 = 0.9, \beta_2 = 0.999$ and the weight decay as 5×10^{-4}. The learning rate is set as 2×10^{-4}. We adopt the same network architecture as that of the digits classification experiments in [140]. In the 2D semantic image segmentation experiments, the hyperparameter λ in Eq. (2.10) is set as 0.001, and δ in Eq. (2.12) and Eq. (2.13) is set as 0.2, 0.5 and 0.4 for SYNTHIA, GTA5 and Cityscapes dataset, respectively. The images are resized to 1024×512. We use the SGD optimizer for training the semantic segmentation network, whose momentum is 0.9, weight decay is 5×10^{-4} and learning rate is 2.5×10^{-4} with polynomial decay of power 0.9. Meanwhile, the Adam optimizer is used for training the discriminator network, whose momentum is $\beta_1 = 0.9, \beta_2 = 0.99$, weight decay is 5×10^{-4} and learning rate is 1×10^{-4} with polynomial decay of power 0.9. We adopt the same semantic segmentation and discriminator network architecture as that of [175]. In the cross-modal semantic segmentation experiments, we follow the exactly same data augmentation and preprocess procedure as that of [83]. The hyperparameter δ in Eq. (2.12) and Eq. (2.13) is set as 0.2. We use the Adam optimizer for training the 2D and 3D semantic segmentation network, with $\beta_1 = 0.9, \beta_2 = 0.999$. The learning rate is set as 1×10^{-3}.

Experiment	Label Space									
Non-Overlapping(Table 2.3)	Domain	Source1	Source2	Target	Source1	Source2	Target	Source1	Source2	Target
	Dataset	SVHN	SYN	MT	MT	SVHN	SYN	MNIST	SYN	SVHN
	Class	0~4	5~9	0~9	0~4	5~9	0~9	0~4	5~9	0~9
Partially-Overlapping(Table 2.5)	Domain	Source1	Source2	Target	Source1	Source2	Target	Source1	Source2	Target
	Dataset	SVHN	SYN	MT	MT	SVHN	SYN	MNIST	SYN	SVHN
	Class	0~6	3~9	0~9	0~6	3~9	0~9	0~6	3~9	0~9

Table 2.2: **Label space** of different source domains and the target domain in the mDALU image classification benchmark.

2.4.1 *Image Classification*

Datasets. MNIST [93] is a hand-written numbers image dataset, SVHN [127] is a street view house numbers image dataset and Synthetic Digits [47] is a synthetic numbers image dataset.

Setup. In the classification benchmark, we adopt the digits classification images from three different datasets, MNIST [93], Synthetic Digits [47], and SVHN [127], coined "MT", "SYN" and "SVHN", resp. Each time, one of them is taken as the target domain, the other two as source domains. There are 10 classes, from '0' to '9', in the target domain. In our main setting, we adopt the most difficult setup to evaluate different methods, where the label spaces of different source domains are non-overlapping. Only half the classes are labeled in each of the source domains. The partially-overlapping situation is also explored. For fair comparison, we adopt the same network architecture used in [140] for all methods. The classification performance is evaluated on all 10 classes in the target domain. The label spaces of different source domains and target domain are detailed in Table 2.2.

Comparison with SOTA. Table 2.3 compares our method with other SOTA methods which include 1) unsupervised domain adaptation method DANN [47], 2) category-shift unsupervised domain adaptation method DCTN [198], 3) multi-source unsupervised domain adaptation method M^{3}SDA [140], and 4) label unification method AENT [219]. It can be seen that without the pseudo label (PL) generation part, other domain adaptation based methods, DANN, DCTN, and M^{3}SDA show the negative transfer effect, or perform similarly to the baseline trained with source data only. This is because each source domain can only provide guidance for a partial label space, and the adaptation in the partial label space guides the prediction on the target domain to the biased label space when training with data from different source domains. This renders the prediction on the target domain contradictory and

Method	MT	SYN	SVHN	Avg
Source	76.76 ± 0.63	61.77 ± 1.05	43.42±1.89	60.65±1.19
DANN[47]	77.30 ± 2.57	60.31 ± 0.99	41.65±2.34	59.75±1.97
DANN *	71.29 ± 0.48	55.94 ± 0.51	35.60 ± 1.63	54.28 ± 0.87
DCTN [198]	68.10±0.2	62.72±0.30	48.11±0.57	59.64±0.36
DCTN *	72.01 ± 1.22	63.33 ± 0.20	49.34 ± 1.28	61.59 ± 0.90
M^{3}SDA [140]	76.56±0.71	61.25±2.33	43.13±3.55	60.31±2.20
M^{3}SDA *	72.50 ± 2.64	55.92 ± 1.04	36.24 ± 1.70	54.89 ± 1.79
AENT[219]	73.24±1.76	68.66±1.32	52.80 ± 0.92	64.90 ± 1.33
Ours w/o PSF	**81.23±0.92**	**78.97±0.45**	**65.20±0.58**	**75.13±0.65**
DCTN w/ PL [198]	73.40±0.85	65.63±0.43	52.12±0.07	63.72 ± 0.45
AENT[219] w/ PL	78.56±1.23	70.25 ± 0.39	59.24 ± 1.01	69.35 ± 0.88
Ours	**86.18±0.45**	**81.91±0.33**	**68.92±0.81**	**79.00 ± 0.53**

Table 2.3: **Quantitative comparison of image classification.** "MT", "SYN", and "SVHN" represent the target domain. "PL" represents to add the pseudo-label training module, which is specifically adjusted according to their own paper's design. * represents to remove the unlabeled samples in the training data. We implement AENT for classification by utilizing the ambiguity cross entropy loss proposed in [219].

the model hard to adapt to the complete label space. In contrast, the label-unification based method AENT obtained a performance gain of 4.25%, from 60.65% to 64.90%, compared with the source-only baseline. This is because it uses an ambiguity cross entropy loss, to avoid the prediction of the source domain data being restricted in a partial label space.

In our first partially-supervised adaptation stage, the performance is further improved to 75.13%, which proves the effectiveness of our DAT, UM and A^3 module for preventing the negative transfer effect. After the second fully-supervised adaptation stage, by adding the PSF module, our model strongly outperforms DCTN [198] and AENT [219], both with pseudo-label training, by 15.28% and 9.65%, resp. This proves the effectiveness of our entire method for domain adaptation, label space completion and supervision fusion. The ablation results in Table 2.4 show that each part of our model contributes to its performance.

Partially Overlapping. In Fig. 2.3, it is shown that the testing accuracy on the target domain increases, as more and more common classes in the source domains are available. In Table 2.5, we compare

MBR	UM	A^3	PSF	MT	SYN	SVHN	Avg
				76.76 ± 0.63	61.77 ± 1.05	43.42±1.89	60.65±1.19
✓				72.21±1.89	62.41±0.58	50.24±1.23	61.62±1.23
✓	✓			84.74±0.54	76.12±0.85	58.39±0.57	73.08± 0.65
✓	✓	✓*		81.38±0.79	78.20±1.3	65.12±0.64	74.90 ± 0.91
✓	✓	✓		81.23±0.92	78.97±0.45	65.20±0.58	75.13 ± 0.65
✓	✓	✓	✓	**86.18±0.45**	**81.91±0.33**	**68.92±0.81**	**79.00 ± 0.53**

Table 2.4: **Ablation study under the image classification setting.** MBR: multi-branch network, *i.e.*, adopts different networks G_i for different source domains. * indicates there is no adversarial part in the A^3 module, *i.e.*, only the DAT module. The best results are denoted in bold.

Method	MT	SYN	SVHN	Avg
Source	82.10±1.50	73.37± 0.67	57.50±1.93	70.99 ± 1.37
DANN[47]	80.13±1.60	72.97±0.49	55.00±0.73	69.37 ± 0.94
DCTN[198]	78.56±0.47	72.33 ± 0.04	60.86±0.21	70.58 ± 0.24
M^3SDA[140]	81.52 ± 1.55	72.91 ± 0.68	54.26±0.66	69.56 ± 0.96
AENT[219]	79.12 ± 1.07	81.99 ± 0.87	69.07 ± 1.93	76.73 ± 1.29
Ours w/o PSF	**85.39 ± 1.32**	**85.33± 1.21**	**76.48±1.31**	**82.40 ± 1.28**

Table 2.5: Quantitative comparison of image classification, under the partial overlap setting with 4 common classes.

the model performance of our method with other SOTA methods when the source domains are partially overlapping, with 4 common classes. It is shown that our method still strongly outperforms the adaptation-based methods, DANN, DCTN, M^3SDA, and the label unification based method, AENT, 82.40% v.s. 69.37%, 70.58%, 69.56%, 76.73%. It further verifies the effectiveness of our model in the partial overlap situation.

2.4.2 *2D Semantic Image Segmentation*

Datasets. *Cityscapes.* Cityscapes is a dataset composed of the street scene images collected from different European cities. We use the training set of Cityscapes covering 2993 images, without the label information, as the target domain during the training stage. And we adopt the validation set of Cityscapes, which are composed of 500 images and densely labeled with 19 classes, to evaluate the semantic

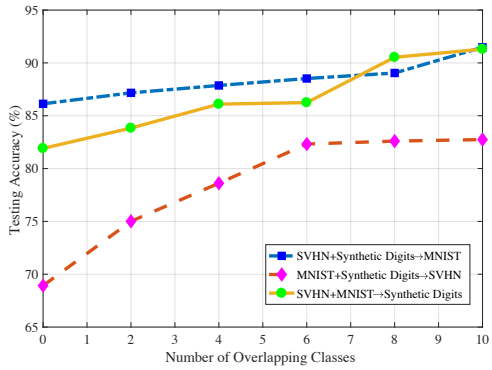


Figure 2.3: Accuracy in the target domain as a function of the number of overlapping classes between the source domains.

segmentation performance of the model on the target domain. *GTA5.* GTA5 is a synthetic urban scene image dataset, whose images are rendered from the game engine. The scene of the images is based on the city of Los Angeles. In our 2D semantic image segmentation benchmark, we use 24966 densely labeled images in the GTA5 dataset as one of our source domains, whose annotation is compatible with that of Cityscapes. *SYNTHIA.* SYNTHIA is a synthetic dataset, containing photo-realistic images rendered from a virtual city. We use the SYNTHIA-RAND-Cityscapes subset, which contains 9400 densely labeled images and the 16 class annotation of which is compatible with that of Cityscapes. In our 2D semantic image segmentation benchmark, the labeled SYNTHIA dataset serves as one of our source domains.

Setup. In the single mode semantic segmentation setting, we adopt the synthetic-to-real image semantic segmentation setup. The synthetic image datasets GTA5 [149] and the SYNTHIA [152] are taken as the source domains, while the real image dataset Cityscapes [34] is used as the target domain. Information of 19 classes needs to be transferred to the Cityscapes dataset. In our main setting, the label spaces of SYNTHIA and GTA5 are non-overlapping. In the SYNTHIA dataset, the label of 7 classes are available, incl. road, sidewalk, building, vegetation, sky, person and car. In GTA5, the labels of 12 classes are available, being wall, fence, pole, light, sign, terrain, rider, truck, bus, train, motorcycle and bicycle. Furthermore, we also explore the performance of our model when the images of the two source domains are fully

labeled. Moreover, we verify the effectiveness of our model when the taxonomies of different source domains are inconsistent. In those inconsistency experiments, for GTA5, the labels wall, fence, pole, light, sign, terrain, truck, bus, train, person (incl. person and rider), cycle (incl. bicycle and motorcycle) are available. In SYNTHIA, the labels road, sidewalk, building, vegetation, sky, person, rider, car, public facilities (incl. wall, fence, pole), motorcycle and bicycle are available. In order to further evaluate the performance of all methods when combined with the pixel-level domain adaptation methods [225, 72], we conduct experiments in two settings; 1) source domain images are not translated with CycleGAN [225], named as "NT"; 2) source domain images are translated with CycleGAN, named as "T". Also, in order to verify model performance combined with output-level adaptation method [175], we conduct additional experiments which include "ADV" in the fully-supervised adaptation stage. "ADV" generates the complete source domain label as in PSF, and then trains the semantic segmentation model via adversarial adaptation between pseudo-complete source domain and unlabeled target domain in the output-level space. For fair comparison, all the methods use the DeepLabv2-ResNet101 [20, 68] semantic segmentation network.

Comparison with SOTA. In Table 2.6a, we show a quantitative comparison for semantic segmentation between our method and other SOTA methods. It is shown that our method without adding PSF strongly outperforms the adaptation-based AdaptSegNet[175], the self-supervision-based MinEnt[184], and the method combining adaptation and self-supervision Advent [184]. Our method achieves 36.3% and 38.1% in the "NT" and "T" settings, resp. Similar to the image classification results, without using the translated source images, the adaptation-based methods suffer from negative transfer and the performance is lower than the source-only baseline. By using the translated source images in "T", different source domain images are all Citysapes-like images. The different source domains can be seen as a larger unified source domain, which can provide guidance for the complete label space to some extent. So all adaptation-based or self-supervision based methods perform much better in the "T" situation, compared with the non-adapted baseline. Yet, even in the "T" situation, our method still provides an advantage by further completing the label space, through our partially supervised adaptation. This proves the effectiveness of our method in preventing negative transfer and in completing the label

Method	NT	T
Source	17.7	24.0
AdaptSegNet[175]	7.7	30.8
MinEnt[184]	27.1	30.1
Advent[184]	11.8	30.3
Ours w/o PSF	**36.3**	**38.1**
Ours (ADV)	40.1	41.5
Ours (PSF)	37.3	42.4
Ours (ADV+PSF)	**40.6**	**42.8**

(a)

MBR	UM	A^3	PSF	ADV	NT	T
					17.7	24.0
✓					20.9	21.4
✓	✓				27.6	36.8
✓	✓	✓*			29.1	37.0
✓	✓	✓			36.3	38.1
✓	✓			✓	35.4	40.9
✓	✓		✓		31.4	41.5
✓	✓	✓		✓	40.1	41.5
✓	✓	✓	✓		37.3	42.4
✓	✓	✓	✓	✓	**40.6**	**42.8**

(b)

Table 2.6: (a) Quantitative comparison of single mode semantic segmentation, SYNTHIA+GTA5→ Cityscapes. The mIoU results are reported for 19 classes. (b) Ablation study for single mode segmentation. * indicates there is no adversarial part in the A^3 module, *i.e.*, only the DAT module. "ADV+PSF" means to combine "ADV" and "PSF" by completing the label space and generating pseudo-labels in the source and target domains, then adversarial alignment in the output space is adopted during the second stage training.

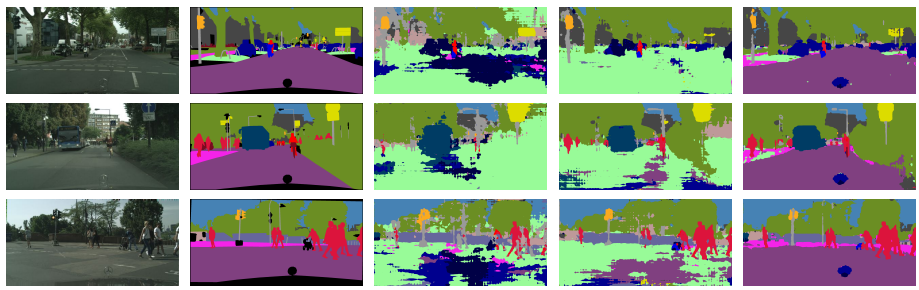
space. By further adding the second "fully-supervised" adaptation stage, the model achieves a new SOTA performance in both the "T" and the "NT" settings. An ablation study, see Table 2.6b, confirms all parts of our method add to its performance, and the output space alignment "ADV" is helpful as well. Fig. 2.4 shows qualitative results on Cityscapes.

Fully labeled. In the fully labeled setting, *i.e.*, the source domain images are labeled with all considered classes - 16 classes in SYNTHIA and 19 classes in GTA5 - Table 2.7 shows that our model still outperforms other unsupervised domain adaptive semantic segmentation methods, 43.1% vs. 40.8%, 42.2%, and 42.9%. Our model also outperforms the SOTA method for multi-source domain adaptive semantic segmentation MADAN [218], 41.9% vs. 41.4%.

Inconsistent Taxonomies. Table 2.8 shows that our method is advantageous when taxonomies are inconsistent, 40.0% vs. 28.1%, 31.9%, 32.2%. In the partially supervised adaptation stage, as in Sec. 2.3.2.6, by adding higher weights to "person", "rider", "motorcycle" and "bicycle" for SYNTHIA and "wall", "fence" and "pole" for GTA5, our method

Method	Base	mIoU*	mIoU
Source	ResNet-101	42.8	39.1
AdaptSegNet[175]		45.2	40.8
Minentropy[184]		46.4	42.2
Advent[184]		46.7	42.9
Ours w/o PSF		**46.8**	**43.1**
Source[218]	VGG-16	37.3	–
MADAN[218]		41.4	–
Ours w/o PSF		**41.9**	**38.0**

Table 2.7: Single mode segmentation results, under fully-labeled setting and "T". mIoU* is the mean IoU of 16 classes in SYNTHIA, while mIoU is that of all 19 classes.



(a) Image (b) Ground Truth (c) Source Only (d) MinEnt (e) Ours

Figure 2.4: Qualitative results of 2D semantic segmentation.

can achieve a higher performance than inference without weighting, 37.2% vs. 35.3%. After the fully supervised adaptation stage, the performance can be further improved to 40.0%. The detailed performance for inconsistent taxonomies classes in Table 2.8 underlines the effectiveness of our method for the inconsistent taxonomies.

Attention visualization for semantic segmentation. During the "partially-supervised adaptation" stage, we introduce the attention map in the domain attention (DAT) module, the attention-guided adversarial alignment (A^3) module and the inference via attention-guided fusion. In order to verify the effectiveness of our attention map prediction, we show the qualitative visualization of the attention map on the target domain images in Fig. 2.5. Corresponding to the Sec. 2.3.2.1, the attention map $\tilde{\mathbf{a}}_1^t$ and $\tilde{\mathbf{a}}_2^t$, are generated by feeding the target domain image $\mathbf{x}^t$ into the attention network M_1 and M_2. It is shown that our predicted attention map $\tilde{\mathbf{a}}_1^t$, corresponding to the source domain $\mathcal{S}_1$, has higher attention value, for the objects belonging to the partial label space $\mathcal{C}_1$,

Method	wall	fence	pole	person	rider	motorcycle	bicycle	mIoU
Source	2.6	12.0	12.3	40.6	0.5	0.1	28.6	19.8
AdaptSegNet[175]	7.1	2.6	4.0	33.2	6.9	1.8	37.6	28.1
Minentropy[184]	6.7	**18.1**	23.0	28.8	6.6	1.0	42.3	31.9
Advent[184]	6.2	11.5	11.4	32.8	12.2	0.9	41.2	32.2
Ours w/o PSF	12.3	15.2	21.2	48.4	3.3	1.3	42.4	35.3
Ours w/o PSF *	**14.1**	15.3	**30.6**	48.1	17.9	13.0	42.1	37.2
Ours (PSF)	13.3	17.9	**30.6**	**53.7**	**18.2**	**19.8**	**43.2**	**40.0**

Table 2.8: Quantitative comparison of single mode segmentation, with inconsistent taxonomies, in the "T" setting. *During inference, an additional weights map is adopted in case of inconsistent taxonomies as in Sec. 2.3.2.6. The detailed performance on inconsistent taxonomies classes is also shown. The mIoU is reported for 19 classes.

such as the road, sidewalk, building, vegetation, sky and car. And the predicted attention map $\tilde{\mathbf{a}}_2^t$, corresponding to the source domain $\mathcal{S}_2$, has higher attention value, for the objects belonging to the partial label space $\mathcal{C}_2$, such as the fence, pole, light, sign, bus, motorcycle and bicycle. It proves the validity of our attention map prediction.

2.4.3 *Cross-Modal Semantic Segmentation*

Datasets. *Nuscenes*. Nuscenes [13] is an autonomous driving dataset covering 1000 driving scenes, which are collected from the Boston and Singapore. Each scene, of 20-second length, is sampled and annotated at 2HZ, resulting in 40K well-annotated keyframes for 3D bounding boxes of the objects. In our cross-modal semantic segmentation benchmark, we adopt the training set of the Nuscenes, including 28130 keyframes 3D LiDAR points, as the 3D source domain. Then as done in [83], we generate the 3D point-wise semantic labels from the 3D bounding boxes, by assigning the object label to the points inside the bounding box and taking the points outside the bounding box as unlabeled points. *A2D2*. A2D2 [50] is an autonomous driving dataset, including simultaneously recorded paired 2D images and 3D LiDAR points. The A2D2 covers

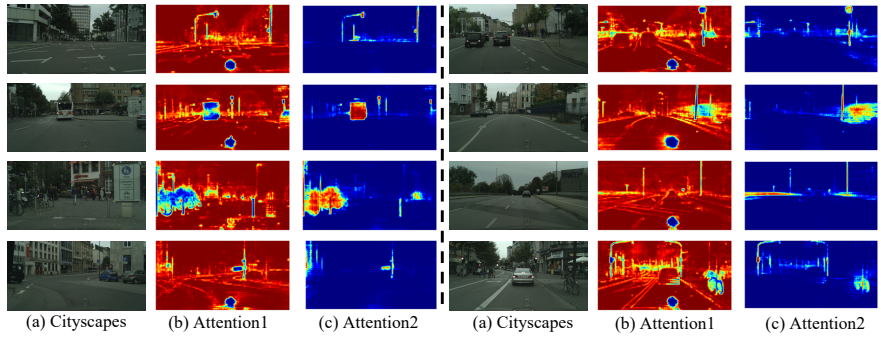


Figure 2.5: Visualization of the attention map $\tilde{\mathbf{a}}_1^t$ and $\tilde{\mathbf{a}}_2^t$ of the target domain images. (a) is the Cityscapes image $\mathbf{x}^t$. (b) is the attention map $\tilde{\mathbf{a}}_1^t$, generated by feeding the $\mathbf{x}^t$ into the attention network M_1. (c) is the attention map $\tilde{\mathbf{a}}_2^t$, generated by feeding the $\mathbf{x}^t$ into the attention network M_2. Red parts are the parts with higher attention value, while the blue parts with lower attention value.

20 scenes, which are corresponding to 28637 frames for training. And the scene 20180807_145028 is used for validation. The 2D images are densely labeled with 38 semantic classes. Following [83], the 3D point-wise semantic labels are generated by the reprojection to the 2D images. In our cross-modal semantic segmentation benchmark, the A2D2 serves as the target domain. We use the training set of A2D2 without the label information during training, including the paired 2D images and 3D LiDAR points. And we use the validation set 20180807_145028 with the ground truth label for evaluating the performance.

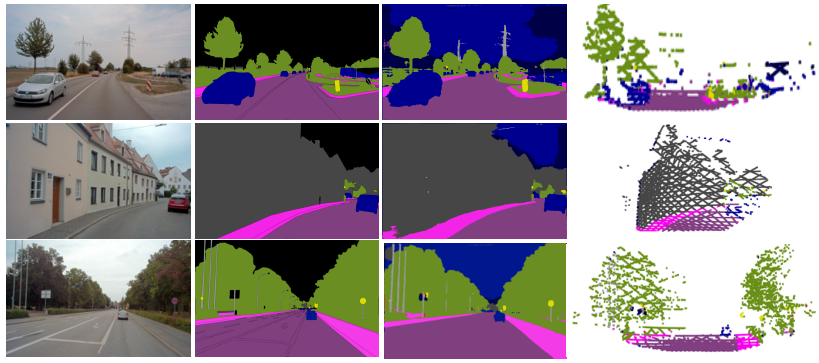
Setup. In the cross-modal semantic segmentation setting, the 2D RGB images from Cityscapes [34], and the 3D LiDAR point clouds from Nuscenes [13] are treated as two different source domains, while the paired but unlabeled 2D RGB images and 3D point clouds from A2D2 [50] are used as the target domain. There are 10 classes in total that need to be transferred to the target domain. In Cityscapes, the label for 6 classes are given, covering road, sidewalk, building, pole, sign and nature. In Nuscenes the labels for 4 classes are given, incl. person, car, truck and bike. The 2D RGB images and 3D point clouds in the target domain are registered via a projection matrix between the 2D pixel and 3D points. Following [83], we adopt U-Net-ResNet34 [151, 68] as the 2D semantic segmentation network, and SparseConvNet [62] for 3D

Cityscapes + Nuscenes → A2D2	2D	3D	Fuse
Source	37.5	2.0	42.5
xMUDA[83]	16.3	1.7	9.1
ES + MinEnt[184]	22.3	1.5	20.8
ES + KL[83]	21.7	1.5	19.7
xMUDA + AKL	27.5	2.3	21.1
xMUDA + AKL + COMP	32.1	2.9	37.7
Ours w/o PSF	38.1	2.4	49.9
Ours	**54.9**	**37.1**	**55.7**

Table 2.9: Quantitative comparison of cross modal segmentation, Nuscenes+Cityscapes→ A2D2. "Fuse" represents the average fusion of the prediction probability from 2D models and 3D models; the final class prediction is the maximum of the fused probability. "ES" means 2D and 3D average fusion ensemble. "KL" means KL-divergence alignment. "AKL" means adaptive KL-divergence alignment. "COMP" means complementary condition constraint for the point. The mIoU is reported over 10 classes on A2D2.

semantic segmentation. Due to the challenge of aligning features for the 3D point clouds, the A^3 module is not included in the cross-modal setting.

Comparison with the SOTA. As shown in Table 2.9, similar to the image classification and the single mode semantic segmentation results, the SOTA cross-modal unsupervised adaptation method xMUDA [83] shows an obvious negative transfer effect, resulting in a performance drop for the 2D model, 3D model and the fused one. Furthermore, we designed reasonable baseline methods for comparison: 1) ES + MinEnt: the prediction from 2D and 3D networks are averaged in the target domain through the 2D and 3D point correspondence during training, and the fused prediction probability is optimized using the minimum entropy loss [184]. 2) ES + KL: the KL-divergence [83] is utilized to align between the 2D/3D prediction and the fused predictions for the corresponding points in the target domain, resp. 3) xMUDA + AKL: the KL-divergence alignment between 2D and 3D in the target domain is weighted adaptively, to reduce the wrong guidance from the unlabeled parts. 4) xMUDA + AKL + COMP: following baseline 3), another constraint, that the weights related to 2D and 3D need to be complementary, is added. It is shown that our method prevents negative transfer without the PSF component, outperforming the non-



(a) A2D2 (b) Ground Truth (c) 2D (Ours) (d) 3D (Ours)

Figure 2.6: Qualitative results of the cross-modal setting.

adapted baseline. Then by adding the PSF module, the 2D and 3D single-model performance is strongly improved, achieving 54.9% and 37.1%, resp. In Fig. 2.6, we show qualitative results in the target domain. The good performance proves the effectiveness of our method for the mDALU with partial modalities. This opens up the avenue to combine datasets collected with different sensors and offers the possibility of cheaply evaluating new combinations of sensors without annotating their data.

2.5 CONCLUSION

In this paper, we proposed the multi-source domain adaptation and label unification with partial datasets problem, called mDALU. Then we proposed a novel multi-stage approach for mDALU, including partially and fully supervised adaptation stages. Our approach is demonstrated through extensive experiments on different benchmarks.

3

META-LEARNING FOR OPEN COMPOUND DOMAIN ADAPTATION

This chapter corresponds to our published article:

Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. „Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation.“ In: *CVPR*. 2021

In this chapter, we investigate the open compound domain adaptive semantic segmentation (OCDA) problem, where target domain is modeled as a compound of multiple unknown homogeneous domains. It brings the advantage of improved generalization to unseen domains. To this end, we propose a principled meta-learning based approach to OCDA for semantic segmentation, MOCDA, by modeling the unlabeled target domain continuously. Our approach consists of four key steps. First, we **cluster** target domain into multiple sub-target domains by image styles, extracted in an unsupervised manner. Then, different sub-target domains are **split** into independent branches, for which batch normalization parameters are learnt to treat them independently. A meta-learner is thereafter deployed to learn to **fuse** sub-target domain-specific predictions, conditioned upon the style code. Meanwhile, we learn to online **update** the model by model-agnostic meta-learning (MAML) algorithm, thus to further improve generalization. We validate the benefits of our approach by extensive experiments on synthetic-to-real knowledge transfer benchmark, where we achieve the state-of-the-art performance in both compound and open domains.

3.1 INTRODUCTION

The traditional domain adaptation problem typically assumes the target domain as a single homogeneous domain. However, this assumption is not valid when the target domain images are collected under mixed, continually varying, and even unseen conditions. This gives rise to the challenge known as open compound domain adaptation [111]. The open compound domain adaptation treats the target as a compound

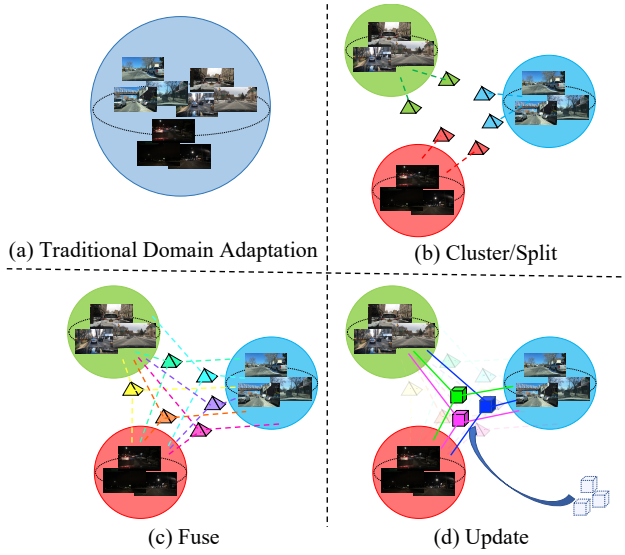


Figure 3.1: (a) The traditional unsupervised domain adaptation (UDA) vs. (b,c,d) the proposed meta-based open compound domain adaptation (MOCDA). Unlike the traditional UDA, MOCDA treats target as a compound of multiple unknown sub-domains. These sub-domains are discovered and processed using the cluster and the split module (b). The fuse module (c) then combines the sub-domain splits as basis (dash lines). On open domains, MOCDA adapts through online update during inference (blue arrow) in (d). Meta-learning serves in the fuse and the update module.

of multiple unknown sub-domains. Such assumption has been shown to be very promising by *Liu et al.* [111] for many practical settings of image classifications. However, the method developed in [111] does not fully exploit the same assumption for the task of image segmentation. [1] In this work, we show that the homogeneous sub-domain assumption can be exploited effectively also for image segmentation. We propose a novel meta-learning based approach to OCDA (abbreviated as MOCDA) that consists of four modules: cluster; split; fuse; and update, as illustrated in Fig. 3.1.

1 Open compound domain adaptation [111] does not fully exploit the domain information for segmentation task due to the inaccessibility of the domain encoder. Refer the original paper [111] for details.

Similar to OCDA, the proposed MOCDA utilizes two image sets for training from: a single labeled source domain; and a diverse unlabeled target domain, which is assumed to be a compound of multiple unknown sub-domains. Such an assumption is suitable for real challenging situations, where the target domain is a combination of many factors including diverse weather, cities, and acquisition time [141, 34, 120]. The considered learning setup not only performs domain adaptation to the compound target domain, but also has generalization potential to unseen open domains. In this context, the process of domain adaptation happens to exhibit a meta-behaviour [97, 5, 32], which learned dynamically makes the open world semantic segmentation possible. In this work, we show that the meta-behaviour of OCDA can be learned using (a) a hypernetwork for dynamic fusion of knowledge, and (b) the online update. On the one hand, the update process – which is carried out using the model-agnostic meta-learning strategy – creates an opportunity for better open set generalization with only one gradient step. On the other hand, the learned dynamic fusion allows images to appear from the continuous manifold of the compound target domain.

In essence, the proposed framework serves in following four steps. (i) From target images, style codes are extracted and grouped into multiple clusters. (ii) For each cluster, a set of batch normalization (BN) parameters are learned. (iii) Corresponding to each cluster, each image can have different domain-specific predictions. The hypernetwork, then, learns to fuse these predictions. (iv) Model-agnostic meta-learning (MAML) [45] is exploited during hypertraining process, endowing the online update ability of the model on open domain during inference stage. The key contributions of this chapter can be summarized as follows:

- We propose a novel framework for semantic segmentation in the OCDA setting. We use meta-learning in the dynamic fusion and MAML strategy based online update, to address the limitations of [111].

- We propose to model the compound target domain continuously, taking the sub-target domain as the basis, which offers the advantage of adapting to target domain and generalizing to unseen open domains.

- We demonstrate the adequacy of image style features, learned in an unsupervised manner, for our meta-based method MOCDA.
- The proposed method provides the state-of-the-art results in synthetic-to-real knowledge transfer benchmark datasets, for both compound and open domains.

3.2 RELATED WORKS

Unsupervised Domain Adaptation and Generalization. Our work is related to domain adaptation [155, 132, 173, 53, 182, 140] and domain generalization [97, 96, 103, 99] works. Unsupervised domain adaptation aims at training a model on the labeled source domain and transferring the learned knowledge to the unlabeled target domain. The traditional unsupervised domain adaptation works [114, 47, 115, 177] typically focus on solving adaptation problem from a single source domain to a single target domain. Even though being effective in several tasks, the single target domain assumption is still restricted in many practical applications. Recently, multiple-target domain adaptation problem [32, 52] has received increasing research interests. The problem investigates knowledge transfer to multiple unlabeled target domains. Yet another important aspect not prioritized by the classical domain adaptation methods is the knowledge transfer to unseen but related open domains [111, 60, 102].

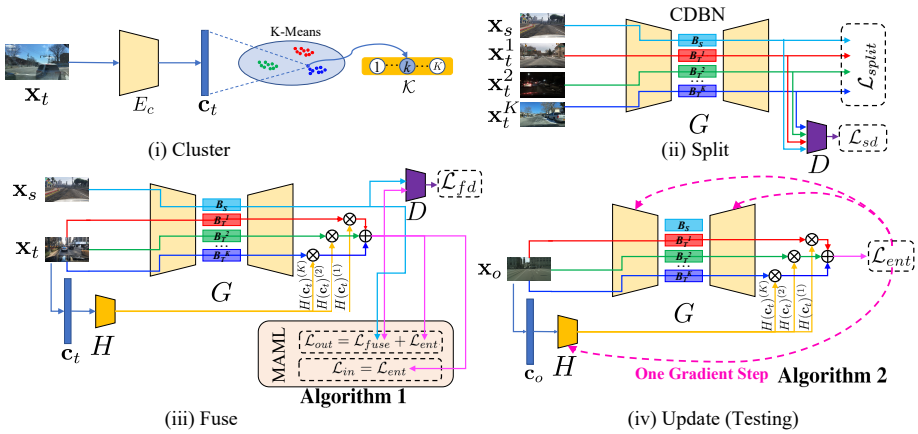
Cross-Domain Semantic Segmentation. In order to improve the adaptation and the generalization ability of the semantic segmentation model [20, 188, 151, 113, 22, 21], cross-domain semantic segmentation topic is extensively studied, both in the domain adaptation setting [214, 162, 227, 31, 29, 184] and in the domain generalization setting [183, 41, 60, 144, 111]. Most works either assume the target domain as a single domain [162, 214, 156, 72, 31, 29, 227], or a composition of multiple known domains [60, 213, 218, 144], with an exception of OCDA [111]. OCDA assumes target domain as a composition of multiple unknown domains, which is more realistic in practice. [111] follows a different approach for semantic segmentation compared to the classification task. The curriculum learning therefore is based on the average class confidence scores, rather than the neatly learned domain-focused factors in case of the classification task. Nevertheless, the experimental setup of our work is inspired by [111]. Concurrently, [136] develops the image

translation based method for the OCDA problem, which is complementary to our method. Besides the open domain in [111, 136], our work further explores the generalization ability of the model when facing more diverse extended open domains.

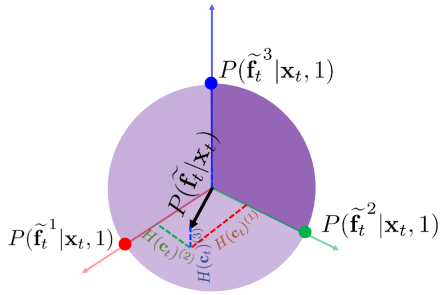
Meta-Learning for Domain Adaptation/Generalization. Meta-learning addresses the problem of learning to learn and has been successfully applied to various applications including image classification [66], image restoration [78], visual tracking [8], and network compression [104]. The principle of meta-learning [163, 70] has also been investigated for the task domain adaptation [143, 95, 32] and generalization [97, 5, 41], with the algorithmic advances [3, 45, 145]. Our work can be related to those works in terms of general methodology. Among those works, the ones most related are [32] and [209]. The similarities are : 1) both of [32] and our MOCDA study the domain adaptation problem when there are multiple unknown target domains through meta-learning. 2) both of [209] and our MOCDA aims at improving the domain generalization performance for semantic segmentation model, with the help of MAML strategy. However, we have significant differences in the following aspects: 1) [32] utilizes the meta-learner for clustering the target domain into different sub-target domains, and the target domain is modeled as a union of multiple sub-target domains. And [32] does not include the open domain. However, our meta-hypernetwork is utilized to fuse the knowledge from different clusters, to model the target domain as a continuous compound target domain. 2) [209] does not study the domain adaptation problem, and only focus on the domain generalization. The MAML strategy in [209] is only used during training stage on the well labeled source domain. By contrast, MOCDA utilizes the MAML strategy in both of the well labeled source domain and the unlabeled target domain during the training stage. During inference, the MAML strategy is exploited for online update.

3.3 THE MOCDA MODEL

Preliminaries. We consider that the labeled source domain $\mathcal{S}$ is composed of the source images $\mathbf{x}_s$, and the corresponding semantic labels $\mathbf{y}_s$, *i.e.*, $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s) | \mathbf{x}_s \in \mathbb{R}^{H \times W \times 3}, \mathbf{y}_s \in \mathbb{R}^{H \times W}\}$, where H, W are height and width of the image, respectively. In OCDA, the unlabeled target domain $\mathcal{T}$ consists of target images $\mathbf{x}_t^i$ from multiple homogeneous sub-target domains, $\mathcal{T}^i = \{\mathbf{x}_t^i | \mathbf{x}_t^i \in \mathbb{R}^{H \times W \times 3}\}, i = 1, \dots N$, where N is



(a) Framework Overview



(b) Compound Domain Modeling

Figure 3.2: (a) The overview of MOCDA framework demonstrating four modules; (i) Cluster, (ii) Split, (iii) Fuse, and (iv) Update. (b) Illustration of compound domain modeling, taking $K = 3$ for example. The sub-target domain $P(\widetilde{\mathbf{f}}_t^1|\mathbf{x}_t, 1)$, $P(\widetilde{\mathbf{f}}_t^2|\mathbf{x}_t, 2)$ and $P(\widetilde{\mathbf{f}}_t^3|\mathbf{x}_t, 3)$ is taken as basis. The cluster/split module models the compound target domain as the union set of three points, *i.e.*, red, green and blue points. But the fuse module models the compound target domain $P(\widetilde{\mathbf{f}}_t|\mathbf{x}_t)$ as the vector $H(\mathbf{c}_t) = [H(\mathbf{c}_t)^{(1)}, H(\mathbf{c}_t)^{(2)}, H(\mathbf{c}_t)^{(3)}]'$, composing the purple half quarter-spherical surface.

number of sub-target domains. In the context of this work (and also in OCDA), these sub-target domains are unknown. Therefore, the images $\mathbf{x}_t^i$ from some unknown sub-target domain $\mathcal{T}^i$ are simply denoted as $\mathbf{x}_t$, for notation convenience and clarity.

In this section, we propose the MOCDA model for semantic segmentation. The MOCDA model is composed of four modules: cluster, split, fuse, and update. The **Cluster** module extracts and clusters the style code from the target domain images automatically, dividing the target domain into multiple sub-target domains. The **Split** module adopts the compound-domain specific batch normalization (CDBN) layer to process different sub-target domain images using different branches. The **Fuse** module exploits a hypernetwork to predict the weights corresponding to each branch adaptively, conditioned on the style code of the input image. The final output of the network is the weighted combination of the outputs of different branches. The MAML method is utilized to train the **Fuse** module, so as to make the model be adapted quickly in **Update** module. Finally, the **Update** is carried out online during the inference time with one-gradient step, which is found to be beneficial for open domains. The framework overview is shown in Fig. 3.2. In the following, we provide the details of all four modules, separately.

3.3.1 *Cluster: Style Code Extraction and Clustering*

The aim of the cluster module is to cluster the target domain $\mathcal{T}$ into different sub-target domains $\mathcal{T}^k, k = 1, \ldots, K$, serving the OCDA's assumptions of unknown multiple sub-target domains of the target domain. As shown in [111, 81], the major differences of the target domain images due to varying conditions, such as the weather, lighting, and inter-dataset, can be effectively reflected by the style of the images. Our cluster module consists of two mappings; $E_c(\cdot)$ and $E_l(\cdot)$. $E_c(\cdot)$ maps the target domain $\mathcal{T}$ to the style code domain $\mathcal{C}_t = \{\mathbf{c}_t | \mathbf{c}_t \in \mathbb{R}^l\}$ as $E_c : \mathcal{T} \rightarrow \mathcal{C}_t$, where l is the dimension of the style code. More specifically, the target domain image $\mathbf{x}_t$ is mapped to a low-dimension style code $\mathbf{c}_t = E_c(\mathbf{x}_t)$. Then a clustering algorithm, K-means [112], is adopted to automatically cluster the style code domain $\mathcal{C}_t$, partitioning into K clusters with centroids $\{\mathbf{c}_t^k\}$. We use the mapping $E_l(\cdot)$ to assign $\mathbf{x}_t$ to one of the sub-target domains, represented by the set $\mathcal{K} = \{k | k = 1, \ldots, K\}$, as $E_l : \mathcal{T} \rightarrow \mathcal{K}$. Here, we adopt the nearest neighbor strategy for $E_l(\cdot)$. More specifically, each target image is

assigned to the nearest cluster, using the Euclidean distance between style codes of the image and the centroids, given by,

$$E_l(\mathbf{x}_t) := \underset{k}{\operatorname{argmin}} \|\mathbf{c}_t - \mathbf{c}_t^k\|. \tag{3.1}$$

The key of our cluster module is to find an adequate mapping $E_c(\cdot)$. In this work, the unsupervised image translation framework MUNIT [81] is trained to translate between the source domain $\mathcal{S}$ and the target domain $\mathcal{T}$. During the translation training process, the style code encoder of MUNIT is trained to extract the style code from images unsupervisedly. The trained style encoder of MUNIT is used as $E_c(\cdot)$. Then, the target domain $\mathcal{T}$ is clustered into K sub-target domains $\mathcal{T}^k$, where the number of sub-target domains K is a hyperparameter. Using the nearest neighbour search, refer Eq. (3.1), each target image $\mathbf{x}_t$ is assigned to one of the sub-target domains $\mathcal{T}^k$. Henceforth, the image $\mathbf{x}_t$ assigned image k^{th} cluster is denoted as $\mathbf{x}_t^k$.

3.3.2 *Split: Domain-Specific Batch Normalization*

In [17], the domain-specific batch normalization (DSBN) is shown to be beneficial for the unsupervised domain adaptation (UDA), by separating the batch normalization layer for the source and target domain.

Similar to DSBN for UDA, the aim of our split module is to separate the multiple sub-target domain-specific information from the domain-invariant information. We propose DSBN for OCDA (abbreviated as CDBN), to conduct such separation for source domain $\mathcal{S}$ and the multiple (clustered) sub-target domains $\{\mathcal{T}^k\}$. Note that DSBN for UDA learns only two sets of BN parameters (with possible extension given more labeled domains). However, the proposed CDBN learns $K+1$ sets of BN parameters for source domain and multiple *unlabeled* sub-target domains, *i.e.*, $B_S, B_T^1, ..., B_T^K$, formulated as,

$$B_S(\mathbf{x}_s, \mu_s, \sigma_s, \beta_s, \gamma_s) = \gamma_s \frac{\mathbf{x}_s - \mu_s}{\sigma_s} + \beta_s, \tag{3.2}$$

$$B_T^k(\mathbf{x}_t^k, \mu_t^k, \sigma_t^k, \beta_t^k, \gamma_t^k) = \gamma_t^k \frac{\mathbf{x}_t^k - \mu_t^k}{\sigma_t^k} + \beta_t^k, \tag{3.3}$$

where k is the sub-target domain label, $k = 1, ..., K$. Our split module replaces BN layers by CDBN. As shown in Fig. 3.2, our split

module includes the multi-branch semantic segmentation network $G = \{G_s, G_1, ..., G_K\}$ and the discriminator D. G_k is formed by selecting the k-th branch B_k of the CDBN layer. Through the adversarial learning, the discriminator D aligns the prediction distributions of source domain and that of the sub-target domains, in the output space. Therefore, the full optimization objective of the split module includes the semantic segmentation loss and the adversarial loss, presented below.

Semantic Segmentation Loss. We train the semantic segmentation network G with a standard cross entropy loss, using the source domain image $\mathbf{x}_s$ and the associated ground truth label $\mathbf{y}_s$,

$$\mathcal{L}_{seg}(G) = -\frac{1}{HW}\sum_{n=1}^{HW}\sum_{m=1}^{M} y_s^{(n,m)} \log(G_s(\mathbf{x}_s)^{(n,m)}), \tag{3.4}$$

where (n, m) represents (pixel, class) indices for M classes.

Multi-Branch Adversarial Loss. Recall the cluster module, each target image $\mathbf{x}_t$ is assigned to a unique sub-target domain label k, *i.e.*, $\mathbf{x}_t^k$. Here in the split module, the image $\mathbf{x}_t^k$ is processed using only the corresponding branch G_k, *i.e.*, $G_k(\mathbf{x}_t^k)$. Our multi-branch adversarial loss is an extension of the adversarial loss [175], which aligns the prediction distributions of the source domain $G_s(\mathbf{x}_s)$, and the sub-target domains $\{G_k(\mathbf{x}_t^k)\}$. The multi-branch adversarial loss $\mathcal{L}_{sadv}$ and the corresponding discriminator training loss $\mathcal{L}_{sd}$ are formulated as,

$$\mathcal{L}_{sadv}(G) = -\sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}_t^k \sim P_{T^k}} \log(D(G_k(\mathbf{x}_t^k))^{(n,1)}), \tag{3.5}$$

$$\begin{aligned}\mathcal{L}_{sd}(D) = &-\mathbb{E}_{\mathbf{x}_s \sim P_S} \log(D(G_s(\mathbf{x}_s))^{(n,1)}) \\ &-\sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}_t^k \sim P_{T^k}} \log(D(G_k(\mathbf{x}_t^k))^{(n,0)}),\end{aligned} \tag{3.6}$$

where P_S and P_{T^k} are the underlying data distributions of $\mathcal{S}$ and $\mathcal{T}_k$, respectively. The following full optimization objective is used for training our split module,

$$\mathcal{L}_{split}(G) = \mathcal{L}_{seg}(G) + \lambda_1 \mathcal{L}_{sadv}(G), \tag{3.7}$$

where λ_1 is a trades-off parameter. During the training process, we alternatively optimize the discriminator D and the generator G with the objective in the Eq. (3.6) and the Eq. (3.7), respectively.

3.3.3 *Fuse: HyperNetwork for Branches Fusion*

The **cluster** and **split** module discretizes the target domain into a few clusters, providing an initial discrete modeling of the target domain. The fuse of the discretized modes forms continuous manifold, the sample on which reflects the continuous change of the target domain and might correspond to an unseen domain. In the **fuse** module, we learn to combine the sub-target domain to model the compound target domain continuously.

Compound Domain Modelling. Here we model the target domain $\mathcal{T}$ in the corresponding feature domain $\mathcal{F}$, which is mapped by $F : \mathcal{T} \rightarrow \mathcal{F}$. Let $P(\widetilde{\mathbf{f}}_t^k|\mathbf{x}_t, k)$ be the feature distribution corresponding to image $\mathbf{x}_t$ when assumed to be from the k^{th} cluster. Then the distribution of the feature $\widetilde{\mathbf{f}}_t$ of the image $\mathbf{x}_t$, *i.e.*, $P(\widetilde{\mathbf{f}}_t|\mathbf{x}_t)$, is expressed as,

$$P(\widetilde{\mathbf{f}}_t|\mathbf{x}_t) = \sum_{k=1}^{K} P(\widetilde{\mathbf{f}}_t^k, k|\mathbf{x}_t) = \frac{1}{N}\sum_{k=1}^{K} P(k|\mathbf{x}_t)P(\widetilde{\mathbf{f}}_t^k|\mathbf{x}_t, k) \tag{3.8}$$

where $N = \int_{\widetilde{\mathbf{f}}_t^k} \sum_{k=1}^{K} P(\widetilde{\mathbf{f}}_t^k|\mathbf{x}_t, k)P(k|\mathbf{x}_t)\mathrm{d}\widetilde{\mathbf{f}}_t^k$. $P(k|\mathbf{x}_t)$ describes the probability distribution of the sub-target domain's label of image $\mathbf{x}_t$. By taking the sub-target domain distributions $P(\widetilde{\mathbf{f}}_t^k|\mathbf{x}_t, k)$ as basis, the compound target domain can be modeled with the vector, *i.e.*, $\{[P(1|\mathbf{x}_t), ..., P(k|\mathbf{x}_t), ..., P(K|\mathbf{x}_t)]'\}$.

HyperNetwork for Branches Fusion. In essence, the cluster and split module can be seen as modeling the sub-target domain label distribution as $P(k|\mathbf{x}_t) = 1$, if $E_l(\mathbf{x}_t) = k$ and $P(k|\mathbf{x}_t) = 0$, if $E_l(\mathbf{x}_t) \neq k$. It models the compound target domain as the discretized points in the vector space, as illustrated in Fig. 3.2. In order to model the compound target domain in the continuous space, in our fuse module, we adopt the categorical distribution for $P(k|\mathbf{x}_t)$, *i.e.*,

$$P(k|\mathbf{x}_t) = w_k, \quad \text{with,} \quad \sum_{k=1}^{K} w_k = 1, w_k > 0, \tag{3.9}$$

where $\mathbf{w} = [w_1, ..., w_k, ..., w_K]^\top$ is the K-dimensional categorical vector, whose element w_k represents the probability that the target image $\mathbf{x}_t$ belongs to the sub-target domain $\mathcal{T}_k$. Then the hypernetwork $H(\cdot)$ is adopted to learn the $P(k|\mathbf{x}_t)$, by taking the style code $\mathbf{c}_t$ of the image

sample $\mathbf{x}_t$ as input, *i.e.*, $[w_1, ..., w_k, ..., w_K]^\top = H(\mathbf{c}_t)$. Substituting the $H(\mathbf{c}_t)$ in Eq. (3.8), the feature distribution $P(\widetilde{\mathbf{f}}_t|\mathbf{x}_t)$ can be derived as,

$$P(\widetilde{\mathbf{f}}_t|\mathbf{x}_t) \sim \sum_{k=1}^{K} H(\mathbf{c}_t)^{(k)} P(\widetilde{\mathbf{f}}_t^k|\mathbf{x}_t, k). \tag{3.10}$$

where $H(\mathbf{c}_t)^{(k)}$ is the k^{th} element of $H(\mathbf{c}_t)$. Eq. (3.10) shows that the compound target domain is modeled in the continuous vector space, $H(\mathbf{c}_t)$, taking the sub-target domain distributions $P(\widetilde{\mathbf{f}}_t^k|\mathbf{x}_t, k)$ as basis, as illustrated in Fig. 3.2.

From above, it is shown that $H(\mathbf{c}_t)$ weights the different sub-target domain distribution differently to get the compound target domain distribution. Here we adopt the network G as our mapping $\mathcal{F}$. Following [80], we reweight each feature sample $\widetilde{\mathbf{f}}_t^k = G_k(\mathbf{x}_t)$ with $H(\mathbf{c}_t)$, so that the feature sample from dominant sub-target domain has higher weight, whereas the sample from non-dominant sub-target domain has lower weight. The final prediction can be represented as,

$$\widetilde{\mathbf{y}}_t = \sum_{k=1}^{K} H(\mathbf{c}_t)^{(k)} G_k(\mathbf{x}_t). \tag{3.11}$$

By combining Eq. (3.11) and Eq. (3.5), the adversarial loss for the fuse module $\mathcal{L}_{fadv}$ and the corresponding discriminator training loss $\mathcal{L}_{fd}$ can be formulated as,

$$\mathcal{L}_{fadv}(G, H) = -\mathbb{E}_{\mathbf{x}_t \sim P_T} \log(D(\widetilde{\mathbf{y}}_t)^{(n,1)}) \tag{3.12}$$

$$\begin{aligned}\mathcal{L}_{fd}(D) = &-\mathbb{E}_{\mathbf{x}_s \sim P_S} \log(D(G_s(\mathbf{x}_s))^{(n,1)}) \\ &-\mathbb{E}_{\mathbf{x}_t \sim P_T} \log(D(\widetilde{\mathbf{y}}_t)^{(n,0)}).\end{aligned} \tag{3.13}$$

The optimization objective of our fuse module is a combination of Eq. (3.4) and Eq. (3.12), which is given by,

$$\mathcal{L}_{fuse}(G, H) = \mathcal{L}_{seg}(G) + \lambda_2 \mathcal{L}_{fadv}(G, H), \tag{3.14}$$

where λ_2 is the hyperparameter to balance between the adversarial loss and the segmentation loss. During the training process, we alternatively optimize the discriminator D and the generator G, the hypernetwork H with the objective in the Eq. (3.13) and the Eq. (3.14), respectively. In our MOCDA model, the training of the fuse module is combined with the MAML strategy, which is explained further in Section 3.3.4 and Algorithm 3.1.

3.3.4 *Update: MAML based Online Update*

In the previous OCDA work [111], the open set is only treated as a testing set to verify the generalization ability of the model. In contrast, in our work, the open set is also used for updating the model online during testing, for better generalization to the unseen domain, realized by MAML.

MAML. The MAML strategy [45] aims at learning the optimal model parameters θ^*, which eases the adaptation process for new tasks. In each iteration of MAML, there are two training loops; inner and outer. Let the data of inner and outer loops be $\mathcal{D}_{in}$ and $\mathcal{D}_{out}$, respectively. In each training iteration, the model parameters θ are first updated with the inner loop loss $\mathcal{L}_{in}$ and data $\mathcal{D}_{in}$. The updated model is then evaluated on the outer loop loss $\mathcal{L}_{out}$ and data $\mathcal{D}_{out}$, to test the generalization ability of the updated model. Furthermore, the evaluation performance $\mathcal{L}_{out}$ is also adopted during update, to better generalize the model. This nested training fashion mimics the training and testing phase of the model. In order to endow adaptation ability, the optimization objective of MAML is formulated as,

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\ \mathcal{L}_{out}(\theta - \alpha \nabla \mathcal{L}_{in}(\theta,\ \mathcal{D}_{in}),\ \mathcal{D}_{out}), \tag{3.15}$$

where α is the learning rate for updating the model.

MAML for OCDA. In our addressed problem of OCDA for semantic segmentation, images from the set $\{\mathbf{x}_o\}$ of the unseen open domain $\mathcal{O}$ are available only during testing. We adopt the MAML algorithm in our MOCDA during training to be combined with the fuse module. MAML then offers us the advantage of quick adaptation to the open set during testing, by means of online update within one gradient step.

In the inner loop, we sample data from the target domain $\mathcal{T}$, *i.e.*, $\mathcal{D}_{in} = \{\mathbf{x}_t\}$. Meanwhile, in order to update the model without supervision, we use the unsupervised self-entropy loss [184] $\mathcal{L}_{ent}$ as the inner loop loss $\mathcal{L}_{in}$ – which mimics the model update process during testing, given by,

$$\mathcal{L}_{in} = \mathcal{L}_{ent} = \ -\frac{1}{HW}\sum_{n=1}^{HW}\sum_{c=1}^{C} \widetilde{\mathbf{y}}_t^{(n,c)} \log \widetilde{\mathbf{y}}_t^{(n,c)}. \tag{3.16}$$

In the outer loop, the data is sampled from both source domain $\mathcal{S}$ and the target domain $\mathcal{T}$, *i.e.*, $\mathcal{D}_{out} = \{\mathbf{x}_s, \mathbf{y}_s, \mathbf{x}_t\}$. In order to evaluate the

Algorithm 3.1 MAML algorithm for OCDA (Training)

Require: Source data $\mathcal{S} = \{(\mathbf{x}_s, \mathbf{y}_s)\}$, target data $\mathcal{T} = \{\mathbf{x}_t\}$, segmentation network G, hypernetwork H, discriminator D, the learning rate α of G, H, and the learning rate ζ of discriminator D.

1: Initialize the parameters θ_{GH} and θ_D, respectively of the segmentation network G, hypernetwork H, and the discriminator D;
2: **while** not done **do**
3: Sample $\mathcal{D}_{in}$ from $\mathcal{T}$ ▷ Inner Loop
4: $\theta^+_{GH} \leftarrow \theta_{GH} - \alpha \nabla_{\theta_{GH}} \mathcal{L}_{in}(\mathcal{D}_{in}, \theta_{GH})$;
5: Sample $\mathcal{D}_{out}$ from $\mathcal{S}$ and $\mathcal{T}$ ▷ Outer Loop
6: $\theta_{GH} \leftarrow \theta_{GH} - \alpha \nabla_{\theta_{GH}} \mathcal{L}_{out}(\mathcal{D}_{out}, \theta^+_{GH})$;
7: $\theta_D \leftarrow \theta_D - \zeta \nabla_{\theta_D} \mathcal{L}_{fd}(\mathcal{D}_{out}, \theta_D)$;
8: **end while**

model's performance on different domains and in different way, the outer loop loss $\mathcal{L}_{out}$ uses the optimization objective of the fuse module in Eq. (3.14) and the self-entropy loss in Eq. (3.16), such that,

$$\mathcal{L}_{out} = \mathcal{L}_{fuse} + \delta \mathcal{L}_{ent}, \tag{3.17}$$

where δ is the hyperparameter to balance between the fuse module loss and the unsupervised self-entropy loss. The MAML algorithm used during OCDA training is presented in Algorithm 3.1. Similarly, the MAML used during the online update, of OCDA testing, is given in Algorithm 3.2.

3.3.5 *Training Protocol of MOCDA*

In total, our MOCDA model is trained in the multi-stage way, consisting of three steps: i) training the MUNIT model for style code extraction and clustering, ii) training with the CDBN layer in split module, iii) the CDBN layer is frozen, adding the hyper-network and the fuse module, and training the hypernetwork H and fine-tuning the semantic segmentation network G with MAML strategy as described in Algorithm 3.1. Then during testing stage, our whole model, except for CDBN layer, is online updated with the MAML strategy as clarified in Algorithm 3.2.

Algorithm 3.2 MAML algorithm for OCDA (Testing)

Require: Data $\{\mathbf{x}_o\}$ from the unseen novel domain $\mathcal{O}$, segmentation network G, hypernetwork H.

1: Use trained parameters θ_{GH} of the segmentation network, G and the hypernetwork H, from the training phase;
2: $F \leftarrow 0$
3: **for** $i = 1, ..., n$ **do**
4: Sample the i^{th} image $\mathbf{x}_o^i$ from $\{\mathbf{x}_o\}$;
5: $\widetilde{\mathbf{y}}_o^i \leftarrow G(\mathbf{x}_o^i)$;
6: $\theta_{GH} \leftarrow \theta_{GH} - \eta \nabla_{\theta_{GH}} \mathcal{L}_{ent}(\widetilde{\mathbf{y}}_o^i, \theta_{GH})$
7: **end for**

3.4 EXPERIMENTS

In this section, we demonstrate the benefits of our MOCDA model under the open compound domain adaptive semantic segmentation setting. We compare our MOCDA model with other state-of-the-art (SOTA) methods on both of the target domain and the open domain. In order to further prove the effectiveness of our MOCDA model for open domain with online update, we introduce more diverse and challenging extended open domains to test the model performance additionally.

3.4.1 *Experiments Setup*

Implementations. *Cluster.* In the cluster module, we train the MUNIT [81] model to translate between the source domain images and the compound target domain images in the unsupervised way. We follow the experimental set up in the urban scene image translation set up in MUNIT [81]. The shortest side of the images are firstly resized to 512, and then the images are randomly cropped with the size of 400×400. The loss weights for image reconstruction loss, style reconstruction loss, content reconstruction loss, and domain-invariant perceptual loss are set as 10, 1, 1, and 1, respectively. The Adam optimizer [88] is adopted with $\beta_1 = 0.5, \beta_2 = 0.999$, and the learning rate is set as 0.0001. Also, the dimension of the style code is set as 8. The number of the clusters K is set as 4. *Split, Fuse, and Update.* In the split and fuse module, we have the semantic segmentation network and the discriminator. We adopt

the DeepLab-VGG16 [20, 166] with synchronized batch normalization layer [82] for the semantic segmentation network. And we adopt the discriminator structure in [175]. The compound target domain images and the open domain images, from BDD100K [203], Cityscapes[34], WildDash [207] and KITTI [2], are resized to 1024×512, and the source domain images from GTA5 [149] and SYNTHIA-SF [69] are resized to 1280×720. The λ_1 in Eq. (3.7), and λ_2 in Eq. (3.14) are set as 0.001. In the update module, during the training stage, the δ in Eq. (3.17) is set as 0.0001. In the split, fuse and update module, we adopt the SGD optimizer to train the hypernetwork and the semantic segmentation network, where the momentum is 0.9 and the weight decay is 5×10^{-4}. The learning rate is set as 2.5×10^{-4}, and uses the polynomial decay strategy with power of 0.9 as done in [175]. We keep the same learning rate for online updating the hypernetwork and the semantic segmentation network. Also, we adopt the Adam optimizer [88] for training the discriminator with $\beta_1 = 0.9, \beta_2 = 0.99$. The learning rate is set as 1.0×10^{-4} and uses the polynomial decay strategy with power of 0.9. And our MOCDA model is implemented with PyTorch [138].

Datasets. Following [111], we adopt the synthetic image dataset GTA5 [149] or SYNTHIA-SF [162] as the source domain, the rainy, snowy, and cloudy images in BDD100K [203] as the target domain, while the overcast images in BDD100K are utilized as the open domain. Besides, more diverse images from other real image datasets, Cityscapes[34], KITTI[2] and WildDash [207] are introduced as extended open domains. The datasets are detailed as follows. 1) *GTA5.* GTA5 [149] is a synthetic urban scene image dataset, rendered from game engine. The scene of the GTA5 images is based on the city of Los Angeles. The GTA5 dataset covers 24966 densely labeled images, the annotation of which is compatible with that of Cityscapes. In OCDA benchmark, GTA5 $\rightarrow$ BDD100K, the GTA5 images, with the ground truth label, serve as source domain. 2) *SYNTHIA-SF.* SYNTHIA-SF [69] is a synthetically rendered image dataset from virtual city. There are 2224 images in the SYNTHIA-SF dataset, featuring different scenarios and traffic conditions. The images are densely labeled and the labels are compatible with Cityscapes. In our OCDA benchmark, SYNTHIA-SF $\rightarrow$ BDD100K, the SYNTHIA-SF dataset and the associated ground truth label serve as the source domain. 3) *BDD100K.* BDD100K [203] is a real urban scene image dataset, mainly taken from US cities. And the images in BDD100K dataset are diverse in different aspects such as weather and

environment. We adopt the C-driving subset of BDD100K proposed in [111], which is composed of rainy, snowy, cloudy and overcast images. During training stage, 14697 images, without the ground truth label, are used as the unlabeled compound target domain, including rainy, snowy and cloudy weather images. All different weather images are mixed and not assigned the weather information. During the testing stage, 803 images covering rainy, snowy and cloudy weather, with ground truth semantic annotation, are used as the validation set of the compound target domain, for evaluating the adaptation performance of the model. Besides, during the testing stage, 627 images with the ground truth semantic label, containing overcast weather, are taken as the validation set of the open domain, for evaluating the generalization performance of the model. The semantic label of the BDD100K dataset is compatible with that of Cityscapes. 4) *Cityscapes.* Cityscapes [34] is a real street scene image dataset, collected from different European cities. In our OCDA benchmark, during the testing stage, the validation set of Cityscapes, covering 500 densely labeled images, is used as one of the extended open domains to evaluate the generalization ability of the model. 5) *KITTI.* KITTI [2] covers the real urban scene images, taken from the mid-size European city, Karlsruhe. In our OCDA benchmark, the validation set of KITTI, including 200 densely labeled images, is used as one of the extended open domains for generalization ability evaluation during the testing stage. The ground truth label of KITTI dataset is compatible with that of Cityscapes. 6) *WildDash.* WildDash [207] is a dataset covering images from diverse driving scenarios under the real-world conditions. The images in WildDash possess the diversity in different aspects, such as the time, weather, data sources and camera characteristics. In our OCDA benchmark, during the testing stage, the validation set of WildDash, containing 70 Cityscapes annotation compatible images, serves as one of the extended open domains for measuring the generalization performance of the model.

3.4.2 *GTA5 to BDD100K*

Comparison with SOTA. In Table 3.1, we present our open compound domain adaptation results, in comparison with other SOTA methods. For fair comparison, all of the methods adopt the DeepLab-VGG16 model with the batch normalization layer. Compared with our baseline method AdaptSegNet[175], our split module achieves 3.1% and

(a) Clustering visualization (b) Example images of different clusters

Figure 3.3: Visualization of clustering results. (a) is the t-SNE visualization of the style code extracted by the cluster module, (b) is example images from different clusters.

2.4% gain on the target domain and the open domain, respectively. Compared with the SOTA method OCDA[111], our split module performance outperforms by 0.9% on the target domain and by 1.6% on the open domain. It proves the effectiveness of our cluster module and the split module, for sub-target domain discovery and sub-target domain-specific information disjointing. The clustering visualization is shown in Fig. 3.3. Then by adopting the meta-learning with the hypernetwork and the MAML training strategy in the fuse module, our MOCDA model achieves the state-of-the-art performance, which improves the split module performance by 2.3% from 25.4% to 27.7%, and by 1.9% from 29.5% to 31.4% on the target domain and the open domain, respectively. It proves the advantage of our MOCDA model on fusing the different sub-target domains knowledge, modeling the target domain continuously through the hypernetwork, and adopting the MAML training strategy. The qualitative comparison of the semantic segmentation results on the target domain is shown in Fig. 3.6.

Online Update. Another meta-learning paradigm in our MOCDA model, besides the fuse module, is the MAML algorithm based online update during testing stage. From Table 3.2, it is shown that our MOCDA model without online update outperforms the baseline method AdaptSegNet [175] on both of the open domain and the extended open domain by 5.6% in average. It proves the effectiveness of our cluster, split and fuse module for open domain generalization. By further using the MAML based online update strategy described in Algorithm 3.2 during the testing stage, our MOCDA model performance on all the open domains improves by 0.7% in average, from 28.1% to 28.8%. Our model w/ or w/o online update has the same

Source	Compound			Open	Avg	
GTA→	Rainy	Snowy	Cloudy	Overcast	C	C+O
Source Only[111]	16.2	18.0	20.9	21.2	18.9	19.1
Source Only *	19.7	18.4	20.5	22.5	19.7	21.0
AdaptSegNet[111]	20.2	21.2	23.8	25.1	22.1	22.5
AdaptSegNet[175] *	21.6	20.5	23.9	27.1	22.3	24.4
CBST[227]	21.3	20.6	23.9	24.7	22.2	22.6
IBN-Net[133]	20.6	21.9	26.1	25.5	22.8	23.5
PyCDA [106]	21.7	22.3	25.9	25.4	23.3	23.8
OCDA [111]	22.0	22.9	27.0	27.9	24.5	25.0
Ours (Split)	23.5	23.5	27.8	29.5	25.4	27.1
Ours (Fuse)	**24.4**	**27.5**	**30.1**	**31.4**	**27.7**	**29.4**

Table 3.1: Semantic segmentation performance comparison with SOTA: GTA→ BDD100K with DeepLab-VGG16 backbone. The results are reported on mIoU over 19 classes. * means our reproduced result.

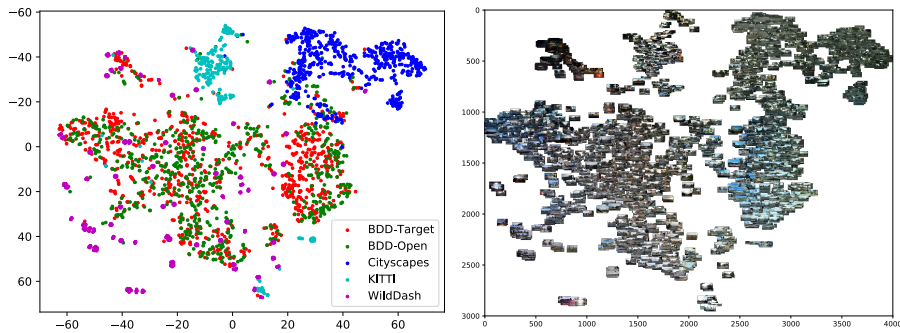


Figure 3.4: Extended open domains, open domain and target domain style code t-SNE visualization. The domain gap between the BDD100K open domain image and the target domain image (red and green points) is narrow due to the similar style. Our introduced extended open domain Cityscapes, KITTI and WildDash images have much larger domain gap from the BDD100K images. And the style code extracted by our cluster module can effectively reflect the domain gap.

performance on the open domain, BDD100K overcast image. It is due to that the BDD100K overcast image is still from the BDD100K dataset, and the style gap between the overcast image and the target domain image is very narrow, whose visualization is shown in Fig. 3.4. The benefit from our cluster, split and fuse module has been able to han-

Source GTA→	Open BDD	Extended Open Cityscapes	KITTI	WildDash	Avg
Source[111]	21.2	–	–	–	–
Source*	22.5	19.3	24.1	16.0	20.5
AdaptSegNet[111]	25.1	–	–	–	–
AdaptSegNet[175] *	27.1	22.0	23.4	17.5	22.5
w/o Online Update	31.4	30.4	29.8	20.6	28.1
w/ Online Update	31.4	31.1	30.9	21.6	**28.8**
Gain of Online Update	–	+0.7	+1.1	+1.0	+0.7

Table 3.2: Open domain semantic segmentation performance comparison w/ or w/o online update: GTA→ BDD100K with DeepLab-VGG16 backbone. The results are reported on mIoU over 19 classes. * means our reproduced result.

dle the narrow style gap and have good generalization performance already. The performance gain, 0.7%, 1.1% and 1.0% on the extended open domains where the style gap is much larger, Cityscapes, KITTI and WildDash dataset, proves that the MAML based meta-learning paradigm, in Algorithm 3.1 for training and Algorithm 3.2 for testing, endows the fast adaptation ability to our model to generalize better on open domains. The qualitative comparison, w/ or w/o online update, on the open domains are shown in Fig. 3.6.

Ablation Study. We show the comparison of ablations and different variants of our model in Table 5.6. From Table 5.6, it is shown that all the modules, the cluster/split module ($\mathcal{L}_{split}$), the fuse module ($\mathcal{L}_{fadv}$) and the MAML training strategy are helpful to our whole MOCDA model. The cluster and split module has been proven to be helpful in the comparison with AdaptSegNet[175] and other SOTA methods. Here we show the effectiveness of our meta-learning paradigm, the hypernetwork and the MAML training strategy through the ablations and variants methods comparison. Firstly, in order to prove the validity of our hypernetwork, we build the baseline methods of the branch fusion in non-adaptive way; 1), averagely fuse for prediction during the testing stage of the split module. 2), averagely fuse during the training and testing stage of the fuse module. 3) use the style code distance from different clusters to weight different branches during the training and testing stage of the fuse module. It is shown that our hypernetwork based branch fusion strategy performance, 27.1%, outperforms all other non-adaptive fusion strategy, 23.1%, 26.1%, 26.6%.

$\mathcal{L}_{seg}$	$\mathcal{L}_{adv}$	$\mathcal{L}_{sadv}$	$\mathcal{L}_{fadv}$	$\mathcal{L}_{ent}$	MAML	mIoU
✓						18.9
✓	✓					22.3
✓		✓				25.4
✓		✓				23.1†
✓			✓			26.1‡
✓			✓			26.6§
✓			✓			27.1
✓			✓	✓		27.3
✓			✓	✓	✓	**27.7**

Table 3.3: Different ablations and variants comparison for OCDA, tested on BDD100k target domain based on DeepLab-VGG16 with batch normalization layer backbone. The results are reported on mIoU over 19 classes. † represents the average fusion only during testing. ‡ represents the average fusion of different branches during training and testing. § represents the style code distance weighted fusion during training and testing.

It benefits from the advantage of adaptive weights predicted from the hypernetwork conditioned on the image sample style code. Secondly, by comparing the performance of training the fuse module using the $\mathcal{L}_{out}$ in the Eq. (3.17) and purely using the $\mathcal{L}_{fuse}$ in Eq.(3.14), it is shown that there is 0.2% performance gain by adding the unsupervised entropy loss, from 27.1% to 27.3%. By further introduce the MAML training strategy in Algorithm 3.1 for the fuse module, as done in our MOCDA model, the performance can be further improved to 27.7%. It proves that the MAML training strategy is not only helpful to the open domain generalization as described above, but also is beneficial to improve the adaptation performance of the model on the target domain. It results from that MAML training strategy mimics the training and testing procedure with the outer loop and inner loop and makes the model more domain adaptive.

Hypernetwork prediction. Besides ablation study and the variants of our model, we provide additional t-SNE [119] visualization of our hypernetwork prediction to prove the validity of the hypernetwork in our MOCDA model. As shown in Fig. 3.5, for the image samples from different sub-target domains, our hypernetwork prediction possesses

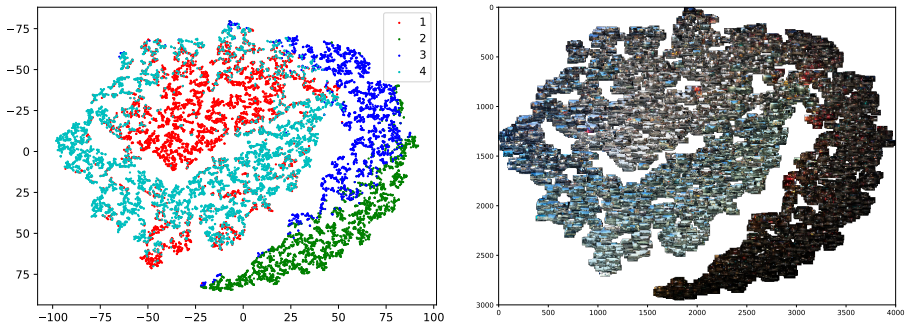


Figure 3.5: t-SNE visualization of hypernetwork prediction. For image samples belonging to different sub-target domains $1, 2, 3, 4$, our hypernetwork prediction shows different attributes even though we do not explicitly input the sub-target domain information during the fuse module training, which proves the validity of our hypernetwork.

different feature attributes, even though we do not explicitly provide the sub-target domain information in this process. It proves that our hypernetwork is able to adaptively adjust the prediction, conditioned on the style code of the image samples.

3.4.3 *SYNTHIA-SF to BDD100K*

In this section, SYNTHIA-SF is used as the source domain. Following [223], we only take 11 main classes in the SYNTHIA-SF dataset to measure the semantic segmentation performance, which are road, sidewalk, building, wall, fence, pole, light, vegetation, sky, person and car.

Comparison with SOTA. In Table 3.4, we report the quantitative comparison results between our MOCDA model and other SOTA methods for the open compound domain adaptation setting, from the SYNTHIA-SF to the BDD100K. From Table 3.4, it is shown that our MOCDA model outperforms MinEnt [184] and AdaptSegNet [175] on both of the target domain and the open domain. It further verifies the effectiveness of our MOCDA model for OCDA.

Online Update. In Table 3.5, the performance of our MOCDA model for the open domain and the extended open domain are shown. Our MOCDA model w/o online update outperforms the AdaptSegNet

Source	Compound			Open	Avg	
SYNTHIA-SF→	Rainy	Snowy	Cloudy	Overcast	C	C+O
Source Only	16.5	18.2	21.4	20.6	19.2	19.8
MinEnt[184]	21.8	22.6	26.2	25.7	23.9	24.7
AdaptSegNet[175]	24.9	26.9	30.7	30.3	28.0	29.0
Ours (Split)	25.2	27.9	32.4	31.8	29.1	30.3
Ours (Fuse)	**26.6**	**30.0**	**33.0**	**32.6**	**30.4**	**31.4**

Table 3.4: Semantic segmentation performance comparison with SOTA: SYNTHIA-SF→ BDD100K with DeepLab-VGG16 backbone. The results are reported on mIoU over 11 classes. The best results are denoted in bold.

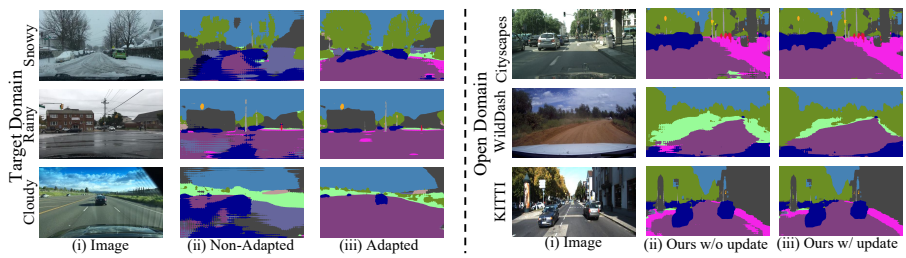


Figure 3.6: Qualitative comparison of semantic segmentation results on the target domain, including the rainy, snowy and cloudy weather, and on the open domains, KITTI, WildDash and Cityscapes.

method by 2.2% in average on all the open domains. By further utilizing the online update in the open domain, the peformance can be further improved by 1.1% in average, from 30.1% to 31.2%. It further proves the validity of the online update for the open domain.

3.5 CONCLUSION

In this chapter, we address the problem of open compound domain adaptation, and propose a meta-learning based model, MOCDA. MOCDA is composed of four modules, cluster, split, fuse and update module. Meta-learning serves in fuse and update module for continuously modeling the compound target domain and online update. The extensive experiments show that our model achieves the state-of-the-art per-

Source SYNTHIA-SF→	Open BDD	Extended Open Cityscapes	KITTI	WildDash	Avg
Source	20.6	24.7	20.7	17.3	20.8
AdaptSegNet[175]	30.3	35.9	24.7	20.7	27.9
w/o Online Update	32.6	29.9	33.2	24.5	30.1
w/ Online Update	32.6	32.2	34.2	25.8	**31.2**
Gain of Online Update	–	+2.3	+1.0	+1.3	+1.1

Table 3.5: Open domain semantic segmentation performance comparison w/ or w/o online update: SYNTHIA-SF→BDD100K with DeepLab-VGG16 backbone. The results are reported on mIoU over 11 classes.

formance on different benchmarks, proving the effectiveness of our proposed MOCDA model.

4

TAXONOMY ADAPTIVE CROSS-DOMAIN SEMANTIC SEGMENTATION

This chapter corresponds to our published article:

Rui Gong, Martin Danelljan, Dengxin Dai, Danda Pani Paudel, Ajad Chhatkuli, Fisher Yu, and Luc Van Gool. „TACS: Taxonomy adaptive cross-domain semantic segmentation.“ In: *ECCV*. 2022

In this chapter, we introduce a taxonomy adaptive cross-domain semantic segmentation (TACS) problem, allowing for inconsistent taxonomies between the two domains. The motivation behind this introduction is to address the limitation of traditional domain adaptation in the output space. While tackling the input domain gap, the traditional domain adaptation settings assume no domain change in the output space. In semantic prediction tasks, different datasets are often labeled according to different semantic taxonomies. In many real-world settings, the target domain task requires a different taxonomy than the one imposed by the source domain. To tackle TACS, we further propose an approach that jointly addresses the image-level and label-level domain adaptation. On the label-level, we employ a bilateral mixed sampling strategy to augment the target domain, and a relabelling method to unify and align the label spaces. We address the image-level domain gap by proposing an uncertainty-rectified contrastive learning method, leading to more domain-invariant and class-discriminative features. We extensively evaluate the effectiveness of our framework under different TACS settings: open taxonomy, coarse-to-fine taxonomy, and implicitly-overlapping taxonomy. Our approach outperforms the previous state-of-the-art by a large margin, while being capable of adapting to target taxonomies.

4.1 INTRODUCTION

Traditional unsupervised domain adaptation (UDA) approaches for semantic segmentation [184, 72, 30, 175, 111, 174] typically focus on the *image level* domain gap, which can involve visual style, weather, lighting conditions, *etc*.. However, these methods are restricted by the

assumption of having consistent taxonomies between source and target domains, *i.e.*, each source domain class can be unambiguously mapped to one target domain class (Fig. 4.1 (a-c)), which is often not the case. In many applications, the label spaces of the source and target domains are inconsistent, due to different scenarios or requirements, inconsistent annotation practices, or the strive towards an increasingly fine-grained taxonomy [128, 92, 34].

The aforementioned considerations motivate us to consider the *label level* domain gap problem. Even though recent open/universal/class-incremental domain adaptation works [134, 202, 90] touched upon the label level domain gap, they 1) only took image classification as test-bed, and 2) only focused on unseen classes in the target domain. However, the label level domain gap in practical scenarios is more complicated than only involving unseen classes. We therefore formulate and explore the label level domain gap problem in a more general and complete setting. We identify three typical types of label taxonomy inconsistency. i) *Open taxonomy*: some classes, *e.g.*, "terrain" in Fig. 4.1(d), appear in the target domain, but are unlabeled or unseen in the source domain. ii) *Coarse-to-fine taxonomy*: some classes in the source domain, *e.g.*, "person", are split into several sub-classes in the target domain, *e.g.*, "pedestrian" and "rider' (Fig. 4.1(e)). iii) *Implicitly-overlapping taxonomy*: for a certain class in the source domain, one or more of its sub-classes are merged into other classes in the target domain. For example, there exists a taxonomic conflict between {"vehicle", "bicycle"} in the source domain and {"car", "cycle"} in the target domain (Fig. 4.1(f)).

We therefore introduce a more general and challenging domain adaptation problem, namely *taxonomy adaptive cross-domain semantic segmentation* (TACS). In traditional UDA for semantic segmentation, the goal is to transfer a model learned on a labelled source domain to an unlabelled target domain, under the consistent taxonomy assumption. In contrast, TACS allows for inconsistent taxonomies between a labeled source domain and a few-shot/partially labeled target domain, where the inconsistent classes of the target domain are exemplified by a few labeled samples. Thus TACS approaches domain adaptation on both the image and label side, under the few-shot/partially labeled setting. Such task setting is realistic, but involves practical challenges. On the one hand, TACS allows methods to make full use of the labeled source domain without annotation costs in the target domain for the consistent classes. On the other hand, for the inconsistent classes the taxonomy

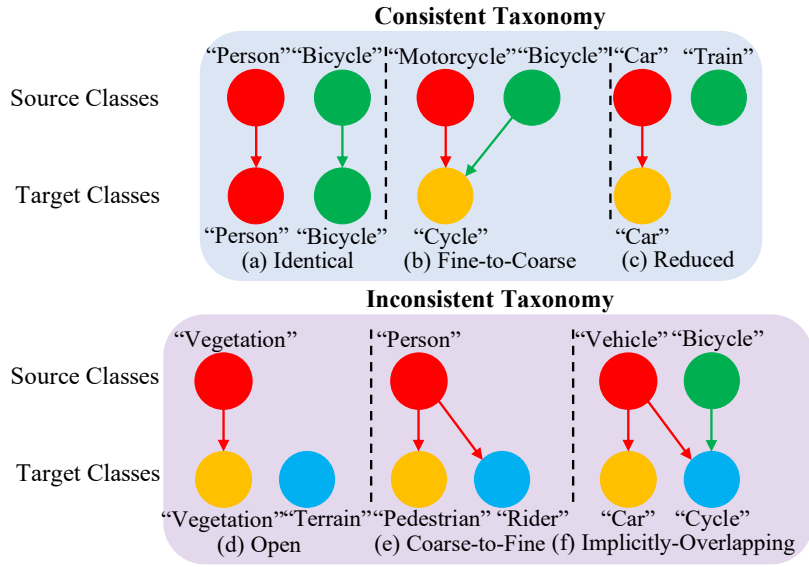


Figure 4.1: Consistent *vs.* inconsistent taxonomy. In (a)-(f), the upper row shows the source domain classes, and the lower row the target domain classes. Circles represent classes while an arrow represents a mapping from a source domain class to a target domain class. (a)-(c) and (d)-(f) are examples of consistent and inconsistent taxonomies, resp. Different from other domain adaptation problems, *e.g.*, universal/partial/open-set domain adaptation [202, 16, 134], that only touch the consistent taxonomy or special case of open taxonomy, our TACS provides a more general problem, including the consistent taxonomy and different inconsistent taxonomies types. More detailed comparisons with other domain adaptation problems are put in Sec. 4.2.

adaptation should only require very limited supervision in the target domain, *i.e.*, only few samples should be labeled there.

We put forward the first approach for TACS, addressing both the image and label domain gaps. As to the latter, we aim to remedy the gap using pseudo-labelling techniques. First, a *bilateral mixed sampling* strategy is proposed to augment unlabeled images by mixing them with both labeled source-domain and target-domain samples. Second, we map inconsistent source domain labels with a *stochastic label mapping* strategy, which encourages a more flexible taxonomy adaptation during the earlier learning phase. Third, a *pseudo-label based relabeling* strategy

is proposed to replace the inconsistent classes in the source-domain according to the model's predictions, to further enforce taxonomy adaptation during the training process. To tackle the image level domain gap, we introduce an *uncertainty-rectified contrastive learning* scheme that facilitates the learning of class-discriminative and domain-invariant features, under the uncertainty-aware guidance of predicted pseudo-labels. Our complete approach for TACS demonstrates strong results in different inconsistent taxonomy settings (*i.e.*, open, coarse-to-fine, and implicitly-overlapping). Moreover, our suggested mixed-sampling and contrastive-learning scheme outperforms current state-of-the-art methods by a large margin in the traditional UDA setting.

To summarize, our contributions are three-fold:

- A new problem – *taxonomy adaptive cross-domain semantic segmentation* (TACS) – of addressing both image and label domain gaps is proposed. It opens up a new avenue for more flexible cross-domain semantic segmentation.
- A generic solution for UDA and TACS is proposed, for which the unified mixed-sampling, pseudo-labeling and uncertainty-rectified contrastive learning scheme is presented to solve both image and label level domain gaps.
- Extensive experiments are conducted under the traditional UDA and the new TACS settings, showing the effectiveness of our approach.

4.2 RELATED WORK

Domain adaptation: The traditional unsupervised domain adaptation (UDA) [175, 214, 73, 47, 227, 114] considers the case when the source and target domain share the same label space and where the target domain is unlabeled. However, this setting does not conform with many practical applications. Some recent works have therefore explored alternative settings. **Open-set/universal domain adaptation** [134, 158, 202] aims at recognizing the new unseen classes in the target domain together as the "unknown" class. **Class-incremental/zero-shot domain adaptation** [90, 11] are proposed to recognize the new unseen classes explicitly and separately in the target domain under the source domain free setting and in the zero-shot segmentation way, resp. These works

touch upon the specific case of the open taxonomy setting in TACS. However, the above works only consider the case where the unseen classes are absent in the source domain. In contrast, the open taxonomy setting in TACS also allows for the unseen classes to exist in the source domain, where they are unlabelled. Besides, the above works do not consider the coarse-to-fine and implicitly-overlapping taxonomy problems, which are covered by the more general TACS formulation. Recent **few-shot/semi-supervised domain adaptation** works [172, 125, 211] aim at improving the domain adaptation performance by introducing few-shot fully labeled target domain samples. However, they still assume a consistent taxonomy between the source and target domain. Moreover, all the aforementioned non-UDA works, except for [11] and [211], only touch upon the image classification task. Instead, our TACS aims at semantic segmentation, which is more challenging and raises particular interest due to its great importance in autonomous driving [175, 184, 174, 123]. Next, we compare our TACS with different domain adaptation problems in more detail, respectively.

TACS vs., Unsupervised Domain Adaptation (UDA). The traditional UDA [175, 184] only focuses on the image-level domain gap, but ignores the label-level domain gap (cf. Fig. 4.1), *i.e.*, assuming the consistent taxonomy between the source domain and the target domain.

TACS vs., Partial Domain Adaptation (PDA). The implicitly-overlapping taxonomy in our TACS is totally different from PDA [16]. PDA only assumes the reduced label space from the source domain to the target domain, *e.g.*, {"vehicle", "bicycle"} $\rightarrow$ {"bicycle"}, which actually still assumes consistent taxonomy between the source domain and the target domain (cf. (c) in Fig. 4.1). However, the implicitly-overlapping taxonomy setting in our TACS touches the problem that, for a certain class in the source domain, one or more of its sub-classes are merged into other classes in the target domain, *e.g.*, {"vehicle", "bicycle"} $\rightarrow$ {"car", "cycle"}, which tackles the inconsistent taxonomy between the source domain and the target domain (cf. (f) in Fig. 4.1).

TACS vs., Few-Shot/Semi-Supervised Domain Adaptation (FS/SS DA). FS/SS DA [211, 125, 172] aims at improving the domain adaptation performance by introducing the few-shot fully labeled target domain samples. However, FS/SS DA still assumes the consistent taxonomy between the source domain and the target domain.

TACS vs., Open-Set/Universal Domain Adaptation (OS/US DA). OS/US DA [202, 134, 158] aims at recognizing the new unseen classes in the

target domain together as an "unknown" class, which can be seen as a special case of our open taxonomy setting in our TACS. Differently, the open taxonomy setting in our TACS aims at recognizing different new unseen classes explicitly and separately. For example, assuming {"terrain", "train"} are the new unseen classes in the target domain, OS/US DA just aims at recognizing the pixels of {"terrain", "train"} classes as the "unknown" class pixel together. However, the open taxonomy setting in our TACS aims at recognizing the pixels of {"terrain", "train"} classes as the "terrain" and "train" classes explicitly and seperately, as the recognition of the seen class. Besides, OS/US DA does not consider the coarse-to-fine taxonomy and implicitly-overlapping taxonomy setting in our TACS.

TACS vs., Zero-Shot/Class-Incremental Domain Adaptation (ZS/CI DA). Similar to the open taxonomy setting of our TACS, ZS/CI DA [11, 90] aims at recognizing the new unseen classes in the target domain explicitly and separately, which can be seen as a specific case of the open taxonomy setting of our TACS. However, ZS/CI DA only considers the case where the unseen classes are absent in the source domain. In contrast, the open taxonomy setting in our TACS also allows for the unseen classes to exist in the source domain, where they are unlabelled. Besides, ZS/CI DA does not consider the coarse-to-fine taxonomy and implicitly-overlapping taxonomy setting in our TACS.

Contrastive learning: Recently, contrastive learning [25, 63, 26, 67, 27, 180] was proven to be successful for unsupervised image classification. Benefiting from the strong representation learning ability, contrastive learning has been applied to different applications, including semantic segmentation [188], image translation [137], object detection [195] and domain adaptation [87]. In [87], contrastive learning is exploited to minimize the intra-class discrepancy and maximize the inter-class discrepancy for the domain adaptive image classification task. However, since the approach is designed for the image classification task, it utilizes the contrastive learning between the whole feature vectors of the different image samples, which is not directly applicable to dense prediction tasks, such as semantic segmentation. Instead, we develop a pseudo-label guided and uncertainty-rectified pixel-wise contrastive learning, to distinguish between positive and negative pixel samples to learn more robust and effective cross-domain representations.

4.3 METHOD

4.3.1 *Problem Statement*

In our taxonomy adaptive cross-domain semantic segmentation (TACS) problem, we are given the labeled source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, where $\mathbf{x}^s \in \mathbb{R}^{H\times W\times 3}$ is the RGB color image, and $\mathbf{y}^s$ is the associated ground truth C_S-class semantic label map, $\mathbf{y}^s \in \{1, ..., C_S\}^{H\times W}$. In the target domain, we are also given a limited number of labeled samples $\mathcal{D}_t = \{(\mathbf{x}_i^t, \mathbf{y}_i^t)\}_{i=1}^{n^t}$, which we refer to as few-shot or partially labeled target domain samples. We are also given a large set of unlabeled target domain samples $\mathcal{D}_u = \{\mathbf{x}_i^u\}_{i=1}^{n^u}$. The target ground truth $\mathbf{y}^t$ follows the C_T-class semantic label map. Denoting the source and target image samples distributions as P_S and P_T, we have $\mathbf{x}^s \sim P_S$, $\mathbf{x}^t, \mathbf{x}^u \sim P_T$. The source and target image distributions are different, *i.e.*, $P_S \neq P_T$. The label set space of $\mathcal{D}_s$ and $\{\mathcal{D}_t, \mathcal{D}_u\}$ are given by $\mathcal{C}_s = \{\mathbf{c}_1^s, \mathbf{c}_2^s, ..., \mathbf{c}_{C_S}^s\}$ and $\mathcal{C}_t = \{\mathbf{c}_1^t, \mathbf{c}_2^t, ..., \mathbf{c}_{C_T}^t\}$ resp., and $\mathcal{C}_s \neq \mathcal{C}_t$. The inconsistent taxonomy subsets of $\mathcal{C}_s, \mathcal{C}_t$ are denoted as $\overline{\mathcal{C}_s}, \overline{\mathcal{C}_t}$, resp. Our goal is to train the model on $\mathcal{D}_s$, $\mathcal{D}_t$ and $\mathcal{D}_u$, and evaluate on the target domain data in the label sets space $\mathcal{C}_t$.

Inconsistent Taxonomy. [1] Specifically, we consider three different cases of inconsistent taxonomy. 1) The *open taxonomy* considers the case where new classes, unseen or unlabeled in the source domain, appear in the target domain. That is, $\exists \mathbf{c}_j^t \in \mathcal{C}_t$ such that $\mathbf{c}_i^s \cap \mathbf{c}_j^t = \emptyset, \forall \mathbf{c}_i^s \in \mathcal{C}_s$. 2) The *coarse-to-fine taxonomy* considers the case where the target domain has a *finer* taxonomy where source classes have been split into two or more target classes. That is, $\exists \mathbf{c}_i^s \in \mathcal{C}_s, \mathbf{c}_{j_1}^t \in \mathcal{C}_t, \mathbf{c}_{j_2}^t \in \mathcal{C}_t, j_1 \neq j_2$ such that $\mathbf{c}_{j_1}^t, \mathbf{c}_{j_2}^t \neq \mathbf{c}_i^s$ and $(\mathbf{c}_{j_1}^t \cup \mathbf{c}_{j_2}^t) \subseteq \mathbf{c}_i^s$. 3) The *implicitly-overlapping taxonomy* considers the case where a class in the target domain has a common part with the class in the source domain, but also owns the private part. That is, $\exists \mathbf{c}_i^s \in \mathcal{C}_s, \mathbf{c}_j^t \in \mathcal{C}_t$ such that $\mathbf{c}_j^t \not\subseteq \mathbf{c}_i^s, \mathbf{c}_i^s \cap \mathbf{c}_j^t \neq \emptyset$, and $(\mathbf{c}_j^t \setminus (\mathbf{c}_i^s \cap \mathbf{c}_j^t)) \notin \{\emptyset, \mathbf{c}_q^s, q = 1, ..., C_S\}$.

Few-shot/Partially Labeled. In TACS, the $\mathcal{D}_t$ is only few-shot/partially labeled for the inconsistent taxonomy classes, in the class-wise way. More specifically, for each of the class $\mathbf{c}_j^t \in \overline{\mathcal{C}_t}$, we have n^t-shot labeled

1 With a slight abuse of notation, each class, *e.g.*, $\mathbf{c}_i^s$, is also considered as a set consisting of its domain of definition. The set operations $\cap, \cup, \setminus, \subset$ thus applies to the underlying definition of the class.

samples $\{(\mathbf{x}_i^{t_j}, \mathbf{y}_i^{t_j})\}_{i=1}^{n^t}$, where only the class $\mathbf{c}_j^t$ is labeled in $\mathbf{y}_i^{t_j}$. When $n^t \ll n^u$, it is called few-shot labeled. When $n^t \not\ll n^u$, it is named partially-labeled. The sample and corresponding semantic map is written as $\mathbf{x}^{t_j}$ and $\mathbf{y}^{t_j}$.

Technical Challenges. The main technical challenge of TACS is to deal with both of the label-level and image-level domain gap. On the **label level**, there are two main problems: i) The inconsistent taxonomy may induce there is the *one-to-many* mapping from the source domain to the target domain classes. If we purely assign the source class of inconsistent taxonomy to one of the corresponding target class, it will generate incorrect supervision, degrading the performance of the model. However, if we instead take the inconsistent source class as unlabeled, the source domain information is not fully exploited. ii) The complete target domain label taxonomy is partially inherited from the source domain for the consistent taxonomy, and partially provided by the few-shot/partially labeled target domain. The problem of how to *unify the consistent and inconsistent taxonomy classes* for the target domain is non-trivial. The naive way is to train the model on the source domain for the consistent taxonomy classes, and on the few-shot/partially labeled target domain for the inconsistent taxonomy classes separately, in the supervised way. However, the few-shot labeled target domain samples are far fewer than the labeled source domain samples, causing the model training to be easily dominated by the consistent taxonomy classes, therefore the inconsistent taxonomy classes are possibly ignored. Meanwhile, most of the pixels in the few-shot/partially labeled target domain samples are unlabeled except for the pixels of class $\mathbf{c}_j^t$, and the arbitrarily incorrect prediction on these unlabeled parts can bring the negative effect since most of these parts belong to the consistent taxonomy classes or other inconsistent taxonomy classes. On the **image level**, the image domain distribution difference between the source and target domain, $P_S \neq P_T$, still exists in TACS.

4.3.2 *Our Approach to the TACS Problem*

Motivation. Motivated by the technical challenge i) of the label level in Sec. 4.3.1, the stochastic label mapping (SLM) and pseudo-label based relabeling (RL) module are proposed to solve the problem of the one-to-many mappings from the source domain to the target domain

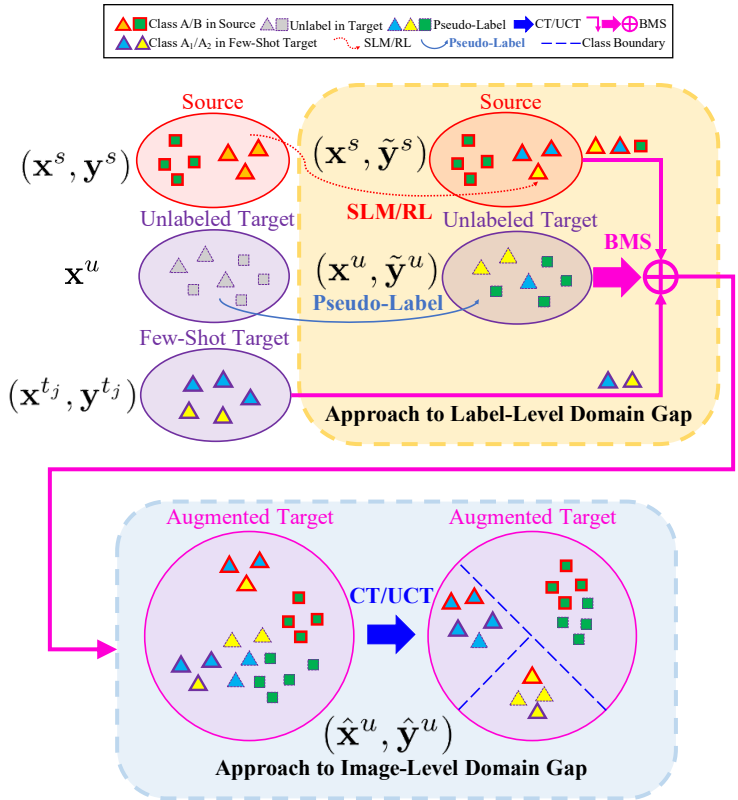


Figure 4.2: Framework overview. Class A is an inconsistent taxonomy class (*e.g.*, "person") in the source domain, related to class A_1 (*e.g.*, "pedestrian") and A_2 (*e.g.*, "rider") in the target domain. Class B is a consistent taxonomy class. On the label level, SLM/RL module maps the inconsistent taxonomy class A in the source domain to the related classes A_1, A_2 in the target domain. BMS module unifies label space and augments the few-shot supervision, by randomly selecting samples from the source domain and the few-shot/partially labeled target domain and then mixing them in the unlabeled target domain. On the image level, CT/UCT module adopts the pseudo-label to distinguish the positive and negative pixel samples, and then conducts the pixel-wise contrastive learning, to learn more domain-invariant and class-discriminative features.

classes. Motivated by the technical challenge ii) of the label level in Sec. 4.3.1, the bilateral mixed sampling (BMS) module is proposed to

unify the consistent and inconsistent taxonomy classes and augment the few-shot supervision for the target domain. Motivated by the technical challenge of the image level in Sec. 4.3.1, the contrastive learning (CT/UCT) module is proposed to train the domain-invariant but class-discriminative features.

Training Strategy. The whole framework adopts the pseudo-label based self-training strategy. Following the self-training structure of [129], there are two components of our framework, namely a student network $\mathcal{F}_\theta$ and a mean-teacher network $\mathcal{F}_{\theta'}$, which are both semantic segmentation networks. The student network $\mathcal{F}_\theta$ is used to backpropagate the gradients and update θ according to the training loss. The pseudo-labels $\tilde{\mathbf{y}}^u = \mathcal{F}_{\theta'}(\mathbf{x}^u)$ are generated by the mean-teacher network $\mathcal{F}_{\theta'}$ by feeding the unlabeled target sample $\mathbf{x}^u$. The parameters θ' are the exponential moving average of the parameters θ during the optimization process, which is proven to bring more stable training [174, 171]. During inference, the mean-teacher network $\mathcal{F}_{\theta'}$ is used to output the final segmentation map.

Framework Overview. The framework overview is shown in Fig. 4.2. The SLM and RL modules (Sec. 4.3.3) are used to map inconsistent taxonomy class labels $\mathbf{y}^s$ in the source domain to target-domain class labels $\tilde{\mathbf{y}}^s$. Then in order to unify the label spaces, the source domain sample $(\mathbf{x}^s, \tilde{\mathbf{y}}^s)$ and the few-shot/partially labeled target domain sample $(\mathbf{x}^{t_j}, \mathbf{y}^{t_j})$ is cut and mixed with the unlabeled target domain sample and corresponding pseudo-label $(\mathbf{x}^u, \tilde{\mathbf{y}}^u)$, to synthesize the sample $(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$ through the BMS module (Sec. 4.3.3). In this way, the synthesized sample $(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$ is a cross-domain mixed sample and covers the consistent taxonomy class from $(\mathbf{x}^s, \tilde{\mathbf{y}}^s)$ and inconsistent taxonomy class from $(\mathbf{x}^{t_j}, \mathbf{y}^{t_j})$. The CT/UCT module (Sec. 4.3.4) is further utilized on the $(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$ to train the domain-invariant and class-discriminative features using pixel-wise contrastive learning. All the modules are thus employed together in a single framework. Next, we detail individual components.

4.3.3 *Approach to the Label Level Domain Gap*

In order to solve the problem of *one-to-many class mappings*, the SLM and RL modules are proposed. In the initial training stage, the model is unable to distinguish the different inconsistent taxonomy classes reliably. Thus, taking the coarse-to-fine taxonomy as example, we pro-

pose the SLM module, and it stochastically assigns the source "coarse class" to different corresponding target "finer classes" to guide the model to predict the uniform distribution over the "finer classes" on the source domain samples. In this way, in the early training stage, the prediction of the model on the "finer classes" will be mainly guided by the few-shot labeled target samples. As the training goes on, with the help of the few-shot labeled target samples, the teacher network gradually has the capacity to distinguish the "finer classes". In the second stage, we then replace the SLM module with the RL module. It relabels the "coarse-class" pixel in the source domain with the "finer class" predicted by the teacher network.

Stochastic Label Mapping (SLM). We propose the SLM module, which maps the source domain classes of inconsistent taxonomy, *e.g.*, "person" in Fig. 1 (e), to the corresponding target domain classes stochastically, *e.g.*, "pedestrian" and "rider" in Fig. 1 (e), in the initial training stage and *in each training iteration*. Under the inconsistent taxonomy setting, there might be the one-to-many class mapping from the source domain classes to the target domain label space. Without loss of generality and for the convenience of clarification, we take the example that the corresponding classes in $\mathcal{C}_t$ of $\mathbf{c}_i^s$ include q classes $\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t$. Then the SLM module can be described as, $\tilde{\mathbf{y}}^{s(m,n)} = \text{rand}(\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t)$, where the (m, n) is the (row, column) index. The $\text{rand}(\cdot)$ represents the uniformly discrete sampling function. With the obtained new labels $\tilde{\mathbf{y}}^s$, we employ the standard cross-entropy loss, $\mathcal{L}_{slm} = CE(\mathcal{F}_\theta(\mathbf{x}_s), \tilde{\mathbf{y}}^s)$ to learn the model.

Pseudo-Label based Relabeling (RL). As the training goes on, the model learns to distinguish the different inconsistent taxonomy classes to some extent. Instead of adopting SLM strategy at the latter part of the training, we introduce an alternative strategy. To exploit the capabilities learned by the model, we perform the pseudo-label based relabeling (RL), which relabels the pixels of inconsistent taxonomy classes in the source domain with the classes predicted by the model. Without loss of generality and for the writing convenience, we take the same example that $\mathbf{c}_i^s$ is related to $\mathbf{c}_p^t, \mathbf{c}_{p+1}^t, ..., \mathbf{c}_{p+q-1}^t$ as in SLM module. We generate predictions $\mathbf{f}^s = \mathcal{F}_{\theta'}(\mathbf{x}^s)$ by feeding the source domain sample $\mathbf{x}^s$ into the mean-teacher network $\mathcal{F}_{\theta'}$. Then the prediction $\mathbf{f}^s$ is used to relabel the source domain sample $\mathbf{x}^s$ for the inconsistent taxonomy classes $\mathbf{c}_i^s$, to generate the complete label $\tilde{\mathbf{y}}^s$ as, $\tilde{\mathbf{y}}^{s(m_i^s,n_i^s)} = \text{argmax}_c\, \mathbf{f}^{s(m_i^s,n_i^s)}$, if $\max_c \mathbf{f}^{s(m_i^s,n_i^s)} > \delta$, and $\text{argmax}_c\, \mathbf{f}^{s(m_i^s,n_i^s)} \in$

$\{\mathbf{c}_p^t, ..., \mathbf{c}_{p+q-1}^t\}$. (m_i^s, n_i^s) is the index of the pixel corresponding to $\mathbf{c}_i^s$. The δ represents the threshold to decide whether the predicted label is used. The pseudo-label based relabeling module loss is written as $\mathcal{L}_{rl} = CE(\tilde{\mathbf{y}}^s, \mathcal{F}_\theta(\mathbf{x}^s))$. The SLM module and the RL module are used in the sequential manner during the training process, *i.e.*, initially SLM and then RL.

Bilateral Mixed Sampling (BMS). In order to *unify the consistent and inconsistent taxonomy classes* and *augment the few-shot supervision* for the target domain, we propose the bilateral mixed sampling (BMS) module, which cuts and mixes the source domain and few-shot/partially labeled target domain samples on the unlabeled target domain. Recently, the mixed sampling based data augmentation approach [208, 51, 205] is proven to be able to generate the synthetic data to combine the samples and corresponding labels, thus provides such a potential to unify the label space. In [174], the cross-domain mixed sampling (DACS) is shown helpful to UDA of consistent taxonomy.

Similar to DACS for UDA, we adopt the class-mixed sampling strategy for TACS. Different from DACS, which only focus on the labeled source domain and the unlabeled target domain, our BMS module conducts the class-mixed sampling in the bilateral way: 1) from labeled source domain samples $\mathbf{x}^s$ to unlabeled target domain samples $\mathbf{x}^u$; 2) from few-shot/partially labeled target domain samples $\mathbf{x}^{t_j}$ to unlabeled target domain samples $\mathbf{x}^u$. The bilateral mixed sampling mask $\mathbf{m}^s$ of $\mathbf{x}^s$ is,

$$\mathbf{m}^{s(m,n)} = \begin{cases} 1, \text{if } \tilde{\mathbf{y}}^{s(m,n)} = \mathbf{c}_r \\ 0, \text{otherwise,} \end{cases} \tag{4.1}$$

where the sampling class $\mathbf{c}_r$ is randomly selected from the available classes in $\tilde{\mathbf{y}}^s$. Following [174], half of all the available classes in $\tilde{\mathbf{y}}^s$ is randomly selected as the sampling class in each training iteration. Similar to $\mathbf{m}^s$, the bilateral mixed sampling mask $\mathbf{m}^{t_j}$ of $\mathbf{x}^{t_j}$ is defined. Then the augmented target domain sample and the corresponding pseudo-label $\hat{\mathbf{x}}^u$, $\hat{\mathbf{y}}^u$ are,

$$\hat{\mathbf{x}}^u = \mathbf{m}^s \odot \mathbf{x}^s + (1 - \mathbf{m}^s) \odot (\mathbf{m}^{t_j} \odot \mathbf{x}^{t_j} + (1 - \mathbf{m}^{t_j}) \odot \mathbf{x}^u), \tag{4.2}$$

$$\hat{\mathbf{y}}^u = \mathbf{m}^s \odot \tilde{\mathbf{y}}^s + (1 - \mathbf{m}^s) \odot (\mathbf{m}^{t_j} \odot \mathbf{y}^{t_j} + (1 - \mathbf{m}^{t_j}) \odot \tilde{\mathbf{y}}^u). \tag{4.3}$$

where $\odot$ denotes element-wise multiplication. On this basis, the pseudo-label based self-training loss of our BMS module is formulated as, $\mathcal{L}_{bms} = CE(\hat{\mathbf{x}}^u, \hat{\mathbf{y}}^u)$.

4.3.4 *Approach to the Image Level Domain Gap*

Besides dealing with the label-level domain gap, we also need to tackle the *image-level domain gap*. We propose a pseudo-label based contrastive learning (CT) module, and further the pseudo-label based uncertainty-rectified contrastive learning (UCT) module. They are easy to be plugged into our self-training pipeline and trained jointly with the BMS, SLM and RL modules.

Contrastive Learning (CT) for Domain Adaptation. The typical strategy of image-level adaptation is to train the domain-invariant but class-discriminative features in the cross-domain embedding space [47, 175, 48]. The pixels of the same class from different or same domains need to have similar features in the feature embedding space, while the pixels of different classes needs be distinguishable in the feature embedding space. This kind of distinction between features can naturally be formulated as a contrastive learning problem, where positive pairs stem from pixels of the same class, irrespective of their domain. In [188], the pixel-wise contrastive learning is proven to be helpful for semantic segmentation. However, it relies on ground truth label, which is unavailable for our unlabeled samples.

In order to exploit contrastive learning to train domain-invariant and class-discriminative features under cross-domain setting, we propose the pseudo-label based contrastive learning for domain adaptation. We employ pseudo-labels as guidance for distinguishing the positive and negative samples. The contrastive learning is conducted on the augmented target domain image sample $\hat{\mathbf{x}}^u$, and corresponding pseudo-label $\hat{\mathbf{y}}^u$ in the BMS module. Our main semantic segmentation network $\mathcal{F}_\theta$ can be decomposed into the encoder $\mathcal{E}_\theta$ and the decoder $\mathcal{M}_\theta$. The decoder is used to map the embedding space $\mathcal{V}$ to the label domain $\mathcal{Y}$. The encoder $\mathcal{E}_\theta$ maps the source image domain $\mathcal{S}$ and the target image domain $\mathcal{T}$ to the embedding space $\mathcal{V}$, *i.e.*, $\mathcal{E}_\theta : \mathcal{S}, \mathcal{T} \rightarrow \mathcal{V}$. The feature embedding corresponding to the sample $\hat{\mathbf{x}}^u$ is denoted as $\hat{\mathbf{v}}^u$, *i.e.*, $\hat{\mathbf{v}}^u = \mathcal{E}_\theta(\hat{\mathbf{x}}^u)$. Then the pseudo-label based contrastive learning module loss $\mathcal{L}_{ct}$ can be described as,

$$\mathcal{L}_{ct} = -\sum_h \sum_w \log \sum_{v^+ \in \mathcal{P}_v} \text{Contrast}(v, v^+), \tag{4.4}$$

$$\text{Contrast}(v, v^+) = \frac{\exp(v \cdot v^+ / \tau)}{\exp(v \cdot v^+ / \tau) + \sum_{v^- \in \mathcal{N}_v} \exp(v \cdot v^- / \tau)}, \tag{4.5}$$

where $v = \hat{\mathbf{v}}^{u(h,w)}$ is the feature vector of $\hat{\mathbf{v}}^u$ at the position (h, w). The positive samples in $\mathcal{P}_v$ are the feature vectors whose corresponding pixels in $\hat{\mathbf{y}}^u$ have the same class label as that of the corresponding pixel of v. The negative samples in $\mathcal{N}_v$ are the feature vectors whose corresponding pixels in $\hat{\mathbf{y}}^u$ have the different class label from that of the corresponding pixel of v. Eq. (4.5) tries to learn similar features for the pixels of the same class, and learn discriminative features for the different class pixels, no matter whether pixels are in the same domain or not.

Uncertainty-Rectified Contrastive Learning (UCT) for Domain Adaptation. There unavoidably exist incorrect predictions in the pseudo-label $\hat{\mathbf{y}}^u$ of the unlabeled part in CT module, resulting in incorrect guidance to the contrastive module for the selection of the positive and negative samples. In order to alleviate the incorrect guidance, we propose the uncertainty-rectified contrastive learning (UCT) module based on the CT module. In our UCT module, we use the prediction uncertainty of the pseudo-label $\hat{\mathbf{y}}^u$ to rectify the contrastive learning, so that the uncertain prediction of $\hat{\mathbf{y}}^u$ has less effect on the contrastive learning. The uncertainty estimation map of $\hat{\mathbf{y}}^u$ is denoted as $\hat{\mathbf{u}}^u$, and the uncertainty measurement function is denoted as $\mathcal{U}(\cdot)$, *i.e.*, $\hat{\mathbf{u}}^u = \mathcal{U}(\hat{\mathbf{y}}^u)$. We adopt the maximum prediction probability of $\hat{\mathbf{x}}^u$ as $\mathcal{U}(\cdot)$, formulated as,

$$\hat{\mathbf{u}}^u = \max_c \mathcal{F}_{\theta'}(\hat{\mathbf{x}}^u). \tag{4.6}$$

Then, based on Eq. (4.5), the uncertainty-rectified CT loss $\mathcal{L}_{uct}$ is formulated as,

$$\mathcal{L}_{uct} = -\sum_h \sum_w \hat{\mathbf{u}}^u(v)\hat{\mathbf{u}}^u(v^+)\,\text{Contrast}(v, v^+), \tag{4.7}$$

where $\hat{\mathbf{u}}^u(v)$, $\hat{\mathbf{u}}^u(v^+)$ are the uncertainty estimation value of the pixel corresponding to v, v^+, resp.

4.3.5 *Joint Training*

With the above proposed BMS, SLM, RL and UCT modules, the total loss function is derived as,

$$\mathcal{L}_{total} = \mathcal{L}_{bms} + \lambda_1 \mathcal{L}_{slm} + \lambda_2 \mathcal{L}_{rl} + \lambda_3 \mathcal{L}_{uct} \tag{4.8}$$

where λ_1 and λ_2 are used to train the SLM and RL module in a sequential manner. When iteration $t < T$, $\lambda_1 = 1, \lambda_2 = 0$. When iteration

$t \geq T$, $\lambda_1 = 0, \lambda_2 = 1$. T is the number of iterations to start training the RL module. λ_3 is the hyper-parameter to balance the UCT module loss and other loss, which is set as 0.01 in our work. Our model is trained end-to-end with the loss in Eq. (4.8).

4.4 EXPERIMENTS

We evaluate the effectiveness of our framework under different scenarios, including the consistent and inconsistent taxonomy settings. For the consistent taxonomy, we follow the traditional UDA setting. For the inconsistent taxonomy, we build different benchmarks for TACS, including the open, coarse-to-fine and implicitly-overlapping taxonomy setting. The DeepLabv2-ResNet101 [20, 68] is adopted as the segmentation network. The baselines in Table 4.3-4.5 adopt the SOTA few-shot cross-domain semantic segmentation training strategy, *i.e.*, fine-tuning [211] and pseudo-label [129], to exploit the supervision from the few-shot labeled target domain.

4.4.1 *Experimental Setup*

Datasets. *SYNTHIA.* SYNTHIA [152] is a syntheic image dataset, consisting of photo-realistic images rendered from a virtual city. We adopt SYNTHIA-RAND-CITYSCAPES subset, including 9400 densely labeled synthetic images. *GTA5.* GTA5 [149] is a synthetic image dataset, containing 24966 urban scene images. The images in GTA5 dataset are rendered from game engine, and densely labeled with pixel-level semantic annotation. The scene of GTA5 dataset is based on the city of Los Angeles. *Synscapes.* Synscapes [192] is a photo-realistic synthetic dataset, created with physically based rendering techniques. Synscapes is built for street scene parsing, composed of 25000 densely pixel-level annotated images. *Cityscapes.* Cityscapes [34] is a real street scene image dataset, collected from different European cities. We adopt the training set of Cityscapes during the training stage, covering 2975 images. And we use the validation set of Cityscapes, including 500 images, to evaluate the performance of the semantic segmentation model.

UDA: Consistent Taxonomy. We adopt the UDA setting for the consistent taxonomy. The target domain is completely unlabeled. In the SYNTHIA→ Cityscapes experiment, SYNTHIA [152] is used as the

Method	Road	SW	Build	Wall*	Fence*	Pole*	TL	TS	Veg	Sky	Person	Rider	Car	Bus	MC	Bike	mIoU*	mIoU
ADVENT[184]	87.0	44.1	79.7	9.6	0.6	24.3	4.8	7.2	80.1	83.6	56.4	23.7	72.7	32.6	12.8	33.7	47.6	40.8
FDA[200]	79.3	35.0	73.2	–	–	–	19.9	24.0	61.7	82.6	61.4	31.1	83.9	40.8	**38.4**	51.1	52.5	–
IAST[123]	81.9	41.5	**83.3**	17.7	4.6	32.3	30.9	28.8	83.4	85.0	65.5	30.8	**86.5**	38.2	33.1	**52.7**	57.0	49.8
DACS[174]	80.56	25.12	81.90	21.46	**2.85**	**37.20**	22.67	23.99	83.69	**90.77**	**67.61**	**38.33**	82.92	38.90	28.49	47.58	54.81	48.34
Ours (DACS+CT)	86.32	26.63	82.71	5.78	1.97	33.87	**34.60**	**40.00**	83.83	86.73	67.52	36.53	83.46	**55.23**	25.03	41.46	57.70	49.47
Ours (DACS+UCT)	**91.54**	**60.41**	82.52	**21.80**	1.48	31.66	31.59	27.95	**84.71**	88.95	66.68	35.78	81.04	42.79	28.49	45.88	**59.10**	**51.45**

Table 4.1: Consistent Taxonomy: SYNTHIA→Cityscapes. The mIoU are over 13 classes and 16 classes, resp. In UDA setting, we adopt the class-mixed sampling strategy in DACS to augment the target domain. *3 classes are not included when calculating mIoU over 13 classes.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
ADVENT[184]	89.4	33.1	81.0	26.6	26.8	27.2	33.5	24.7	83.9	36.7	78.8	58.7	30.5	84.8	38.5	44.5	1.7	31.6	32.4	45.5
FDA[200]	92.5	53.3	82.4	26.5	27.6	36.4	40.6	38.9	82.3	39.8	78.0	62.6	34.4	84.9	34.1	53.1	16.9	27.7	46.4	50.5
IAST[123]	**93.8**	**57.8**	85.1	**39.5**	26.7	26.2	43.1	34.7	84.9	32.9	88.0	62.6	29.0	87.3	39.2	49.6	**23.2**	34.7	39.6	51.5
DACS[174]†	89.90	39.66	87.87	30.71	**39.52**	38.52	**46.43**	52.79	**87.98**	43.96	**88.76**	**67.20**	**35.78**	84.45	45.73	50.19	0.00	27.25	33.96	52.14
DACS[174]*	93.25	50.20	87.21	36.75	34.80	38.83	39.80	48.68	87.06	44.06	88.76	65.19	34.38	89.25	**51.64**	52.71	0.00	28.59	48.42	53.66
Ours (DACS+UCT)	93.03	55.92	**87.91**	38.19	38.76	**40.44**	42.14	**54.50**	87.53	**46.67**	87.77	66.26	33.67	**90.18**	47.54	**54.15**	0.00	**41.24**	**53.34**	**55.75**

Table 4.2: Consistent Taxonomy: GTA5→Cityscapes. The mIoU is over 19 classes. In the UDA setting, we adopt the class-mixed sampling strategy in DACS to augment the target domain. The best results are denoted in bold. † is the performance reported in the DACS [174]. * is the peak performance model publicly provided by the author of DACS [174].

source domain, while Cityscapes [34] is treated as the target domain. The source domain and target domains share the same label space, where there are 16 classes in total: *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, *rider*, *car*, *bus*, *motorcycle* and *bike*. In the GTA5→Cityscapes experiment, we adopt the GTA5 [149] dataset as the source domain, and the Cityscapes dataset as the target domain. The label space of source domain is composed of *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *terrain*, *vegetation*, *sky*, *moving objects*. The *moving objects* class in the source domain is further divided into 8 classes, including *person*, *rider*, *car*, *truck*, *bus*, *train*, *motorcycle* and *bicycle* in the target domain.

TACS: Open Taxonomy. The SYNTHIA dataset [152] is used as the source domain, and the Cityscapes dataset [34] is adopted as the target domain. In the SYNTHIA dataset, the main 13 classes are labeled: *road*, *sidewalk*, *building*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, *rider*, *car*, *bus*, *motorcycle* and *bike*. In the Cityscapes dataset, the 6 classes *wall*, *fence*, *pole*, *terrain*, *truck* and *train* are few-shot labeled, with 30 image samples per class.

TACS: Coarse-to-Fine Taxonomy. The GTA5 dataset [149] is utilized as the source domain, and the Cityscapes dataset [34] as the target domain. The label space of source domain is composed of *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *sky*, *person*, *car*, *truck*, *bus*, *train*, *cycle*. The *vegetation* class of source domain is further divided into *vegetation* and *terrain* in the target domain, *person* in source domain is mapped to *person* and *rider* in the target domain, and *cycle* in the source domain is fine-grained labeled into *bicycle* and *motorcycle* in the target domain. In Cityscapes, each of the fine-grained 6 classes is 30-shot labeled.

TACS: Implicitly-Overlapping Taxonomy. The Synscapes dataset [192] is treated as the source domain, while the Cityscapes dataset [34] is seen as the target domain. The label space of the source domain contains the *road*, *sidewalk*, *building*, *wall*, *fence*, *pole*, *traffic light*, *traffic sign*, *vegetation*, *terrain*, *sky*, *person*, *rider* and *vehicle*. The *vehicle* class in source domain can be seen as the union of the *car*, *truck*, *bus*, and *motorcycle* classes. In the target domain, each of 3 classes are few-shot labeled in 15 image samples, including the vehicle, public transport and cycle. The *vehicle* class in the target domain is the union of *car* and *truck*, the *public transport* is the union of *bus* and *train*, and *cycle* is the union of the *bicycle* and *motorcycle*.

Batch Size. For the open taxonomy, coarse-to-fine taxonomy and implicitly-overlapping taxonomy experiments of TACS, in each training batch, there are 2 source domain images, 2 unlabeled target domain images and 2 few-shot labeled target domain images mixed in the bilateral mixed sampling module. For the consistent taxonomy experiments of UDA, we strictly follow the traditional UDA setting, and the target domain is completely unlabelled. Therefore, under UDA setting, in each training batch, there are 2 source domain images and 2 unlabelled target domain images mixed in the class mixed sampling way [174].

Parameters. The source domain images are resized to 1280×720, and the target domain images are resized to 1024×512. And the random crop with size 512×512 is then adopted. We adopt the SGD optimizer to train the semantic segmentation network, whose momentum is set as 0.9 and the weight decay is set to 5×10^{-4}. The learning rate is set as 2.5×10^{-4}, with polynomial decay of power 0.9. The iteration T in Sec. 4.3.5 for starting training the RL module is set as 130000. The total training iteration is set as 250000.

Contrastive Learning. We adopt the 2048-dim output vector of the final layer of feature extractor, *i.e.*, , the layer before the classifier, of the Deeplab-v2 framework. The 2048-dim vector is mapped to a 256-dim vector with a projection head, composed of 1x1 Conv, Batchnorm, ReLU, 1x1 Conv layers. The 256-dim vector is then adopted as the pixel-wise feature. For each mini-batch, we use 100 anchor pixel samples per category. The 100 pixel samples of the same category are taken as positive samples, while the other pixel samples of different categories are all taken as negative samples.

Baseline Setup. In the baseline methods setup of Table 4.3, Table 4.4 and Table 4.5, we add the additional supervised loss to train the model in the supervised way, with the few-shot/partially labeled samples in the target domain. For the baseline methods which adopt the pseudo-label based training strategy, such as FDA [200], IAST [123], and DACS [174], the few-shot/partial label on the target domain samples is combined with the generated pseudo-label to attain the final pseudo-label. I.e., in the pseudo-label generation process on the few-shot/partially labeled samples, we adopt the ground-truth label for the labeled parts, while we adopt the generated pseudo-label for other unlabeled parts.

Compute Resources. The code is implemented with PyTorch [138]. Experiments are conducted on an NVIDIA GeForce RTX 2080 Ti GPU, with 11GB memory, where it takes 3 days for training the whole 250000

iterations. In the whole investigation process of our paper, the total compute used is around 390×3 GPU days.

4.4.2 *Experimental Results*

Comparison with the SOTA. In Table 4.1, it is shown that our proposed contrastive-learning based scheme outperforms the previous SOTA methods under the UDA setting, including the adversarial learning based ADVENT [184], the image translation based FDA [200], the self-training based IAST [123], and the data augmentation based DACS [174]. It proves the effectiveness of our contrastive learning for dealing with the domain gap on the image level. In Table 4.3, Table 4.4, and Table 4.5, it is shown that our proposed framework improves other SOTA methods performance by a large margin, under the open, coarse-to-fine and implicitly-overlapping taxonomy settings. It validates the proposed framework for dealing with both of the image- and label-level domain gap. In Fig. 4.6, we show qualitative semantic segmentation results on the target domain.

Ablation Study. The ablation study in Table 4.3, Table 4.4, and Table 4.5 proves that each module, BMS, SLM, RL, CT/UCT, all contributes to the final performance under open, coarse-to-fine, and implicitly-overlapping taxonomy settings. In different settings, the improvement brought by different modules are different. It is mainly because different settings in TACS touch diverse and broad aspects of inconsistent taxonomy. For example, the open taxonomy setting includes the new classes which are unseen or unlabeled in the source domain. The RL module is especially helpful to those unlabeled classes, *e.g.*, "wall" class. The SLM module is significantly beneficial under the coarse-to-fine taxonomy setting since each fine class is corresponding to one coarse class unambiguously. The CT/UCT module contribution difference is mainly related to the image-level difference, *e.g.*, the style difference of SYNTHIA, GTA, Synscapes. Besides, it is shown that the UCT module is able to reach higher performance than the CT module, verifying the help of our uncertainty rectification for contrastive learning. It is also observed that the combination of SLM and other baseline methods, *e.g.*, ADVENT, FDA, IAST and DACS, does not necessarily bring the performance improvement. It is because the model prediction, when using SLM, is guided by the few-shot labeled target samples, but the baseline methods cannot effectively extract and exploit few-shot supervision

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU	mIoU
Source	29.22	6.58	55.48	4.79	8.71	10.11	4.04	12.93	64.06	5.09	71.90	43.26	11.93	22.43	6.04	6.96	2.42	2.61	16.41	6.19	20.26
ADVENT[184]	75.72	24.62	74.94	0.00	0.17	18.98	11.30	16.01	76.87	**21.93**	78.91	48.24	14.20	54.97	2.54	18.38	17.58	12.22	20.90	10.20	30.97
FDA[200]	28.87	13.22	67.10	4.63	14.52	18.94	10.99	14.75	51.56	12.48	78.85	56.78	25.81	70.10	14.24	20.85	21.27	19.22	41.14	14.35	30.81
IAST[123]	70.73	29.60	75.49	6.90	0.00	1.36	36.43	25.37	66.17	7.65	83.96	60.72	19.99	82.51	0.00	39.52	0.09	27.42	23.55	2.67	34.60
DACS[174]	66.48	1.42	6.55	10.26	9.47	4.39	0.47	2.09	33.38	3.75	36.45	46.75	18.23	20.90	1.91	2.78	7.18	1.30	5.08	6.16	14.68
Ours (M)	87.59	27.18	80.98	5.99	15.74	7.13	37.09	18.51	83.68	0.08	87.46	65.89	**37.45**	86.55	24.76	40.58	37.71	**37.57**	43.44	15.24	43.44
Ours (M+CT)	86.33	32.57	82.62	9.49	12.78	5.10	**37.49**	**39.32**	82.00	0.73	**88.03**	65.70	33.09	78.92	33.55	**62.53**	**41.90**	29.83	49.35	17.26	45.86
Ours (M+UCT)	90.84	57.64	80.77	5.79	**16.67**	8.40	32.82	33.21	83.68	1.68	86.89	63.54	26.57	86.87	33.43	48.65	35.57	31.51	49.29	16.92	45.99
Ours (M+UCT+RL)	**92.64**	**58.66**	**84.21**	**20.55**	15.04	**29.47**	35.26	32.41	**84.63**	4.45	87.91	**66.16**	34.07	**87.52**	**36.37**	57.63	31.21	34.17	**52.28**	**22.85**	**49.72**
n^t=2975	89.19	41.08	86.14	37.54	33.68	33.45	32.25	39.99	85.39	31.64	89.51	67.02	35.61	80.49	50.54	49.43	51.70	32.41	47.90	39.76	53.42
Oracle [191]	96.7	75.7	88.3	46.0	41.7	42.6	47.9	62.7	88.8	53.5	90.6	69.1	49.7	91.6	71.0	73.6	45.3	52.0	65.5	50.0	65.9

Table 4.3: Open Taxonomy: SYNTHIA→Cityscapes. There are 13 classes labeled in the SYNTHIA dataset, and 6 new classes few-shot labeled in Cityscapes. The gray columns are the 6 new classes and mean IoU of 6 new classes in Cityscapes. "M" represents BMS module.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU	mIoU
Source	54.12	16.20	70.08	13.07	19.37	22.56	28.59	20.59	75.87	13.49	74.36	47.91	5.35	36.15	16.08	9.71	1.61	8.77	21.34	28.79	29.22
Source*	63.38	20.95	67.65	15.07	18.60	23.03	27.74	18.00	76.03	14.11	75.19	38.36	10.25	49.01	26.32	9.23	2.68	9.93	27.26	29.32	31.20
ADVENT[184]	88.91	38.93	79.18	26.22	22.65	25.45	31.24	25.42	75.22	0.03	78.91	55.76	0.00	77.76	28.22	33.19	0.55	13.02	7.15	25.20	37.25
ADVENT*	86.72	34.02	79.22	22.32	23.60	26.92	31.36	24.89	59.86	3.39	75.47	41.83	7.73	69.62	32.71	20.39	0.49	12.06	39.25	27.35	36.41
FDA[200]	90.83	45.07	81.62	28.37	31.04	32.56	34.00	29.80	83.09	6.31	72.61	60.67	10.13	82.71	29.06	51.51	0.11	15.69	45.61	36.92	43.73
FDA *	88.96	39.53	80.23	22.58	29.73	32.78	33.64	26.66	80.06	25.39	73.63	36.78	10.91	77.82	26.35	46.14	1.37	22.80	50.31	37.71	42.40
IAST[123]	83.20	37.84	82.63	36.00	21.59	32.34	43.48	44.69	84.92	36.51	88.77	59.71	28.04	84.34	32.64	38.66	2.52	31.27	35.57	46.00	47.62
IAST*	76.62	32.39	83.04	**37.52**	23.43	28.96	39.11	39.47	81.33	26.02	**89.10**	56.83	26.41	82.36	18.95	38.16	**23.03**	21.14	44.22	42.66	45.69
DACS[174]	82.93	29.50	69.67	31.58	24.87	18.17	20.71	17.43	69.69	8.54	64.06	32.17	9.78	76.99	36.40	44.26	0.00	8.64	30.39	26.54	35.57
DACS *	45.03	18.55	24.01	9.80	12.25	10.14	13.08	5.62	46.05	4.23	23.95	14.94	8.64	52.14	36.28	12.43	0.00	8.35	15.08	16.22	18.98
Ours(M)	93.60	60.14	85.64	34.57	25.27	33.67	34.67	41.84	83.03	2.67	86.96	60.15	2.34	87.25	52.06	47.66	0.00	17.81	42.53	34.76	46.94
Ours(M+SLM)	93.33	57.28	86.14	36.66	29.25	36.84	43.25	43.09	85.50	**39.17**	85.85	63.47	26.95	88.71	52.76	53.06	0.00	**41.46**	57.13	52.28	53.68
Ours(M+SLM+CT)	93.83	60.53	86.37	30.73	35.05	36.69	41.74	47.82	**85.70**	38.69	85.75	62.65	36.28	87.89	51.00	52.84	0.00	39.71	59.11	53.69	54.34
Ours(M+SLM+UCT)	**94.51**	**62.40**	87.15	29.95	35.96	37.96	44.17	52.17	84.56	34.33	84.80	65.79	37.41	**90.03**	**56.10**	52.57	0.00	40.46	59.82	53.73	55.27
Ours(M+SLM+UCT+RL)	93.97	59.71	**87.58**	29.81	**36.26**	**38.81**	**45.38**	**52.53**	85.26	35.18	87.28	**66.58**	**38.74**	89.74	55.23	**54.72**	0.00	40.72	**60.47**	**54.49**	**55.68**
n^t=2975	93.65	56.25	86.48	27.37	39.02	37.59	43.73	50.49	87.08	49.25	86.38	67.71	43.83	89.40	50.98	47.01	0.09	45.42	63.96	59.54	56.09
Oracle [191]	96.7	75.7	88.3	46.0	41.7	42.6	47.9	62.7	88.8	53.5	90.6	69.1	49.7	91.6	71.0	73.6	45.3	52.0	65.5	63.1	65.9

Table 4.4: Coarse-to-Fine Taxonomy: GTA5→Cityscapes. There are 3 classes in the GTA5 dataset fine-grained into 6 classes in the Cityscapes dataset. The gray columns are the 6 fine-grained classes in the Cityscapes and corresponding mean IoU of these classes. "M": BMS. "*" with SLM module.

with the previous SOTA few-shot cross domain semantic segmentation strategy, *i.e.*, fine-tuning [211] and pseudo-label [129]. Instead, our proposed BMS can augment and utilize the few-shot supervision effectively, guiding the model prediction when using SLM.

Partially Labeled/Oracle. In Table 4.3, Table 4.4, and Table 4.5, under the open, coarse-to-fine and implicitly-overlapping taxonomy settings, we report the partially labeled performance where inconsistent taxonomy classes are labeled in all the available target domain image samples, *i.e.*, $n^t = 2975$. Compared to the few-shot performance, the partially labeled performance is further improved due to more labeled samples on the target domain being available. But there is still gap to the fully supervised oracle performance on the target domain. It shows that our method serves as a strong baseline, but still provides the potential to develop stronger algorithms for the TACS problem.

Effect of Few-shot Samples Number. In order to analyze the effect of the number of few-shot samples in the target domain for the inconsistent taxonomy adaptation performance, we take the open taxonomy setting as the example, and show the performance change with different number of few-shot samples in Fig. 4.3. It is shown that the inconsistent taxonomy class adaptation performance is improved, when more few-shot labeled samples are available.

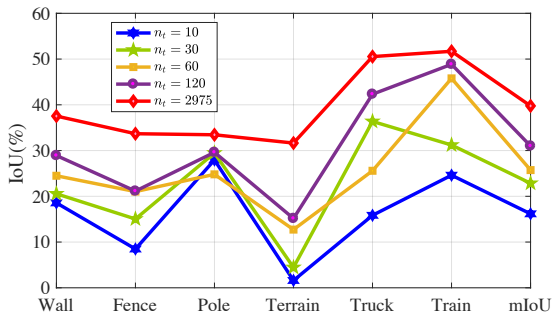


Figure 4.3: Performance of inconsistent taxonomy classes under open taxonomy setting, varying n^t.

Contrastive Learning. In Fig. 4.4, the performance when varying the number of negative samples in the contrastive learning is shown. It is observed that the performance increases as more samples are taken. Balancing the performance and memory, we adopt 100 samples per class. In Fig. 4.5, we compare the t-SNE visualization [179] of the feature embedding of the model trained with/without UCT, taking

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Vehicle	PT	Cycle	mIoU	mIoU
Source	82.74	43.14	70.95	29.04	19.24	33.99	34.47	36.29	81.90	28.67	86.61	55.17	28.25	54.75	1.75	34.99	30.50	45.12
Source*	87.95	40.99	74.68	24.35	22.67	32.17	31.86	34.74	81.53	27.52	83.74	55.08	26.68	67.51	11.34	21.56	33.47	45.27
ADVENT[184]	92.84	54.32	82.54	31.40	25.90	37.67	38.92	40.55	85.46	35.95	87.69	58.12	29.75	73.19	2.42	3.23	26.28	48.75
ADVENT*	90.02	46.16	80.37	27.90	24.56	35.69	31.48	37.81	83.96	38.81	84.83	54.73	30.69	73.67	16.02	18.80	36.16	48.47
FDA[200]	89.45	44.66	75.82	28.3	27.91	37.89	41.09	49.91	83.78	26.17	83.50	61.24	39.37	65.35	6.32	26.56	32.74	49.21
FDA *	86.86	43.56	75.32	28.01	27.68	38.50	39.50	50.31	83.80	21.69	83.93	63.45	**42.32**	80.99	10.96	42.64	44.86	51.22
IAST[123]	91.65	54.26	81.82	31.61	28.48	35.33	42.83	46.74	85.67	41.89	89.47	57.51	32.77	75.78	31.13	50.45	52.45	54.84
IAST *	93.00	55.31	83.55	32.80	30.49	38.21	46.04	53.09	86.46	41.91	88.57	60.58	29.17	83.18	39.01	36.76	52.98	56.13
DACS[174]	89.72	61.93	57.59	28.87	26.87	33.42	41.44	41.14	84.57	41.96	86.49	57.94	25.36	59.88	2.13	19.63	27.21	47.43
DACS *	82.27	41.83	13.43	17.67	18.84	23.23	23.93	23.54	56.89	18.20	68.49	44.60	13.75	22.09	2.39	16.75	13.74	30.49
Ours(M)	91.35	59.29	**86.81**	**34.60**	32.14	43.9	49.29	55.8	83.51	**42.28**	90.44	**67.98**	37.27	83.01	16.89	**43.92**	47.94	57.40
Ours(M+SLM)	93.66	65.25	81.31	28.81	26.43	44.96	51.70	55.84	87.59	38.47	88.80	67.93	35.10	87.71	35.55	36.29	53.18	57.84
Ours(M+SLM+CT)	**95.70**	**70.24**	85.42	29.16	25.78	42.10	49.77	54.14	**87.67**	42.11	90.10	66.59	36.67	87.55	34.97	40.43	54.32	58.65
Ours(M+SLM+UCT)	92.43	66.46	82.25	32.24	**32.47**	45.37	**52.29**	**57.15**	87.20	36.48	**91.85**	65.03	37.87	88.53	41.95	38.11	56.20	59.23
Ours(M+SLM+UCT+RL)	92.47	65.40	83.21	33.33	30.87	**45.94**	49.86	55.86	87.23	39.50	91.30	66.56	39.87	**88.75**	**42.59**	39.64	**56.99**	**59.52**
n^t=2975	94.62	63.90	85.13	28.52	31.03	46.46	53.44	50.16	86.98	41.21	91.00	67.61	35.04	89.98	74.72	52.85	72.52	62.04
Oracle	96.79	76.53	87.75	49.21	41.14	40.64	43.82	60.49	88.01	52.68	89.16	68.68	49.33	91.05	74.69	64.26	76.67	67.14

Table 4.5: Implicitly-Overlapping Taxonomy: Synscapes→Cityscapes. There are 3 classes (in gray) in the Cityscapes corresponding to the implicitly-overlapping taxonomy. "M": BMS. "*": with SLM.

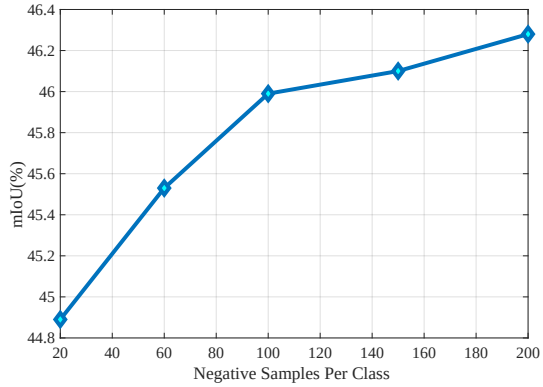
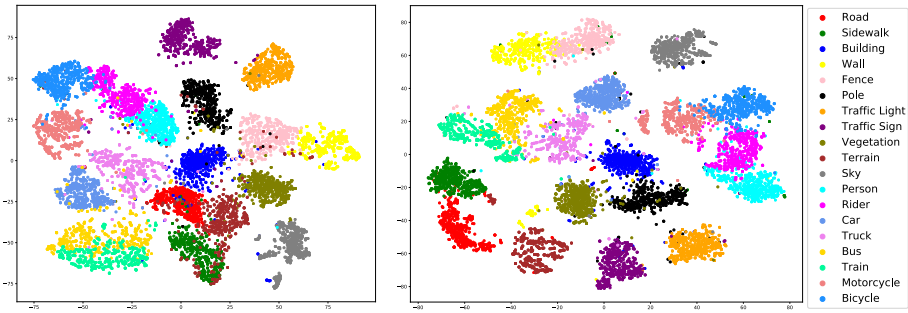


Figure 4.4: Negative samples number study for contrastive learning, under M+UCT in Table 4.3.



(a) Without Contrastive (b) With Contrastive

Figure 4.5: t-SNE visualization of the features with/without contrastive learning under the open taxonomy setting.

open taxonomy setting as example. It verifies the contrastive learning is helpful to train the cross-domain invariant and class-discriminative features.

4.5 CONCLUSION

We propose the new TACS problem, allowing inconsistent taxonomies between the source and the target domain in the cross-domain semantic segmentation. Three typical types of inconsistent taxonomies are identified. To resolve TACS, the mixed-sampling, pseudo-label

Figure 4.6: Qualitative semantic segmentation results on the target domain under different inconsistent taxonomy settings, open taxonomy, coarse-to-fine taxonomy and implicitly-overlapping taxonomy. (a) shows the RGB target domain image. (b) gives the ground truth semantic segmentation map. (c) is the semantic segmentation result without adaptation. (d) is the semantic segmentation result adapted by the IAST [123] method. (e) is the semantic segmentation result adapted by our proposed method. Refer to the red box region for the adaptation results of the inconsistent taxonomy classes.

and contrastive learning based techniques are developed. Extensive experiments prove the effectiveness of our approach.

5

PSEUDO-LABEL RECTIFICATION WITH IMPLICIT NEURAL REPRESENTATIONS

This chapter corresponds to our published article: Rui Gong, Qin Wang, Martin Danelljan, Dengxin Dai, and Luc Van Gool. „Continuous Pseudo-Label Rectified Domain Adaptive Semantic Segmentation With Implicit Neural Representations." In: *CVPR*. 2023

In this chapter, we present a novel approach based on implicit neural representations to rectify pseudo-labels in domain adaptation, boosting the effectiveness of knowledge transfer across different domains. While previous methods have shown impressive progress using pseudo-labels on unlabeled target domain, the presence of low-quality pseudo-labels, stemming from domain discrepancies, poses a challenge to adaptation. Addressing this issue requires effective and accurate strategies for estimating the reliability of pseudo-labels to rectify them. Motivated by this challenge, we propose to estimate the rectification values of the predicted pseudo-labels with implicit neural representations. We view the rectification value as a signal defined over the continuous spatial domain. Taking an image coordinate and the nearby deep features as inputs, the rectification value at a given coordinate is predicted as an output. This allows us to achieve high-resolution and detailed rectification values estimation, important for accurate pseudo-label generation at mask boundaries in particular. The rectified pseudo-labels are then leveraged in our rectification-aware mixture model (RMM) to be learned end-to-end and help the adaptation. We demonstrate the effectiveness of our approach on different UDA benchmarks, including synthetic-to-real and day-to-night.

5.1 INTRODUCTION

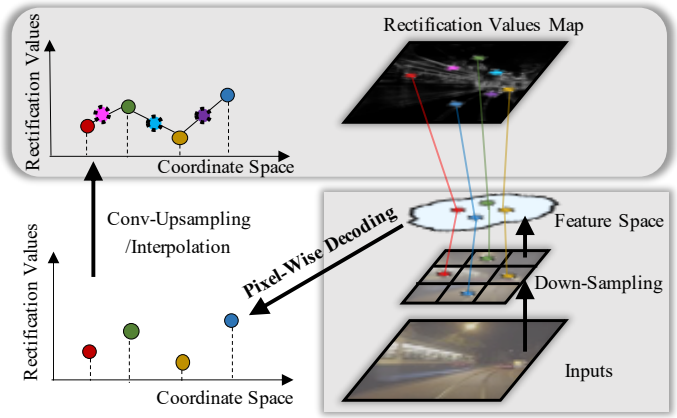
Different from the predominant UDA methods that explicitly align the source and target distributions on the image-level [200, 72, 60, 105] or the feature-level [184, 175, 176], pseudo-labeling or self-training [227, 226, 212, 75, 76, 174] has recently emerged as a simple yet effective approach for UDA. Pseudo-labeling approaches typically first generate

pseudo-labels on the unlabeled target domain using the current model. The model is then fine-tuned with target pseudo-labels in an iterative manner. However, some pseudo-labels are inevitably incorrect because of the domain shift. Therefore, pseudo-label correction, or rectification, is critical for the adaptation process. This is typically implemented in the literature by removing [227, 226] or assigning a smaller weight [220, 212, 76, 190] to pixels with low-quality and potentially incorrect pseudo-labels. The key problem is thus to formulate a *rectification function that estimates the pseudo-label quality*. We identify two important issues with current approaches.

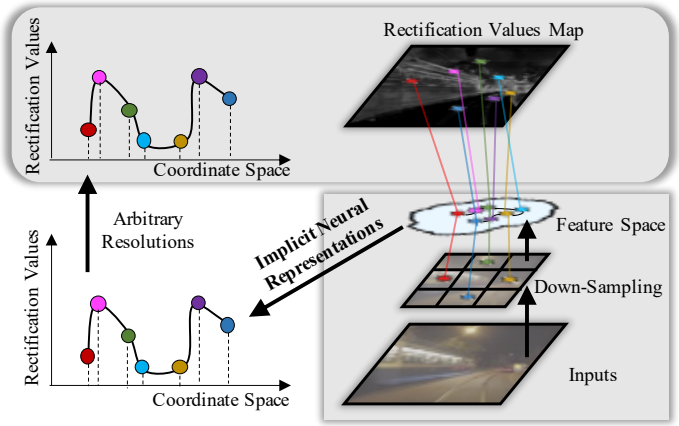
First, most existing methods use *hard-coded heuristics* as the rectification function, *e.g.*, hard thresholding of the softmax confidence [226, 227], prediction variances of different learners [220], or distance to prototypes [212, 190]. These heuristic rectification functions assume on strong correlations between the function and the pseudo-label quality, which may not be the case. For example, the rectification function that uses the variance of multiple learners [220] to suppress disagreement on the pseudo-labels can be sensitive to small objects in the adaptation [76].

The second issue is that the existing works [76] typically model the rectification function in a *discrete* spatial grid (see Fig.5.1). Rectification values are predicted by the pixel-wise decoding from the fixed-grid feature space, which is constrained by the limited resolution. This is especially harmful when the objects in the test images are of a different scale than in the training, since the rectification function cannot generalize well on these unseen scales (see Fig.5.1). Existing approaches also lose vital high-frequency information through down-/up-sampling operations [76, 164, 77, 121], which may lead to poorer pseudo-labels, in particular close to mask boundaries.

To address these two issues, we propose a novel continuous rectification-aware mixture model (RMM). **First**, instead of formulating the rectification function with heuristics and priors, we propose a principled mixture model representation, *i.e.*, rectification-aware mixture model (RMM), ensuring a probabilistic end-to-end ***learnable*** formulation. **Second**, the rectification function in RMM is represented by our proposed implicit rectification-representative function (IR^{2}F), to model the pixel-wise rectification of pseudo-labels in ***continuous*** spatial coordinates, *i.e.*, *continuous RMM*. The primary idea of IR^{2}F is to learn pixel-wise rectification values as latent codes, which are decoded at arbitrary continuous spatial coordinates. Given a queried coordinate, our IR^{2}F inputs latent



(a) Discrete Modeling



(b) Continuous Modeling

Figure 5.1: **Discrete *vs.*, Continuous Rectification Function Modeling.** Discrete modeling suffers from the convolutional pixel-wise decoding in the fixed-grid, where some coordinates are missing (see dashed circle in (a)). Thus, the rectification values corresponding to these coordinates can only be obtained by upsampling/interpolation, which is constrained by the blurring effect and induces the inaccurate rectification values estimation in some areas, *e.g.*, mask boundaries. In contrast, our continuous modeling decodes the features – in the continuous coordinate space – into rectification values, which can be generalized to arbitrary resolution and preserve finer details. (The coordinate space and rectification values are shown in 1-D axis just for better viewing.)

codes around the given coordinate from the different learners (*e.g.*, high-/low-resolution decoder in [76] and primary/auxiliary classifier in [220]) along with their spatial coordinates. IR^{2}F then predicts the rectification value at the queried coordinate. Our *principled* formulation is a general *plug-in* module, compatible with different rectification-aware UDA architectures.

We thoroughly analyze our continuous RMM on different UDA benchmarks, including *synthetic-to-real* and *day-to-night* settings. Extensive experimental results demonstrate the effectiveness of continuous RMM, outperforming the previous state-of-the-art (SOTA) methods by a large margin, including on SYNTHIA→Cityscapes (**+1.9%** mIoU), Cityscapes→Dark Zurich (**+3.0%** mIoU) and ACDC-Night (**+3.4%** mIoU). Overall, continuous RMM reveals the significant potential of modeling pseudo-labels rectification for UDA in the learnable and continuous manner, inspiring further research in this field.

5.2 RELATED WORK

Unsupervised Domain Adaptation (UDA). UDA for semantic segmentation aims at adapting the model from the labeled source domain to the unlabeled target domain. To this end, different strategies are proposed, which can be generally categorized into two classes: 1) *adversarial learning* based algorithms make use of domain discriminator to align the domain distributions on the images inputs space [142, 43, 126], features space [73] and outputs space [118, 175, 185]; 2) *pseudo-labeling (or self-training)* based algorithms typically generate pseudo-labels on the unlabeled target domain. To avoid the error accumulation caused by noisy pseudo-label drift, different approaches have been developed for pseudo-label rectification, *e.g.*, confidence thresholding [227, 226], uncertainty estimation [220, 190] and pseudo-label prototypes [212, 190]. These methods formulate the pseudo-label rectification function as hard-coded heuristics, while our method formulates the rectification function in the end-to-end learnable manner.

Implicit Neural Representations (INR). Implicit neural representations are originally proposed for 3D reconstruction, where object shapes [29, 4, 64, 135, 201], scene surfaces [168, 85, 139, 204] and structure appearances [124, 6, 122, 221] are represented as a multi-layer perceptron (MLP). The core idea is to map coordinates to signals with MLP. Very recently, the vast success of implicit neural representations

in 3D reconstruction motivates the further exploration in 2D tasks, *e.g.*, image representations [167, 29], image super-resolution [28, 199], and feature alignment [77]. Different from previous methods that explore the in-domain learning, we focus on leveraging implicit neural representations to rectify pseudo-labels to help the cross-domain adaptation.

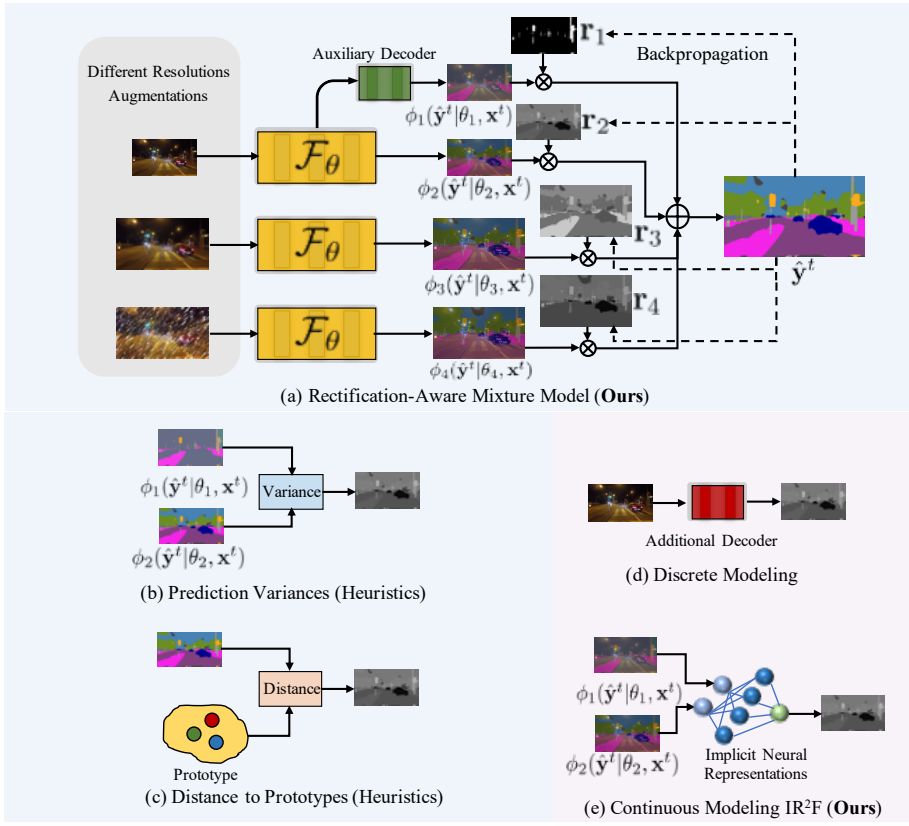


Figure 5.2: **Rectification-Aware Mixture Model (RMM) and Different Rectification Function Modeling.** Our rectification function is learned end-to-end by our proposed RMM as shown in (a), without relying on the predefined heuristics in (b) and (c). Moreover, rectification function in our RMM is modeled in the continuous manner, by the proposed implicit rectification-representative function (IR^{2}F) in (e), to overcome the resolution limitation of the fixed-grid discrete modeling in (d).

5.3 METHOD

5.3.1 *Preliminary*

In UDA problem, we are given the well-labeled source domain, $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$, and the unlabeled target domain, $\mathcal{D}_t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$, where $\mathbf{x}^s, \mathbf{x}^t \in \mathbb{R}^{H\times W\times 3}$ are RGB images while $\mathbf{y}^s \in \{0,1\}^{H\times W\times C}$ is the C-class semantic label map associated with $\mathbf{x}^s$. The goal of UDA is to train the semantic segmentation model $\mathcal{F}_\theta$ on $\mathcal{D}_s, \mathcal{D}_t$ and evaluate $\mathcal{F}_\theta$ on the target domain testing data.

Since the ground truth label $\mathbf{y}^t$ corresponding to $\mathbf{x}^t$ is not available, the pseudo-labeling (or self-training) strategy for UDA generates pseudo-labels by, $\hat{\mathbf{y}}^{t(i,j,c)} = [c = \text{argmax}\, \mathcal{F}_\theta(\mathbf{x}^t)^{(i,j)}]$, where (i, j, c) represents (row, column, class) index and $[\cdot]$ is the Iverson bracket. Then $\mathcal{F}_\theta$ is trained by, $\mathcal{L}_{ce} = CE(\mathcal{F}_\theta(\mathbf{x}^t), \hat{\mathbf{y}}^t) + CE(\mathcal{F}_\theta(\mathbf{x}^s), \mathbf{y}^s)$, where $CE(\cdot)$ denotes the cross-entropy loss. As pseudo labels $\hat{\mathbf{y}}^t$ are not necessarily correct, different schemes are advocated to rectify pseudo labels, where the rectification function is denoted as $\mathcal{H}(\cdot)$. Most existing pseudo-label rectifying methods can be categorized into one of the following three types, 1) weighting pseudo-label based cross-entropy loss with the estimated rectification values $\mathcal{H}(\mathbf{x}^t)$ [220], *i.e.*, $\mathcal{L}_{ce}^t = \mathcal{H}(\mathbf{x}^t) \odot CE(\mathcal{F}_\theta(\mathbf{x}^t), \hat{\mathbf{y}}^t)$; 2) weighting soft pseudo-labels with the estimated rectification values $\mathcal{H}(\mathbf{x}^t)$ [212, 190], *i.e.*, $\hat{\mathbf{y}}^{t(i,j,c)} = [c = \text{argmax}(\mathcal{H}(\mathbf{x}^t)^{(i,j)} \odot \mathcal{F}_\theta(\mathbf{x}^t)^{(i,j)})]$; 3) averaging pseudo-labels from multiple K learners (*e.g.*, decoders) [79, 7, 187] to rectify pseudo labels of each single learner, *i.e.*, $\hat{\mathbf{y}}^{t(i,j)} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{F}_{\theta^k}(\mathbf{x}^t)^{(i,j)}$, where $\odot$ denotes the element-wise multiplication.

In general, such pseudo-labeling-based approaches can be categorized into *non-ensemble* (type 1 and 2) and *ensemble* based solutions (type 3). In the domain adaptation and generalization field, numerous empirical and theoretic comparisons [224, 18, 1, 107, 118, 79] between these two classes have been conducted before and after the deep learning revolution. The consensus is that ensembles can take advantage of different ensemble members (*e.g.*, different data augmentation, different resolutions image and different level features as shown in Fig. 5.2) to adaptively filter pseudo-label noise, and have the potential to overcome the problem of mode collapse/overfitting [18, 89, 187, 150] in non-ensemble methods. Thus, the *ensemble* method is particularly remarkable and taken as the test-bed in this work.

5.3.2 *Rectification-Aware Mixture Model*

The key is *how to formulate a rectification function to estimate the pseudo-labels quality*. Instead of utilizing hard-coded heuristics and priors as the rectification function, we propose a *principled end-to-end learnable* formulation. Based on the fact that existing methods make use of multiple members (auxiliary classifiers/decoders, prototypes, different images resolutions/augmentations) to rectify the models [220, 76], we reformulate the pseudo-labels rectification problem in principled manner as learning a rectification-aware mixture model (RMM), drawing inspiration from mixture density networks (MDN) [9] and deep ensembles [91, 131]. In RMM, each mixture member is weighted by the rectification function, which is the measurement of pseudo-labels quality of the corresponding member, formulated as,

$$p(\hat{\mathbf{y}}^t|\mathbf{x}^t) = \sum_{k=1}^{K} \mathbf{r}_k \phi_k(\hat{\mathbf{y}}^t|\theta_k, \mathbf{x}^t)\,, \tag{5.1}$$

where K is the number of mixture members, $\phi_k(\cdot|\theta_k)$ denotes an arbitrary parametric distribution conditioned on parameters θ_k, and $\mathbf{r}_k = \mathcal{H}(\mathbf{x}^t)$ are the estimated rectification values by the rectification function $\mathcal{H}(\cdot)$, satisfying $\sum_{k=1}^{K} \mathbf{r}_k = 1$. Specifically, the primary/auxiliary decoders in [220], the high-resolution/low-resolution image decoders in [76] and the different data augmentation techniques in [1] can be seen as $\phi(\cdot|\theta_k)$ in Eq. (5.1), as shown in Fig. 5.2. Benefiting from RMM, the rectification function $\mathcal{H}(\cdot)$ can be learned in the end-to-end way.

5.3.3 *Implicit Rectification-Representative Function*

In this section, we first introduce how to model the rectification function continuously with implicit neural representations, and then leverage the continuous rectification function in RMM to obtain the continuous RMM.

Continuous Rectification Function Modeling with IR^{2}F. *Representing rectification function* $\mathcal{H}(\cdot)$ is the core part of building a rectification-aware mixture model. Current approaches essentially model rectification function in a *discrete* way. They compute rectification values on a pre-defined discrete grid, often using convolutional decoders and disregarding intermediate locations. For example, as shown in Fig. 5.2, [76] introduces an

additional convolutional decoder, as $\mathcal{H}(\cdot)$, to predict $\mathbf{r}_k$ on the discrete fixed-grid. However, this leads to coarse and over-smoothed outputs due to the low resolution and up/down-sampling stages in the decoder. On the other had, spatially detailed rectification values are important in order to achieve high-quality pseudo labels, especially at mask boundaries [76, 19]. To overcome the problems and get spatially accurate rectification values, the key idea of this work is to employ the *continuous* rectification function modeling mechanism, which is *learnable* and then decoded at continuous spatial coordinates in *arbitrary* resolution.

To this end, our proposed implicit rectification-representative function (IR^{2}F) views the pixel-wise rectification value $\mathbf{r}_k$ as a *continuous* signal in the 2D coordinate space. Inspired by implicit neural representations [124, 168] for 3D shape reconstruction and 2D image super-resolution [28, 199], our implicit rectification-representative function (IR^{2}F) aims at learning the implicit function $f_{\theta'}$ to decode the feature map $\mathcal{G}(\mathbf{x}^t)$ into the pixel-wise rectification values $\mathbf{r}_k$. That is, $\mathcal{H}(\cdot)$ in Sec. 5.3.2 is represented by $f_{\theta'}$. $\mathbf{r}_k$ is continuously decoded in the 2D coordinate space $\mathcal{O}$, formulated as,

$$\mathbf{r}^{\mathbf{o}_q} = (\mathbf{r}_k^{\mathbf{o}_q})_{k=1}^K = f_{\theta'}(\mathcal{G}(\mathbf{x}^t)^*, \mathbf{o}_q - \mathbf{o}^*), \tag{5.2}$$

where $\mathbf{o}_q \in \mathcal{O}$ is a queried 2D coordinate in the continuous coordinate space $\mathcal{O}$, and $(\mathbf{r}_k^{\mathbf{o}_q})_{k=1}^K = (\mathbf{r}_1^{\mathbf{o}_q}, \ldots, \mathbf{r}_K^{\mathbf{o}_q})$ is the predicted rectification values for all ensamble members at location $\mathbf{o}_q$. $f_{\theta'}$ is parameterized by θ' as a multi-layer perceptron (MLP). $\mathcal{G}(\mathbf{x}^t)^*$ is the nearest feature vector from $\mathbf{o}_q$ in $\mathcal{G}(\mathbf{x}^t)$, and $\mathbf{o}^*$ is the 2D coordinate of $\mathcal{G}(\mathbf{x}^t)^*$ in $\mathcal{O}$. IR^{2}F can be seen as the mapping from the coordinate space to the rectification value space, *i.e.*, $f_{\theta'}(\mathcal{G}(\mathbf{x}^t), \cdot) : \mathcal{O} \rightarrow \mathcal{R}$.

Spatial Encoding. As noticed by previous works [167, 199], directly inputting the spatial coordinates to an MLP of the implicit neural representation leads to a loss of high-frequency content. However, the high-frequency information, *e.g.*, the edge information between the objects, is crucial to UDA for semantic segmentation as pointed out in [19, 76, 109]. In order to overcome this shortcoming, following [199, 77], we employ a spatial encoding of the spatial coordinates, before it is fed into the MLP of our IR^{2}F in Eq. (5.2). We use a sinusodal positional encoding,

$$\psi(\mathbf{o}) = (\sin(\omega_1 \mathbf{o}), \cos(\omega_1 \mathbf{o}), \ldots, \sin(\omega_n \mathbf{o}), \cos(\omega_n \mathbf{o})), \tag{5.3}$$

$$\mathbf{r}^{\mathbf{o}_q} = f_{\theta'}(\mathcal{G}(\mathbf{x}^t)^*, \psi(\mathbf{o}_q - \mathbf{o}^*), \mathbf{o}_q - \mathbf{o}^*). \tag{5.4}$$

where the frequencies $\omega_1, \omega_2, \ldots, \omega_n$ are learnable during training and n is the spatial encoding dimension.

Continuous RMM based on IR^{2}F. Benefiting from the continuous rectification function modeling with IR^{2}F in Sec. 5.3.3, rectification values of our proposed RMM in Sec. 5.3.2 are predicted in the continuous coordinate space, and can be generalizable to arbitrary resolution. Moreover, to take advantage of multiple learners in our RMM, the input representation $\mathcal{G}(\mathbf{x}^t)$ in Eq. (5.4) is obtained by stacking the feature information from different ensemble members,

$$\mathcal{G}(\mathbf{x}^t) = \text{Concat}(\mathcal{G}_1(\mathbf{x}^t), \mathcal{G}_2(\mathbf{x}^t), \ldots, \mathcal{G}_K(\mathbf{x}^t)), \tag{5.5}$$

Then rectification values for RMM are obtained by substituting Eq. (5.5) into Eq. (5.4). Therefore, considering Eq. (5.4) and Eq. (5.1), the continuous RMM can be formulated as,

$$p(\hat{\mathbf{y}}^t|\mathbf{x}^t, \mathbf{o}_q) = \sum_{k=1}^{K} \mathbf{r}_k^{\mathbf{o}_q} \phi_k(\hat{\mathbf{y}}^t|\theta_k, \mathbf{x}^t). \tag{5.6}$$

Here, $p(\hat{\mathbf{y}}^t|\mathbf{x}^t, \mathbf{o}_q)$ is the predicted class distribution at spatial location $\mathbf{o}_q$. The rectification values $\mathbf{r}_k^{\mathbf{o}_q}$ can thus be queried at any pixel coordinate, by the continuous implicit neural representations $f_{\theta'}$.

5.3.4 *IR^{2}F-RMM Rectified Self-Training*

Our proposed continuous RMM based on IR^{2}F can be used as a *plug-in* strategy, to promote and rectify the pseudo-labels used for self-training in UDA. In this section, we introduce how our continuous RMM can be plugged into two popular UDA frameworks.

HRDA. HRDA [76] is a multi-resolution inputs framework for UDA semantic segmentation, fusing the predictions of low-/high-resolution (LR/HR) inputs to capture both the long-range context from LR and the detailed knowledge from HR. Our continuous RMM module can be plugged into the HRDA framework by considering the two resolution branches as two mixture members, as shown in Fig. 5.3. Rectified pseudo-labels $\hat{\mathbf{y}}^t$ can then be formally written as,

$$\begin{aligned}
&\mathbf{r}^{\mathbf{o}_q} = f_{\theta'}(\text{Concat}(\mathcal{G}_{lr}(\mathbf{x}^t), \mathcal{G}_{hr}(\mathbf{x}^t))^*, \psi(\mathbf{o}_q - \mathbf{o}^*), \mathbf{o}_q - \mathbf{o}^*),\\
&\tilde{\mathbf{y}}_{lr}^t = \phi_1(\hat{\mathbf{y}}^t|\theta_k, \mathbf{x}^t), \quad \tilde{\mathbf{y}}_{hr}^t = \phi_2(\hat{\mathbf{y}}^t|\theta_k, \mathbf{x}^t),\\
&\tilde{\mathbf{y}}^t = \mathbf{r}\tilde{\mathbf{y}}_{lr}^t + (1 - \mathbf{r})\tilde{\mathbf{y}}_{hr}^t,\\
&\hat{\mathbf{y}}^{t(i,j,c)} = [c = \text{argmax}\, \tilde{\mathbf{y}}^{t(i,j)}],
\end{aligned} \tag{5.7}$$

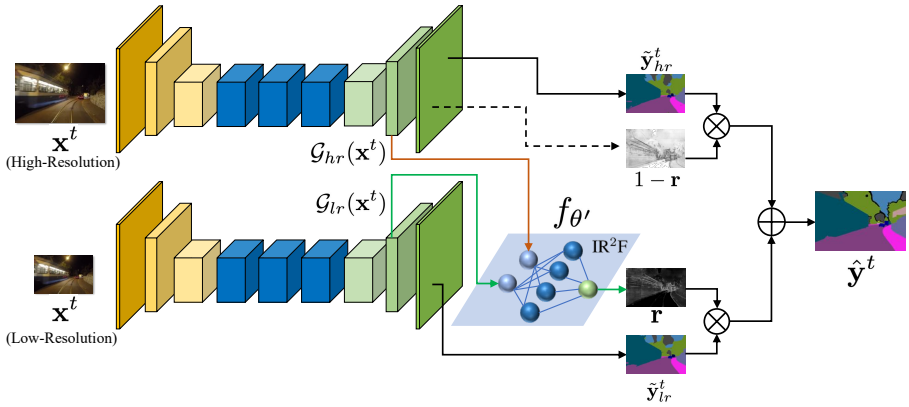


Figure 5.3: **Plugging continuous RMM into HRDA.**

where $\tilde{\mathbf{y}}^t_{lr}, \tilde{\mathbf{y}}^t_{hr}$ are soft pseudo-labels predicted by low-/high-resolutions branches, resp. $\mathcal{G}_{lr}(\mathbf{x}^t), \mathcal{G}_{hr}(\mathbf{x}^t)$ are feature maps from low-/high-resolutions branches, resp. Concat is realized by firstly up-sampling with bi-linear interpolation, and then pixel-wise concatenation. (i, j, c) are the (row, column, class) index, and $[\cdot]$ is the Iverson bracket.

MRNet. MRNet [220] is a rectification-aware UDA framework, where there are primary and auxiliary classifiers. In MRNet, the variances between the primary and auxiliary classifiers are used as the rectification values. Our continuous RMM can be used to replace this rule and instead learn the rectification. By inserting the continuous RMM into MRNet, the pseudo-labels $\hat{\mathbf{y}}^t$ can be written as,

$$
\begin{aligned}
&\mathbf{r}^{\mathbf{o}_q} = f_{\theta'}(\text{Concat}(\mathcal{G}_{pr}(\mathbf{x}^t), \mathcal{G}_{aux}(\mathbf{x}^t))^*, \psi(\mathbf{o}_q - \mathbf{o}^*), \mathbf{o}_q - \mathbf{o}^*),\\
&\tilde{\mathbf{y}}^t_{pr} = \phi_1(\hat{\mathbf{y}}^t | \theta_k, \mathbf{x}^t), \quad \tilde{\mathbf{y}}^t_{aux} = \phi_2(\hat{\mathbf{y}}^t | \theta_k, \mathbf{x}^t),\\
&\tilde{\mathbf{y}}^t = \mathbf{r}\tilde{\mathbf{y}}^t_{pr} + (1-\mathbf{r})\tilde{\mathbf{y}}^t_{aux},\\
&\hat{\mathbf{y}}^{t(i,j,c)} = [c = \text{argmax}\, \tilde{\mathbf{y}}^{t(i,j)}],
\end{aligned}
\tag{5.8}
$$

where $\tilde{\mathbf{y}}^t_{pr}, \tilde{\mathbf{y}}^t_{aux}$ are the soft pseudo-labels from the primary and auxiliary classifiers, resp. $\mathcal{G}_{pr}(\mathbf{x}^t), \mathcal{G}_{aux}(\mathbf{x}^t)$ are feature maps from primary and auxiliary classifiers, resp.

Rectified Pseudo-Labels based Self-Training Loss. With pseudo-labels $\hat{\mathbf{y}}^t$ rectified by our continuous RMM, the semantic segmentation network $\mathcal{F}_\theta$ and our implicit neural representations $f_{\theta'}$ are trained jointly in

the end-to-end manner, through the standard cross-entropy loss written as,

$$\mathcal{L} = CE(\mathcal{F}_\theta(\mathbf{x}^t), \hat{\mathbf{y}}^t). \tag{5.9}$$

5.4 EXPERIMENTS

In this section, we demonstrate the effectiveness of our continuous RMM for UDA semantic segmentation on different benchmarks, synthetic-to-real and day-to-night. We compare our continuous RMM to other heuristics-based and/or discrete rectification modeling methods, to show the benefits of our learnable and continuous rectification function modeling based on RMM and IR^{2}F.

5.4.1 *Experimental Setup*

Datasets. We use the conventional notation A→B do describe the domain adaptation task, where A is the labeled source domain and B is the unlabeled target domain. We consider four different tasks in two categories. *Syntheic-to-Real*: There are two settings, GTA [149] → Cityscapes [34] and SYNTHIA [152] → Cityscapes [34]. *Day-to-Night*: There are also two tasks, Cityscapes [34] → Dark Zurich [161] and Cityscapes [34] → ACDC-Night [160]. The datasets are described as follows.

GTA. GTA [149] is a synthetic urban-scene image dataset, rendered from the game engine. There are 24966 images included in the GTA dataset, which are of 1914×1052 pixels and are densely labeled with pixel-wise semantic segmentation annotations. The urban scene of GTA dataset is built based on the city of Los Angeles, thus with typical U.S. urban scene layout. Following previous UDA works [75, 76, 220, 175, 184, 227, 174, 212], the GTA images are resized to 1280×720 for low-resolution inputs [76], and to 2560×1440 for high-resolution inputs [76].

SYNTHIA. SYNTHIA [152] is a synthetic photo-realistic image dataset, whose images are rendered from a virtual city. We adopt SYNTHIA-RAND-Cityscapes dataset, which is built for street scene parsing and consists of 9400 densely labeled images. The images are of 1280×760 pixels. In accordance with previous UDA works [75, 76, 220, 175, 184, 227, 174, 212], the SYNTHIA images are resized to 2560×1520

for high-resolution inputs [76], and keep 1280×760 for low-resolution inputs.

Cityscapes. Cityscapes [34] is a real street-scene image dataset, collected from different European cities. We utilize the training set of Cityscapes during the training stage, consisting of 2975 images. And we use the validation set of Cityscapes, covering 500 images, to evaluate the model performance. Cityscapes images are of 2048×1024 pixels. The resolution is maintained for high-resolution inputs in experiments, and resized to 1024×512 for low-resolution inputs.

Dark Zurich. Dark Zurich [161] is a real nighttime urban-scene image dataset, which is captured in Zurich. We use the training set of Dark Zurich during the training stage, including 2416 images. And we utilize the test set of Dark Zurich, consisting of 151 images, to evaluate the model performance. The evaluation on the test set of Dark Zurich is only accessible through the online benchmark, where the ground truth is not publicly available. The images in Dark Zurich is of 1920×1080 pixels. The resolution is kept for high-resolution inputs, and is resized to 960×540 for low-resolution inputs.

ACDC-Night. ACDC [160] is a real street-scene image dataset under adverse conditions, *e.g.*, fog, snow, rain and nighttime. We adopt the nighttime subset of ACDC, *i.e.*, ACDC-Night, where there are 400 images as training set and 500 images as test set. Similar to the evaluation of Dark Zurich, the evaluation on the test set can only be conducted through the online benchmark, and the ground truth is not publicly available. The images in ACDC-Night is of 1920×1080 pixels. The resolution is kept for high-resolution inputs, and is resized to 960×540 for low-resolution inputs.

NightCity+. NightCity [170] is a real urban driving scene dataset, for nighttime scene parsing. The images in NightCity are collected from different cities around the world. NightCity+ [39] is the extended version of NightCity, where more accurate annotations in the validation set are provided compared to NightCity. We utilize the validation set of NightCity+, including 1299 images, to evaluate the model performance.

Implementation Details. *Framework and Backbone*: Our default framework is based on HRDA [76] with the MiT-B5 [196] backbone. We utilize AdamW [88, 117] optimizer, where betas of AdamW optimizer are (0.9, 0.999), the weight decay is 0.01, and learning rates of the encoder and decoder are set as 6×10^{-5}, 6×10^{-4}, respectively. The batch size is set as 2, and the linear learning rate warmup and DACS [174] data

augmentation in [76, 75] are adopted. In addition, the method is also evaluated with other backbones such as MRNet [220] (in Table 5.5), and ResNet-101 [68] (in Table 5.4). For all experiments, we simply insert our IR^{2}F based continuous RMM into the decoder without modifying the backbone architecture. *Implicit Neural Representations*: f'_θ in IR^{2}F is implemented with 4-layer MLP, with ReLU activation and hidden dimension as 256. *Training Details*: By default, we follow the training details of HRDA. In Table 5.5, we follow the training details of MRNet. The framework is implemented with PyTorch [138], and all the experiments are conducted on a TITAN RTX GPU.

5.4.2 *Experimental Results*

Comparison with SOTA UDA Methods. In Table 5.1 and Table 5.2, we compare our proposed IR^{2}F-based continuous RMM with other existing UDA semantic segmentation methods, under the synthetic-to-real and day-to-night benchmarks, respectively. As observed in Table 5.1, our IR^{2}F-based continuous RMM outperforms other SOTA methods for UDA semantic segmentation on the synthetic-to-real benchmark, especially by 1.9% mIoU under SYNTHIA $\rightarrow$ Cityscapes setting. As shown in Table 5.2, on the challenging day-to-night benchmark with a larger domain gap, our IR^{2}F-based continuous RMM shows a stronger performance improvement over existing SOTA methods for UDA nighttime segmentation, by 3.0% and 3.4% mIoU over the previous state-of-the-art under the Cityscapes $\rightarrow$ Dark Zurich and Cityscapes $\rightarrow$ ACDC-Night settings, respectively. Note that, the existing SOTA methods for UDA nighttime segmentation always require the day images in the target domain as the reference for adaptation (see Table 5.2). Instead, our IR^{2}F-based continuous RMM method does not need these auxiliary data, and still outperforms the SOTA methods by a large margin. It verifies the strong generalization ability of our proposed IR^{2}F-based continuous RMM compared to the existing SOTA UDA semantic segmentation methods, under different scenarios.

Generalization. Following [94], we further evaluate the adapted model (after domain adaptation) performance on another unseen dataset, to show the generalization ability of our proposed continuous RMM. More specifically, we take the trained model under Cityscapes$\rightarrow$Dark Zurich, and evaluate the trained model on the third dataset NightCity+. As shown in Table 5.3, our proposed approach strongly outperforms

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
GTA → Cityscapes																				
CBST [227]	91.8	53.5	80.5	32.7	21.0	34.0	28.9	20.4	83.9	34.2	80.9	53.1	24.0	82.7	30.3	35.9	16.0	25.9	42.8	45.9
MRNet [220]	90.4	31.2	85.1	36.9	25.6	37.5	48.8	48.5	85.3	34.8	81.1	64.4	36.8	86.3	34.9	52.2	1.7	29.0	44.6	50.3
DACS [174]	89.9	39.7	87.9	30.7	39.5	38.5	46.4	52.8	88.0	44.0	88.8	67.2	35.8	84.5	45.7	50.2	0.0	27.3	34.0	52.1
TACS [57]	93.0	55.9	87.9	38.2	38.8	40.4	42.1	54.5	87.5	46.7	87.8	66.3	33.7	90.2	47.5	54.2	0.0	41.2	53.3	55.8
CorDA [186]	94.7	63.1	87.6	30.7	40.6	40.2	47.8	51.6	87.6	47.0	89.7	66.7	35.9	90.2	48.9	57.5	0.0	39.8	56.0	56.6
BAPA [109]	94.4	61.0	88.0	26.8	39.9	38.3	46.1	55.3	87.8	46.1	89.4	68.8	40.0	90.2	60.4	59.0	0.0	45.1	54.2	57.4
ProDA [212]	87.8	56.0	79.7	46.3	44.8	45.6	53.5	53.5	88.6	45.2	82.1	70.7	39.2	88.8	45.5	59.4	1.0	48.9	56.4	57.5
EHTD [100]	95.4	68.8	88.1	37.1	41.4	42.5	45.7	60.4	87.3	42.6	86.8	67.4	38.6	90.5	66.7	61.4	0.3	39.4	56.1	58.8
UndoUDA [110]	92.9	52.7	87.2	39.4	41.3	43.9	55.0	52.9	89.3	48.2	91.2	71.4	36.0	90.2	67.9	59.8	0.0	48.5	59.3	59.3
CPSL [101]	92.3	59.9	84.9	45.7	29.7	52.8	61.5	59.5	87.9	41.5	85.0	73.0	35.5	90.4	48.7	73.9	26.3	53.8	53.9	60.8
DDB [23]	95.3	67.4	89.3	44.4	45.7	38.7	54.7	55.7	88.1	40.7	90.7	70.7	43.1	92.2	60.8	67.6	34.2	48.7	63.7	62.7
DAFormer [75]	95.7	70.2	89.4	53.5	48.1	49.6	55.8	59.4	89.9	47.9	92.5	72.2	44.7	92.3	74.5	78.2	65.1	55.9	61.8	68.3
HRDA [76]	96.4	74.4	**91.0**	**61.6**	51.5	**57.1**	**63.9**	69.3	91.3	48.4	**94.2**	**79.0**	**52.9**	93.9	84.1	85.7	75.9	**63.9**	67.5	73.8
IR^{2}F-RMM (Ours)	**97.5**	**80.0**	**91.0**	60.0	**53.3**	56.2	**63.9**	**72.4**	**91.7**	**51.0**	**94.2**	**79.0**	51.1	**94.3**	**84.7**	**86.7**	**75.9**	62.6	**67.8**	**74.4**
SYNTHIA → Cityscapes																				
CBST [227]	68.0	29.9	76.3	10.8	1.4	33.9	22.8	29.5	77.6	–	78.3	60.6	28.3	81.6	–	23.5	–	18.8	39.8	42.6
MRNet [220]	87.6	41.9	83.1	14.7	1.7	36.2	31.3	19.9	81.6	–	80.6	63.0	21.8	86.2	–	40.7	–	23.6	53.1	47.9
DACS [174]	80.6	25.1	81.9	21.5	2.9	37.2	22.7	24.0	83.7	–	90.8	67.6	38.3	82.9	–	38.9	–	28.5	47.6	48.3
TACS [57]	91.5	60.4	82.5	21.8	1.5	31.7	31.6	28.0	84.7	–	89.0	66.7	35.8	81.0	–	42.8	–	28.5	45.9	51.5
BAPA [109]	91.7	53.8	83.9	22.4	0.8	34.9	30.5	42.8	86.6	–	88.2	66.0	34.1	86.6	–	51.3	–	29.4	50.5	53.3
CorDA [186]	**93.3**	**61.6**	85.3	19.6	5.1	37.8	36.6	42.8	84.9	–	90.4	69.7	41.8	85.6	–	38.4	–	32.6	53.9	55.0
ProDA [212]	87.8	45.7	84.6	37.1	0.6	44.0	54.6	37.0	88.1	–	84.4	74.2	24.3	88.2	–	51.1	–	40.5	45.6	55.5
UndoUDA [110]	82.5	37.2	81.1	23.8	0.0	45.7	57.2	47.6	87.7	–	85.8	74.1	28.6	88.4	–	66.0	–	47.0	55.3	56.7
EHTD [100]	93.0	69.8	84.0	36.6	9.1	39.7	42.2	43.8	**88.2**	–	88.1	68.3	29.0	85.5	–	54.1	–	37.1	56.3	57.8
CPSL [101]	87.2	43.9	85.5	33.6	0.3	47.7	57.4	37.2	87.8	–	88.5	79.0	32.0	90.6	–	49.4	–	50.8	59.8	57.9
DAFormer [75]	84.5	40.7	88.4	41.5	6.5	50.0	55.0	54.6	86.0	–	89.8	73.2	48.2	87.2	–	53.2	–	53.9	61.7	60.9
HRDA [76]	85.2	47.7	88.8	**49.5**	4.8	57.2	**65.7**	60.9	85.3	–	92.9	79.4	52.8	89.0	–	64.7	–	63.9	64.9	65.8
IR^{2}F-RMM (Ours)	90.4	54.9	**89.4**	48.0	**7.4**	**59.0**	65.5	**63.2**	87.8	–	**94.1**	**80.5**	**55.8**	**90.0**	–	**65.9**	–	**64.5**	**66.8**	**67.7**

Table 5.1: **Synthetic-to-Real**: GTA → Cityscapes, SYNTHIA → Cityscapes. Best results are denoted in bold.

Method	Road	SW	Build	Wall	Fence	Pole	TL	TS	Veg	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	MC	Bike	mIoU
Cityscapes → Dark Zurich																				
ADVENT [184]	85.8	37.9	55.5	27.7	14.5	23.1	14.0	21.1	32.1	8.7	2.0	39.9	16.6	64.0	13.8	0.0	58.8	28.5	20.7	29.7
AdaptSeg [175]	86.1	44.2	55.1	22.2	4.8	21.1	5.6	16.7	37.2	8.4	1.2	35.9	26.7	68.2	45.1	0.0	50.1	33.9	15.6	30.4
BDL [105]	85.3	41.1	61.9	32.7	17.4	20.6	11.4	21.3	29.4	8.9	1.1	37.4	22.1	63.2	28.2	0.0	47.7	39.4	15.7	30.8
DMAda [37]*	75.5	29.1	48.6	21.3	14.3	34.3	36.8	29.9	49.4	13.8	0.4	43.3	50.2	69.4	18.4	0.0	27.6	34.9	11.9	32.1
DACS [174]	83.1	49.1	67.4	33.2	16.6	42.9	20.7	35.6	31.7	5.1	6.5	41.7	18.2	68.8	76.4	0.0	61.6	27.7	10.7	36.7
GCMA [159]*	81.7	46.9	58.8	22.0	20.0	41.2	40.5	41.6	64.8	31.0	32.1	53.5	47.5	75.5	39.2	0.0	49.6	30.7	21.0	42.0
MGCDA [161]*	80.3	49.3	66.2	7.8	11.0	41.4	38.9	39.0	64.1	18.0	55.8	52.1	53.5	74.7	66.0	0.0	37.5	29.1	22.7	42.5
CDAda [197]*	90.5	60.6	67.9	37.0	19.3	42.9	36.4	35.3	66.9	24.4	79.8	45.4	42.9	70.8	51.7	0.0	29.7	27.7	26.2	45.0
DANNet [193]*	90.4	60.1	71.0	33.6	22.9	30.6	34.3	33.7	**70.5**	31.8	80.2	45.7	41.6	67.4	16.8	0.0	73.0	31.6	22.9	45.2
GLASS [94]*	91.6	63.1	71.2	34.7	**26.7**	41.4	39.7	38.4	68.6	34.8	**83.7**	41.3	40.8	69.6	21.5	0.0	63.5	32.1	19.4	46.4
DANIA [194]*	91.5	62.7	**73.9**	39.9	25.7	36.5	35.7	36.2	71.4	35.3	82.2	48.0	44.9	73.7	11.3	0.1	64.3	36.7	22.7	47.0
CCDistill [49]*	89.6	58.1	70.6	36.6	22.5	33.0	27.0	30.5	68.3	33.0	80.9	42.3	40.1	69.4	58.1	0.1	72.6	47.7	21.3	47.5
DAFormer [75]	93.5	65.5	73.3	39.4	19.2	53.3	44.1	44.0	59.5	34.5	66.6	53.4	52.7	82.1	52.7	9.5	89.3	50.5	38.5	53.8
HRDA [76]	90.4	56.3	72.0	39.5	19.5	57.8	**52.7**	43.1	59.3	29.1	70.5	60.0	**58.6**	84.0	**75.5**	11.2	90.5	51.6	40.9	55.9
IR^{2}F-RMM (Ours)	**94.7**	**75.1**	73.2	**44.4**	25.7	**60.6**	39.0	**47.4**	70.2	**41.6**	77.3	**62.4**	55.5	**86.4**	55.5	**20.0**	**92.0**	**55.3**	**42.8**	**58.9**
Cityscapes → ACDC-Night																				
DMAda [37]*	74.7	29.5	49.4	17.1	12.6	31.0	38.2	30.0	48.0	22.8	0.2	47.0	25.4	63.8	12.8	46.1	23.1	24.7	24.6	32.7
MGCDA [161]*	74.5	52.5	69.4	7.7	10.8	38.4	40.2	43.3	61.5	36.3	37.6	55.3	25.6	71.2	10.9	46.4	32.6	27.3	33.8	40.8
GCMA [159]*	78.6	45.9	58.5	17.7	18.6	37.5	43.6	43.5	58.7	39.2	22.5	57.9	29.9	72.1	21.5	56.3	41.8	35.7	35.4	42.9
DANNet [193]*	90.7	61.2	75.6	35.9	28.8	26.6	31.4	30.6	**70.8**	39.4	78.7	49.9	28.8	65.9	24.7	44.1	61.1	25.9	34.5	47.6
DANIA [194]*	91.0	60.9	**77.7**	40.3	**30.7**	34.3	37.9	34.5	70.0	37.2	79.6	45.7	32.6	66.4	11.1	37.0	60.7	32.6	37.9	48.3
GALSS [94]*	91.8	**65.0**	76.4	38.1	30.0	35.8	38.5	37.6	69.2	41.4	**79.8**	45.8	31.2	69.6	38.0	59.9	45.7	24.9	37.2	50.3
HRDA [76]	87.3	46.2	76.0	35.7	17.5	**52.0**	**50.3**	**53.6**	53.1	44.0	41.7	**64.8**	**40.9**	76.3	49.1	64.8	83.1	**36.0**	**51.5**	53.9
IR^{2}F-RMM (Ours)	**92.8**	64.8	74.5	**42.4**	15.0	51.7	36.7	52.4	66.6	**46.7**	62.7	64.1	36.3	**80.3**	**59.8**	**72.1**	**87.7**	32.0	50.5	**57.3**

Table 5.2: **Day-to-Night**: Cityscapes → Dark Zurich, Cityscapes → ACDC-Night. * indicates auxiliary daytime/twilight images corresponding to night images on the target domain are needed for training. But our IR^{2}F-RMM does not need. Best results are denoted in bold.

Figure 5.4: **Qualitative Comparisons for UDA Semantic Segmentation,** under Cityscapes → Dark Zurich. (a) shows the example of Cityscapes images. (b) includes the Dark Zurich images. (c) covers the day images corresponding to the night images in (b) for better visual references. Note that, the day images in (c) are only used for visualization references, but are not used for training and testing. (d) and (e) are the segmentation results for (b) from HRDA [76] and our method, respectively.

Method	PSPNet [217]	DANNet [193]	DANIA [194]	GLASS [94]	HRDA [76]	Ours
mIoU (%)	19.0	29.9	28.9	31.8	36.7	**38.5**

Table 5.3: **Quantitative Generalization Comparisons**, on NightCity+ dataset. The model is trained on the day-to-night benchmark, Cityscapes→Dark Zurich, and is tested on NightCity+ dataset.

other UDA methods on the generalization ability evaluation, 38.5% *vs.*, 36.7%, 31.8%, 28.9%, 29.9%, 19.0%. It proves that our model trained on Cityscapes→Dark Zurich generalizes well to other unseen nighttime datasets.

Rectification Values Prediction on Unseen Coordinates. In order to test the generalization ability of rectification values prediction method to the unseen coordinates, we conduct the experiments to predict the rectification values on the 2× image coordinates. Since only the image coordinates are utilized during the training, the 2× image coordinates are unseen for training and are only used for testing. For the discrete modeling method HRDA [76], the prediction on unseen 2× image coordinates is realized by first predicting on the original image coordinate

and then up-sampling (*e.g.*, bilinear sampling) to the 2× image coordinates. For our continuous modeling method, IR^{2}F can directly output the rectification values by inputting 2× image coordinates. As shown in Fig. 5.5, it is observed that our IR^{2}F-based continuous modeling method generalizes well to the unseen coordinates and preserves finer details compared to the discrete modeling method, especially the boundary parts (see Fig. 5.5).

(a) RGB (b) Discrete (c) Continuous (**Ous**)

Figure 5.5: **Rectification Values Prediction on Unseen Coordinates.** Rectification values are predicted on the 2× image coordinates during the testing stage, which are unseen during the training stage. (b) is realized through the bilinear sampling of the output of HRDA [76], which is the discrete modeling method. (c) is realized by directly predicting rectification values on the 2× image coordinates with our IR^{2}F, which is continuous modeling method. It is shown that our IR^{2}F-based continuous modeling method can generalize well to the unseen 2× image coordinates, preserving finer details especially the boundary parts (see red dashed box).

Different Backbones. Besides the experimental results in Table 5.1 and Table 5.2, we show more quantitative comparisons between our

Method	GTA → Cityscapes	SYNTHIA → Cityscapes
HRDA-ResNet	64.6	60.0
IR^{2}F-ResNet (Ours)	**65.4**	**61.4**

Table 5.4: **Comparisons to HRDA,** with ResNet-101 backbone. As the reference, the highest performance with ResNet-101 backbone, other than HRDA and our method, for GTA, SYNTHIA → Cityscapes are 62.7% in [23] and 57.9% in [101], respectively.

method and the existing SOTA UDA method HRDA in Table 5.4, with the ResNet101 [68] backbone, to further verify the advantage of modeling rectification function in a continuous manner. As reported in Table 5.4, by simply plugging our proposed learnable continuous rectification model, our method outperforms HRDA in the GTA, SYNTHIA → Cityscapes benchmarks. Moreover, as the reference, the highest performance with ResNet-101 backbone, other than HRDA and our method, for GTA, SYNTHIA → Cityscapes are 62.7% in [23] and 57.9% in [101], resp. It means both HRDA and our IR^{2}F, *learnable rectification function modeling methods*, outperform other heuristics-based rectification function modeling methods under the ResNet-101 backbone, and supporting the validity and rationality of modeling the rectification function in the learnable manner as done by our RMM in Sec. 5.3.2.

Insertion of IR^{2}F-based Continuous RMM into MRNet. Our proposed IR^{2}F-based continuous RMM is in principle a plug-in module, which can be inserted into different UDA frameworks. In order to prove its compatibility with other UDA frameworks, we insert our IR^{2}F-based continuous RMM into MRNet [220]. In MRNet, pseudo-labels are originally rectified by the uncertainty measurement, which is formulated as prediction variances between the primary and auxiliary classifiers. The inputs into the primary and auxiliary classifiers are different-level features. In Table 5.5, it is shown that our IR^{2}F-RMM improves MRNet by 2.0% and 1.8% under GTA, SYNTHIA → Cityscapes, resp.

Ablation Study. In order to prove the effectiveness of different components in our proposed IR^{2}F-based continuous RMM, we conduct a set of ablation studies under the synthetic-to-real benchmarks. In Table 5.6, we ablate different ways of estimating rectification values $\mathbf{r}_k$ in Eq. (1), under the HRDA [76] framework. In our proposed IR^{2}F, $\mathbf{r}_k$ is learned by the INR from the features of different mixture members.

Method	GTA → Cityscapes	SYNTHIA → Cityscapes
MRNet	50.3	47.9
IR^{2}F-RMM (Ours)	**52.3**	**49.7**

Table 5.5: **Combination with MRNet.** Our IR^{2}F-based continuous RMM is inserted into MRNet, to replace the original uncertainty based pseudo-labels rectification adopted by MRNet.

Method	GTA → Cityscapes	SYNTHIA → Cityscapes
HRDA	73.8	65.8
w/o. IR^{2}F *w.* AVE	71.0	61.9
w/o. IR^{2}F *w.* Conv	72.9	65.6
w/o. IR^{2}F *w.* IRE	73.3	66.3
IFA [77]	73.1	65.5
Ours	**74.4**	**67.7**

Table 5.6: **Ablation Study.** "AVE" means the average ensemble in Eq. (5.7). "Conv" means to replace the MLP structure of IR^{2}F with the convolutional neural networks. "IRE" means ensemble of the last-layer outputs instead of features from different mixture members with implicit neural representations, *i.e.*, $\mathcal{G}_{lr}(\mathbf{x}^t) = \tilde{\mathbf{y}}^t_{lr}, \mathcal{G}_{hr}(\mathbf{x}^t) = \tilde{\mathbf{y}}^t_{hr}$ in Eq.(5.7). "IFA" leverages the INR-based semantic segmentation decoder head, as done in [77].

Other ways to estimate $\mathbf{r}_k$ can be, 1) *AVE*: setting $\mathbf{r}_k = 1/K$, *i.e.*, average ensemble; 2) *Conv*: replacing IR^{2}F with 5 convolutional blocks without using the coordinate information; 3) *IRE*: taking the last-layer output (before softmax) instead of features from each mixture member as input to the IR^{2}F. Besides, in Table 5.6, we compare to another alternative, "IFA", which leverages the INR-based segmentation decoder head as done in [77]. It is shown that "IFA" does not bring obvious benefits to UDA compared to HRDA [76], 73.1%, 65.5% *vs.*, 73.8%, 65.8%, verifying the necessity and importance of rectifying incorrect pseudo-labels for UDA compared to a stronger decoder.

Comparisons to Heuristics-based/ Discrete Rectification Function Modeling. In order to showcase the advantage of our learnable and continuous rectification function modeling over the heuristics-based/ discrete one, we employ different heuristics-based/ discrete rectification function modeling methods under the HRDA framework as the baselines. As shown in Table 5.7, we compare our continuous IR^{2}F to different

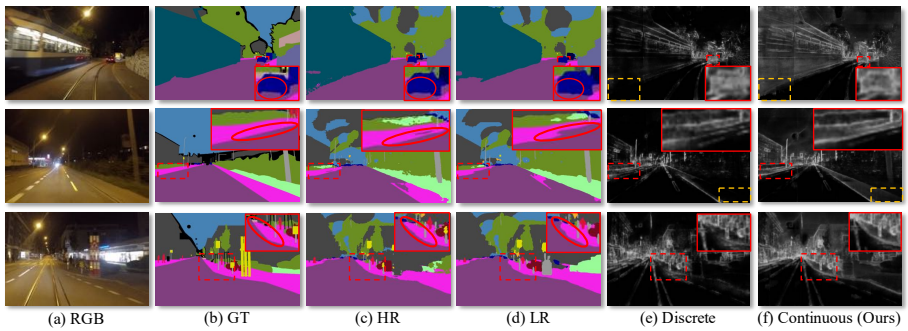


Figure 5.6: **Qualitative Comparisons between Discrete and Continuous Rectification Function Modeling.** (a) and (b) are the RGB inputs and corresponding ground truth semantic segmentation maps, respectively. (c) and (d) are the outputs of the HR, LR branches (see Sec. 5.3.4), *i.e.*, $\text{argmax}\,\tilde{\mathbf{y}}^t_{hr}$, $\text{argmax}\,\tilde{\mathbf{y}}^t_{lr}$ in Eq. (5.7), respectively. (e) and (f) are the estimated rectification values, *i.e.*, $\mathbf{r}$ in Eq. (5.7), by discrete modeling method (*i.e.*, additional decoder in HRDA [76]) and our continuous modeling method, IR^{2}F. In (e) and (f), the brighter the part is, the ensemble result in RMM relies more on HR branch result in (c). It is shown that our continuous modeling method can rectify some areas, which are ignored by the discrete modeling method (see orange dashed boxes), and other areas, where the the discrete modeling method is affected by the blurring effect and does not perform well (see red dashed boxes). The red dashed boxes are enlarged to red solid boxes for better visualization, especially the red circle parts. *Best viewed with zooming.*

rectification function modeling methods, including the *heuristics-based* method, 1) prediction variances [220] of $\tilde{\mathbf{y}}^t_{lr}$ and $\tilde{\mathbf{y}}^t_{hr}$ in Eq. (5.7), 2) Monte Carlo Dropout (MC-Dropout) [46], activating dropout function during inference to obtain different predictions for ensemble, and the *discrete* method, 3) an additional convolutional decoder is exploited to estimate rectification value as done in HRDA [76]. It is shown that our learnable and continuous rectification function modeling method, IR^{2}F-RMM, outperforms all heuristics-based and discrete modeling methods by a large margin. Furthermore, we provide the qualitative comparisons for the discrete and continuous rectification function modeling in Fig. 5.6. Benefiting from continuous modeling, the rectification

Method	GTA → Cityscapes	SYNTHIA → Cityscapes
AVE + Variance	73.7	65.1
AVE + MC-Dropout	71.8	63.9
Additional Conv Decoder	73.8	65.8
IR^{2}F	**74.4**	**67.7**

Table 5.7: **Comparisons to Heuristics-based/ Discrete Rectification Modeling.** "AVE" represents the average the ensemble. Heuristics-based modeling methods include, (1) "Variance": prediction variances are used to rectify pseudo-labels as done in [220], (2) "MC-Dropout": dropout is enabled during inference to get different predictions for ensemble [46], and discrete modeling method has (3) "Additional Decoder": an additional convolutional decoder is utilized to decode the rectification value as done in [76].

values of IR^{2}F are more accurate and insensitive to the blurring effect of down-/up-sampling operations in DNNs (see Sec. 5.3.3), especially at mask boundaries.

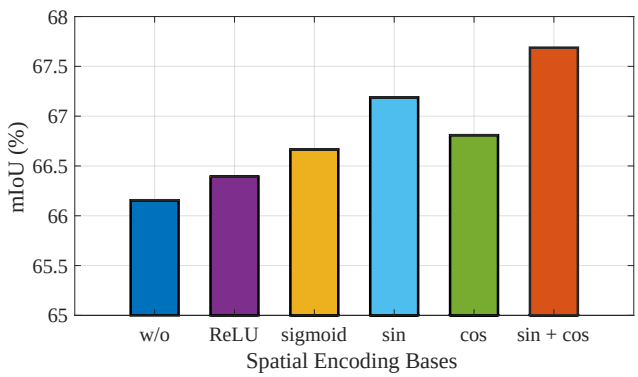


Figure 5.7: **Spatial Encoding Study.** Different spatial encoding bases are compared, and the combination of sin and cos reaches the highest performance.

Spatial Encoding Study. As analyzed in Sec. 5.3.3, the implicit neural representations are insensitive to the high-frequency signal in the image, *e.g.*, boundaries in the image. To overcome the shortcomings, we introduce the spatial encoding in Eq. (5.3), where the combination of sin and cos is adopted as encoding basis. To study the effectiveness of spatial encoding with both sin and cos, we compare to different encoding bases

in Fig. 5.7, including without spatial encoding, leakyReLU, sigmoid, pure sin and pure cos. It is observed that all the spatial encoding bases outperform the one without spatial encoding, proving the effectiveness of the spatial encoding. Among different spatial encoding bases, the combination of sin and cos reaches the highest performance, taken as the spatial encoding basis in IR^{2}F.

5.5 CONCLUSION

In this work, we presented continuous rectification-aware mixture model (RMM) based on implicit neural representations, which rectifies pseudo-labels for UDA in a learnable, continuous and end-to-end manner. As a principled and plug-in module, continuous RMM can be combined with different UDA frameworks, boosting the quality of pseudo-labels. Overall, our proposed continuous RMM achieves superior results compared to state-of-the-art, on synthetic-to-real and day-to-night UDA benchmarks.

6 CONCLUSIONS AND OUTLOOK

6.1 CONTRIBUTIONS

This dissertation proposed a series of novel and practical domain adaptation problems along with their corresponding methods. The primary goal of this dissertation is to overcome the limitations inherent in traditional domain adaptation, with a focus on advancing towards a more comprehensive and effective real-world scene understanding. To achieve this objective, this dissertation addressed two key questions: 1) *How can we make domain adaptation more practical?* (Chapter 2, 3, 4) 2) *How can we make domain adaptation more reliable?* (Chapter 5) In response to question 1), Chapter 2, 3, 4 relaxed the traditional single-source-single-target and compatible label assumptions from the perspectives of multi-source, multi-target and incompatible label space, respectively. In addressing question 2), Chapter 5 rectifies negative knowledge transfer within the domain adaptation process. Further details regarding specific contributions are outlined below.

In Chapter 2, we deviated from the single-source assumption of traditional domain adaptation, by proposing a new problem named as multi-source domain adaptation and label unification (mDALU). In mDALU problem, multiple source domains coexist with an unlabeled target domain, where each source domain only labels a subset of classes. The primary objective of mDALU is to develop a model that encompasses all classes within the target domain. In addressing this challenge, we proposed novel and effective two-phase framework. The initial phase encompasses a partially-supervised adaptation stage, followed by a fully-supervised adaptation stage. During the partially-supervised stage, we developed domain attention, uncertainty maximization, and attention-guided adversarial alignment modules to facilitate the partial transfer of knowledge from different source domains to the target domain. These modules prevented negative transfer resulting from the mismatched label space between different source domains. Following this, in the fully-supervised stage, we devised pseudo-label based supervision fusion module to further enhance the knowledge transfer within

the unified label space across all domains jointly. We demonstrated the effectiveness of our proposed method across a spectrum of tasks, encompassing 2D image semantic segmentation, 2D-3D cross-modal semantic segmentation, and extending to image classification. This chapter showcased the feasibility, validity, and effectiveness of extending single-source assumption of domain adaptation to multiple-source scenario.

In Chapter 3, we relaxed the single-target assumption of traditional domain adaptation, investigating the open compound domain adaptation (OCDA) problem. OCDA views the target domain as a compound of multiple unknown sub-domains. To address OCDA, we proposed a meta-learning based approach, MOCDA, which involves four key steps – cluster, split fuse and update. In the initial *cluster* step, style codes from target domain images are extracted and grouped into distinct clusters in the unsupervised manner. Subsequently, in the *split* step, different batch normalization layers are learned for different sub-domains of the target domain, based on the clustering results of previous step. Following this, in the *fuse* step, a meta-learner is proposed to learn to fuse sub-target domain-specific predictions, conditioned on the corresponding style codes. Meanwhile, the training of the meta-learner utilizes the model-agnostic meta-learning (MAML) algorithm, facilitating online *update* and thereby enhancing generalization ability. Extensive experiments on synthetic-to-real benchmarks validated the benefits of our proposed method for OCDA, outperforming other competing methods significantly. This chapter explored the practical multi-target scenario, further demonstrating enhanced generalization in complex applications. This marked a departure from the conventional assumption of a single-target condition in traditional domain adaptation.

In Chapter 4, we departed from the compatible label space assumption of traditional domain adaptation, newly proposing a problem – taxonomy adaptive cross-domain semantic segmentation (TACS). In TACS setting, a source domain class can be corresponding to multiple classes in the target domain, leading to open, coarse-to-fine and partially-overlapping taxonomy. Furthermore, we proposed the first approach to TACS, which addressed both the label-level and image-level domain gaps. At the label level, our proposed strategies included bilateral mixed sampling, stochastic label mapping, and pseudo-label-based relabeling modules to augment and align the target domain. On the image level, we presented the uncertainty-rectified contrastive

learning module, enabling the model to learn features that are both class-discriminative and domain-invariant, thereby remedying the domain gap. Across various benchmarks, our method demonstrated significant performance gains over the existing state-of-the-art, showcasing superior adaptability to target taxonomies. This chapter challenged the compatible label space assumption of traditional domain adaptation, accommodating inconsistent taxonomies between source and target domains and aligning with the dynamic requirements encountered in practical scenarios.

In Chapter 5, we proposed a novel approach based on implicit neural representations to enhance the quality of pseudo-labels in domain adaptation. This improvement contributed to the overall reliability and effectiveness of domain adaptation in practical applications. Pseudo-labels or self-training, commonly utilized in contemporary domain adaptation frameworks, are often inevitably noisy resulting from domain discrepancy. To address this issue, we proposed a continuous rectification-aware mixture model (RMM) based on implicit neural representations. The model was designed to rectify pseudo-labels for domain adaptation in a learnable, continuous, and end-to-end manner. Serving as a principled and plug-in module, the continuous RMM seamlessly integrates with different domain adaptation frameworks, thereby improving the quality of pseudo-labels. Extensive evaluations demonstrated the consistent superiority of our proposed continuous RMM over state-of-the-art methods across synthetic-to-real and day-to-night domain adaptation benchmarks. This chapter suggested a plug-in module to boost the reliability and effectiveness of pseudo-labels in domain adaptation, facilitating more dependable knowledge transfer.

6.2 CHALLENGES AND OUTLOOK

This dissertation extends traditional domain adaptation to multi-source, multi-target and taxonomy-adaptive scenarios and improves the reliability of knowledge transfer in domain adaptation, pushing towards real-world applications of domain adaptation for scene understanding. Despite notable progress being made, several challenges and limitations persist and necessitate further exploration to achieve a comprehensive cross-domain scene understanding. In this section, we point out some of these challenges and delve into potential avenues for future research.

Adaptation Efficiency. While numerous domain adaptation studies strive to enhance performance towards the oracle level (i.e., in-domain supervised learning performance), current approaches often demand substantial computational resources (e.g., GPUs) and sufficient data for training or fine-tuning. This poses challenges in terms of both computational efficiency and data efficiency, particularly in extreme scenarios. For instance, the model update during the adaptation process necessitates gradient backpropagation, which may be impractical in low-computing edge device scenarios, such as smartwatches. Furthermore, in certain domains such as medical image research, the scarcity of examples for rare diseases complicates adaptation efforts. Although some recent works [10, 169] have acknowledged this problem, adapting the model in scenarios with computational and data insufficiency remains a largely unexplored direction.

Automatic Data Factory. A practical challenge in domain adaptation is the acquisition of target domain data for adaptation. Manual collection is straightforward but often expensive, and in extreme scenarios like accidents, extreme weather, or volcanic eruptions, data collection becomes nearly impossible. This challenge underscores the need for an automatic data factory, where data can be synthesized based on requirements (e.g., text description), and corresponding labels can be generated accordingly. The synthesized data and labels become valuable resources for the domain adaptation process. While some recent works [169, 98, 215] have started addressing this strategy, there is a need for increased flexibility and enhancement in the synthesis methods.

Unified Multi-Modal Adaptation. In Chapter 2 of this dissertation, we partly tackle cross-modal domain adaptation, focusing on knowledge transfer between images and LiDAR points. In practical scenarios, various modalities exist, such as images, LiDAR, text, and audio. While individual works have explored domain adaptation within each modality, the challenge lies in developing a unified and extensible framework for multi-modal domain adaptation, an open problem that requires further exploration. The advantage of establishing such a unified framework is not only in addressing multiple problems within the same framework, but also in leveraging knowledge from different modalities to mutually enhance and improve model robustness across various scenarios. Notably, recent advancements in multi-modal foundational models [130, 42] demonstrate strong prior knowledge acquisition across

different modalities, presenting a promising avenue for designing a unified multi-modality adaptation framework.

BIBLIOGRAPHY

[1] Waqar Ahmed, Pietro Morerio, and Vittorio Murino. „Cleaning Noisy Labels by Negative Ensemble Learning for Source-Free Unsupervised Domain Adaptation.“ In: *WACV*. 2022 (cit. on pp. 88, 89).

[2] Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. „Augmented reality meets computer vision: Efficient data generation for urban driving scenes.“ In: *IJCV* 126.9 (2018), pp. 961–972 (cit. on pp. 47, 48).

[3] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando De Freitas. „Learning to learn by gradient descent by gradient descent.“ In: *NeurIPS*. 2016 (cit. on p. 37).

[4] Matan Atzmon and Yaron Lipman. „Sal: Sign agnostic learning of shapes from raw data.“ In: *CVPR*. 2020 (cit. on p. 86).

[5] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. „Metareg: Towards domain generalization using meta-regularization.“ In: *NeurIPS*. 2018 (cit. on pp. 35, 37).

[6] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. „Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields.“ In: *ICCV*. 2021 (cit. on p. 86).

[7] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. „Mixmatch: A holistic approach to semi-supervised learning.“ In: *NeurIPS* (2019) (cit. on p. 88).

[8] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. „Learning discriminative model prediction for tracking.“ In: *ICCV*. 2019 (cit. on p. 37).

[9] Christopher M Bishop. „Mixture density networks.“ In: (1994) (cit. on p. 89).

[10] Ondrej Bohdal, Da Li, Shell Xu Hu, and Timothy Hospedales. „Feed-Forward Latent Domain Adaptation." In: *WACV*. 2024 (cit. on p. 108).

[11] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. „Handling new target classes in semantic segmentation with domain adaptation." In: *arXiv preprint arXiv:2004.01130* (2020) (cit. on pp. 60–62).

[12] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. „nuscenes: A multimodal dataset for autonomous driving." In: *CVPR*. 2020 (cit. on p. 2).

[13] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. „nuScenes: A multimodal dataset for autonomous driving." In: *arXiv preprint arXiv:1903.11027* (2019) (cit. on pp. 29, 30).

[14] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. „COCO-Stuff: Thing and Stuff Classes in Context." In: *CVPR*. 2018 (cit. on p. 2).

[15] Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. „Partial Transfer Learning With Selective Adversarial Networks." In: *CVPR*. 2018 (cit. on p. 13).

[16] Zhangjie Cao, Lijia Ma, Mingsheng Long, and Jianmin Wang. „Partial Adversarial Domain Adaptation." In: *ECCV*. 2018 (cit. on pp. 12, 13, 59, 61).

[17] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. „Domain-Specific Batch Normalization for Unsupervised Domain Adaptation." In: *CVPR*. 2019 (cit. on p. 40).

[18] Chen-Hao Chao, Bo-Wun Cheng, and Chun-Yi Lee. „Rethinking ensemble-distillation for semantic segmentation based unsupervised domain adaption." In: *CVPR Workshops*. 2021 (cit. on p. 88).

[19] Hongruixuan Chen, Chen Wu, Yonghao Xu, and Bo Du. „Unsupervised domain adaptation for semantic segmentation via low-level edge information transfer." In: *arXiv preprint arXiv:2109.08912* (2021) (cit. on p. 90).

[20] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. „Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.4 (2017), pp. 834–848 (cit. on pp. 1, 26, 36, 47, 71).

[21] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. „Rethinking atrous convolution for semantic image segmentation." In: *arXiv preprint arXiv:1706.05587* (2017) (cit. on p. 36).

[22] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. „Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation." In: *ECCV*. 2018 (cit. on p. 36).

[23] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. „Deliberated Domain Bridging for Domain Adaptive Semantic Segmentation." In: *NeurIPS*. 2022 (cit. on pp. 96, 100).

[24] Minghao Chen, Hongyang Xue, and Deng Cai. „Domain adaptation for semantic segmentation with maximum squares loss." In: *CVPR*. 2019 (cit. on p. 17).

[25] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. „A simple framework for contrastive learning of visual representations." In: *ICML*. 2020 (cit. on p. 62).

[26] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. „Big self-supervised models are strong semi-supervised learners." In: *NeurIPS*. 2020 (cit. on p. 62).

[27] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. „Improved Baselines with Momentum Contrastive Learning." In: *arXiv preprint arXiv:2003.04297* (2020) (cit. on p. 62).

[28] Yinbo Chen, Sifei Liu, and Xiaolong Wang. „Learning continuous image representation with local implicit image function." In: *CVPR*. 2021 (cit. on pp. 87, 90).

[29] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. „Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach." In: *CVPR*. 2019 (cit. on pp. 36, 86, 87).

[30] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. „Domain Adaptive Faster R-CNN for Object Detection in the Wild.“ In: *CVPR*. 2018 (cit. on pp. 2, 10, 57).

[31] Yuhua Chen, Wen Li, and Luc Van Gool. „Road: Reality oriented adaptation for semantic segmentation of urban scenes.“ In: *CVPR*. 2018 (cit. on p. 36).

[32] Ziliang Chen, Jingyu Zhuang, Xiaodan Liang, and Liang Lin. „Blending-target domain adaptation by adversarial meta-adaptation networks.“ In: *CVPR*. 2019 (cit. on pp. 35–37).

[33] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. „Masked-attention mask transformer for universal image segmentation.“ In: *CVPR*. 2022 (cit. on p. 1).

[34] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. „The cityscapes dataset for semantic urban scene understanding.“ In: *CVPR*. 2016 (cit. on pp. 1, 3, 25, 30, 35, 47, 48, 58, 71, 73, 93, 94).

[35] Koby Crammer, Michael Kearns, and Jennifer Wortman. „Learning from Multiple Sources.“ In: *JMLR* 9.57 (2008), pp. 1757–1774 (cit. on p. 13).

[36] Dengxin Dai, Christos Sakaridis, Simon Hecker, and Luc Van Gool. „Curriculum model adaptation with synthetic and real data for semantic foggy scene understanding.“ In: *IJCV* 128.5 (2020), pp. 1182–1204 (cit. on p. 11).

[37] Dengxin Dai and Luc Van Gool. „Dark model adaptation: Semantic image segmentation from daytime to nighttime.“ In: *ITSC*. 2018 (cit. on p. 97).

[38] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. „Imagenet: A large-scale hierarchical image database.“ In: *CVPR*. 2009 (cit. on p. 1).

[39] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. „NightLab: A Dual-level Architecture with Hardness Detection for Segmentation at Night.“ In: *CVPR*. 2022 (cit. on p. 94).

[40] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. „An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In: *ICLR* (2021) (cit. on p. 1).

[41] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. „Domain generalization via model-agnostic learning of semantic features." In: *NeurIPS*. 2019 (cit. on pp. 36, 37).

[42] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. „PaLM-E: An Embodied Multimodal Language Model." In: *arXiv preprint arXiv:2303.03378*. 2023 (cit. on p. 108).

[43] Aysegul Dundar, Ming-Yu Liu, Ting-Chun Wang, John Zedlewski, and Jan Kautz. „Domain stylization: A strong, simple baseline for synthetic to real image domain adaptation." In: *arXiv preprint arXiv:1807.09384* (2018) (cit. on p. 86).

[44] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. „The pascal visual object classes (voc) challenge." In: *International Journal of Computer Vision* 88 (2010), pp. 303–338 (cit. on p. 1).

[45] Chelsea Finn, Pieter Abbeel, and Sergey Levine. „Model-agnostic meta-learning for fast adaptation of deep networks." In: *ICML*. 2017 (cit. on pp. 35, 37, 44).

[46] Yarin Gal and Zoubin Ghahramani. „Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In: *ICML*. 2016 (cit. on pp. 102, 103).

[47] Yaroslav Ganin and Victor Lempitsky. „Unsupervised domain adaptation by backpropagation." In: *ICML*. 2015 (cit. on pp. 11, 12, 22–24, 36, 60, 69).

[48] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario March, and Victor Lempitsky. „Domain-adversarial training of neural networks." In: *Journal of Machine Learning Research* 17.59 (2016), pp. 1–35 (cit. on pp. 1, 69).

[49] Huan Gao, Jichang Guo, Guoli Wang, and Qian Zhang. „Cross-Domain Correlation Distillation for Unsupervised Domain Adaptation in Nighttime Semantic Segmentation." In: *CVPR*. 2022 (cit. on p. 97).

[50] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. „A2d2: Audi autonomous driving dataset." In: *arXiv preprint arXiv:2004.06320* (2020) (cit. on pp. 2, 29, 30).

[51] Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. „Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation." In: *arXiv preprint arXiv:2012.07177* (2020) (cit. on p. 68).

[52] Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. „Unsupervised multi-target domain adaptation: An information theoretic approach." In: *TIP* 29 (2020), pp. 3993–4002 (cit. on p. 36).

[53] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. „Geodesic flow kernel for unsupervised domain adaptation." In: *CVPR*. 2012 (cit. on p. 36).

[54] Rui Gong, Yuhua Chen, Danda Pani Paudel, Yawei Li, Ajad Chhatkuli, Wen Li, Dengxin Dai, and Luc Van Gool. „Cluster, split, fuse, and update: Meta-learning for open compound domain adaptive semantic segmentation." In: *CVPR*. 2021 (cit. on pp. ix, 33).

[55] Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, Danda Pani Paudel, and Luc Van Gool. „Analogical image translation for fog generation." In: *AAAI Conference on Artificial Intelligence*. 2021 (cit. on p. ix).

[56] Rui Gong, Dengxin Dai, Yuhua Chen, Wen Li, and Luc Van Gool. „mDALU: Multi-source domain adaptation and label unification with partial datasets.“ In: *ICCV*. 2021 (cit. on pp. ix, 9).

[57] Rui Gong, Martin Danelljan, Dengxin Dai, Danda Pani Paudel, Ajad Chhatkuli, Fisher Yu, and Luc Van Gool. „TACS: Taxonomy adaptive cross-domain semantic segmentation.“ In: *ECCV*. 2022 (cit. on pp. ix, 57, 96).

[58] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. „Prompting diffusion representations for cross-domain semantic segmentation.“ In: *arXiv preprint arXiv:2307.02138* (2023) (cit. on p. x).

[59] Rui Gong, Wen Li, Yuhua Chen, Dengxin Dai, and Luc Van Gool. „Dlow: Domain flow and applications.“ In: *International Journal of Computer Vision* 129.10 (2021), pp. 2865–2888 (cit. on p. ix).

[60] Rui Gong, Wen Li, Yuhua Chen, and Luc Van Gool. „Dlow: Domain flow for adaptation and generalization.“ In: *CVPR*. 2019 (cit. on pp. 36, 83).

[61] Rui Gong, Qin Wang, Martin Danelljan, Dengxin Dai, and Luc Van Gool. „Continuous Pseudo-Label Rectified Domain Adaptive Semantic Segmentation With Implicit Neural Representations.“ In: *CVPR*. 2023 (cit. on pp. ix, 83).

[62] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. „3D Semantic Segmentation with Submanifold Sparse Convolutional Networks.“ In: *CVPR* (2018) (cit. on p. 30).

[63] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. „Bootstrap your own latent: A new approach to self-supervised learning.“ In: *arXiv preprint arXiv:2006.07733* (2020) (cit. on p. 62).

[64] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. „Implicit geometric regularization for learning shapes.“ In: *arXiv preprint arXiv:2002.10099* (2020) (cit. on p. 86).

[65] Edward Gunther, Rui Gong, and Luc Van Gool. „Style Adaptive Semantic Image Editing with Transformers.“ In: *ECCV Workshops*. 2022 (cit. on p. ix).

[66] David Ha, Andrew Dai, and Quoc V Le. „Hypernetworks.“ In: *arXiv preprint arXiv:1609.09106* (2016) (cit. on p. 37).

[67] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. „Momentum contrast for unsupervised visual representation learning.“ In: *CVPR*. 2020 (cit. on pp. 10, 62).

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. „Deep residual learning for image recognition.“ In: *CVPR*. 2016 (cit. on pp. 1, 26, 30, 71, 95, 100).

[69] Daniel Hernandez-Juarez, Lukas Schneider, Antonio Espinosa, David Vazquez, Antonio M. Lopez, Uwe Franke, Marc Pollefeys, and Juan Carlos Moure. „Slanted Stixels: Representing San Francisco's Steepest Streets.“ In: *BMVC*. 2017 (cit. on p. 47).

[70] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. „Learning to learn using gradient descent.“ In: *ICANN*. 2001 (cit. on p. 37).

[71] Judy Hoffman, Mehryar Mohri, and Ningshan Zhang. „Algorithms and Theory for Multiple-Source Adaptation.“ In: *NeurIPS*. 2018 (cit. on p. 13).

[72] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. „Cycada: Cycle-consistent adversarial domain adaptation.“ In: *ICML*. 2018 (cit. on pp. 2, 26, 36, 57, 83).

[73] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. „Fcns in the wild: Pixel-level adversarial and constraint-based adaptation.“ In: *arXiv preprint arXiv:1612.02649* (2016) (cit. on pp. 60, 86).

[74] Lukas Hoyer, Dengxin Dai, Yuhua Chen, Adrian Koring, Suman Saha, and Luc Van Gool. „Three Ways To Improve Semantic Segmentation With Self-Supervised Depth Estimation.“ In: *CVPR*. 2021 (cit. on p. 10).

[75] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. „Daformer: Improving network architectures and training strategies for domain-adaptive semantic segmentation.“ In: *CVPR*. 2022 (cit. on pp. 2, 3, 83, 93, 95–97).

[76] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. „HRDA: Context-aware high-resolution domain-adaptive semantic segmentation.“ In: *ECCV*. 2022 (cit. on pp. 3, 83, 84, 86, 89–91, 93–103).

[77] Hanzhe Hu, Yinbo Chen, Jiarui Xu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. „Learning implicit feature alignment function for semantic segmentation.“ In: *ECCV*. 2022 (cit. on pp. 84, 87, 90, 101).

[78] Xuecai Hu, Haoyuan Mu, Xiangyu Zhang, Zilei Wang, Tieniu Tan, and Jian Sun. „Meta-SR: A magnification-arbitrary network for super-resolution.“ In: *CVPR*. 2019 (cit. on p. 37).

[79] Zijian Hu, Zhengyu Yang, Xuefeng Hu, and Ram Nevatia. „Simple: similar pseudo label exploitation for semi-supervised classification.“ In: *CVPR*. 2021 (cit. on p. 88).

[80] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. „Correcting sample selection bias by unlabeled data.“ In: *NeurIPS*. 2007 (cit. on p. 43).

[81] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. „Multimodal unsupervised image-to-image translation.“ In: *ECCV*. 2018 (cit. on pp. 39, 40, 46).

[82] Sergey Ioffe and Christian Szegedy. „Batch normalization: Accelerating deep network training by reducing internal covariate shift.“ In: *ICML*. 2015 (cit. on p. 47).

[83] Maximilian Jaritz, Tuan-Hung Vu, Raoul de Charette, Emilie Wirbel, and Patrick Perez. „xMUDA: Cross-Modal Unsupervised Domain Adaptation for 3D Semantic Segmentation.“ In: *CVPR*. 2020 (cit. on pp. 12, 21, 29–31).

[84] Hu JIan, Hongya Tuo, Chao Wang, Lingfeng Qiao, Haowen Zhong, Yan Junchi, Zhongliang Jing, and Henry Leung. „Discriminative Partial Domain Adversarial Network.“ In: *ECCV*. 2020 (cit. on p. 13).

[85] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. „Local implicit grid representations for 3d scenes.“ In: *CVPR*. 2020 (cit. on p. 86).

[86] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C.V. Jawahar. „Universal Semi-Supervised Semantic Segmentation.“ In: *ICCV*. 2019 (cit. on p. 13).

[87] Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. „Contrastive adaptation network for unsupervised domain adaptation.“ In: *CVPR*. 2019 (cit. on p. 62).

[88] Diederik P Kingma and Jimmy Ba. „Adam: A method for stochastic optimization.“ In: *ICLR*. 2015 (cit. on pp. 21, 46, 47, 94).

[89] Abhishek Kumar, Prasanna Sattigeri, Kahini Wadhawan, Leonid Karlinsky, Rogerio Feris, Bill Freeman, and Gregory Wornell. „Co-regularized alignment for unsupervised domain adaptation.“ In: *NeurIPS*. 2018 (cit. on p. 88).

[90] Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R Venkatesh Babu. „Class-incremental domain adaptation.“ In: *ECCV*. 2020 (cit. on pp. 58, 60, 62).

[91] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. „Simple and scalable predictive uncertainty estimation using deep ensembles.“ In: *NeurIPS*. 2017 (cit. on p. 89).

[92] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. „MSeg: A composite dataset for multi-domain semantic segmentation.“ In: *CVPR*. 2020 (cit. on pp. 2, 13, 58).

[93] Yann LeCun, Corinna Cortes, and CJ Burges. „MNIST handwritten digit database.“ In: *ATT Labs [Online].* 2 (2010) (cit. on p. 22).

[94] Hongjae Lee, Changwoo Han, and Seung-Won Jung. „GPS-GLASS: Learning Nighttime Semantic Segmentation Using Daytime Video and GPS data.“ In: *arXiv preprint arXiv:2207.13297* (2022) (cit. on pp. 95, 97, 98).

[95] Da Li and Timothy Hospedales. „Online Meta-Learning for Multi-Source and Semi-Supervised Domain Adaptation.“ In: *arXiv preprint arXiv:2004.04398* (2020) (cit. on p. 37).

[96] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. „Deeper, broader and artier domain generalization.“ In: *ICCV*. 2017 (cit. on p. 36).

[97] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. „Learning to generalize: Meta-learning for domain generalization.“ In: *arXiv preprint arXiv:1710.03463* (2017) (cit. on pp. 35–37).

[98] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. „BigDatasetGAN: Synthesizing ImageNet with pixel-wise annotations.“ In: *CVPR*. 2022 (cit. on p. 108).

[99] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. „Domain generalization with adversarial feature learning.“ In: *CVPR*. 2018 (cit. on p. 36).

[100] Junjie Li, Zilei Wang, Yuan Gao, and Xiaoming Hu. „Exploring High-quality Target Domain Information for Unsupervised Domain Adaptive Semantic Segmentation.“ In: *ACM MM*. 2022 (cit. on p. 96).

[101] Ruihuang Li, Shuai Li, Chenhang He, Yabin Zhang, Xu Jia, and Lei Zhang. „Class-Balanced Pixel-Level Self-Labeling for Domain Adaptive Semantic Segmentation.“ In: *CVPR*. 2022 (cit. on pp. 96, 100).

[102] Wen Li, Zheng Xu, Dong Xu, Dengxin Dai, and Luc Van Gool. „Domain generalization and adaptation using low rank exemplar SVMs.“ In: *TPAMI* 40.5 (2017), pp. 1114–1127 (cit. on p. 36).

[103] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. „Deep domain generalization via conditional invariant adversarial networks.“ In: *ECCV*. 2018 (cit. on p. 36).

[104] Yawei Li, Shuhang Gu, Kai Zhang, Luc Van Gool, and Radu Timofte. „DHP: Differentiable Meta Pruning via HyperNetworks.“ In: *arXiv preprint arXiv:2003.13683* (2020) (cit. on p. 37).

[105] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. „Bidirectional learning for domain adaptation of semantic segmentation.“ In: *CVPR*. 2019 (cit. on pp. 83, 97).

[106] Qing Lian, Fengmao Lv, Lixin Duan, and Boqing Gong. „Constructing self-motivated pyramid curriculums for cross-domain semantic segmentation: A non-adversarial approach.“ In: *ICCV*. 2019 (cit. on p. 50).

[107] Christopher Liao, Theodoros Tsiligkaridis, and Brian Kulis. „Pick up the PACE: Fast and Simple Domain Adaptation via Ensemble Pseudo-Labeling.“ In: *arXiv preprint arXiv:2205.13508* (2022) (cit. on p. 88).

[108] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. „Microsoft coco: Common objects in context.“ In: *ECCV*. 2014 (cit. on p. 2).

[109] Yahao Liu, Jinhong Deng, Xinchen Gao, Wen Li, and Lixin Duan. „Bapa-net: Boundary adaptation and prototype alignment for cross-domain semantic segmentation.“ In: *ICCV*. 2021 (cit. on pp. 90, 96).

[110] Yahao Liu, Jinhong Deng, Jiale Tao, Tong Chu, Lixin Duan, and Wen Li. „Undoing the Damage of Label Shift for Cross-domain Semantic Segmentation.“ In: *CVPR*. 2022 (cit. on p. 96).

[111] Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. „Open Compound Domain Adaptation.“ In: *CVPR*. 2020 (cit. on pp. 33–37, 39, 44, 47–51, 57).

[112] Stuart Lloyd. „Least squares quantization in PCM.“ In: *TIT* 28.2 (1982), pp. 129–137 (cit. on p. 39).

[113] Jonathan Long, Evan Shelhamer, and Trevor Darrell. „Fully convolutional networks for semantic segmentation.“ In: *CVPR*. 2015 (cit. on pp. 1, 36).

[114] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. „Learning transferable features with deep adaptation networks.“ In: *ICML*. 2015 (cit. on pp. 11, 36, 60).

[115] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. „Deep transfer learning with joint adaptation networks.“ In: *ICML*. 2017 (cit. on p. 36).

[116] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. „Unsupervised domain adaptation with residual transfer networks.“ In: *NeurIPS*. 2016 (cit. on p. 1).

[117] Ilya Loshchilov and Frank Hutter. „Decoupled Weight Decay Regularization.“ In: *ICLR*. 2018 (cit. on p. 94).

[118] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. „Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation.“ In: *CVPR*. 2019 (cit. on pp. 86, 88).

[119] Laurens van der Maaten and Geoffrey Hinton. „Visualizing data using t-SNE.“ In: *JMLR* 9.Nov (2008), pp. 2579–2605 (cit. on p. 52).

[120] Will Maddern, Geoffrey Pascoe, Chris Linegar, and Paul Newman. „1 year, 1000 km: The Oxford RobotCar dataset.“ In: *IJRR* 36.1 (2017), pp. 3–15 (cit. on p. 35).

[121] Dmitrii Marin, Zijian He, Peter Vajda, Priyam Chatterjee, Sam Tsai, Fei Yang, and Yuri Boykov. „Efficient segmentation: Learning downsampling near semantic boundaries.“ In: *ICCV*. 2019 (cit. on p. 84).

[122] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. „Nerf in the wild: Neural radiance fields for unconstrained photo collections.“ In: *CVPR*. 2021 (cit. on p. 86).

[123] Ke Mei, Chuang Zhu, Jiaqi Zou, and Shanghang Zhang. „Instance Adaptive Self-Training for Unsupervised Domain Adaptation.“ In: *ECCV*. 2020 (cit. on pp. 61, 72, 74–77, 79, 81).

[124] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. „NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis.“ In: *ECCV*. 2020 (cit. on pp. 86, 90).

[125] Saeid Motiian, Quinn Jones, Seyed Mehdi Iranmanesh, and Gianfranco Doretto. „Few-shot adversarial domain adaptation.“ In: *NeurIPS*. 2017 (cit. on p. 61).

[126] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. „Image to image translation for domain adaptation.“ In: *CVPR*. 2018 (cit. on p. 86).

[127] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. „Reading digits in natural images with unsupervised feature learning.“ In: *NeurIPS workshops*. 2011 (cit. on p. 22).

[128] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. „The mapillary vistas dataset for semantic understanding of street scenes.“ In: *ICCV*. 2017 (cit. on pp. 2, 3, 58).

[129] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. „Classmix: Segmentation-based data augmentation for semi-supervised learning.“ In: *WACV*. 2021 (cit. on pp. 66, 71, 78).

[130] OpenAI. „GPT-4 Technical Report.“ In: *arXiv preprint arXiv: 2303.08774* (2023) (cit. on p. 108).

[131] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. „Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift.“ In: *NeurIPS*. 2019 (cit. on p. 89).

[132] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. „Domain adaptation via transfer component analysis.“ In: *TNN* 22.2 (2010), pp. 199–210 (cit. on p. 36).

[133] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. „Two at once: Enhancing learning and generalization capacities via ibn-net.“ In: *ECCV*. 2018 (cit. on p. 50).

[134] Pau Panareda Busto and Juergen Gall. „Open Set Domain Adaptation.“ In: *ICCV*. 2017 (cit. on pp. 12, 13, 58–61).

[135] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. „Deepsdf: Learning continuous signed distance functions for shape representation.“ In: *CVPR*. 2019 (cit. on p. 86).

[136] Kwanyong Park, Sanghyun Woo, Inkyu Shin, and In-So Kweon. „Discover, Hallucinate, and Adapt: Open Compound Domain Adaptation for Semantic Segmentation.“ In: *NeurIPS*. 2020 (cit. on pp. 36, 37).

[137] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. „Contrastive learning for unpaired image-to-image translation.“ In: *ECCV*. 2020 (cit. on p. 62).

[138] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. „Pytorch: An imperative style, high-performance deep learning library.“ In: *NeurIPS*. 2019 (cit. on pp. 47, 74, 95).

[139] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. „Convolutional occupancy networks." In: *ECCV*. 2020 (cit. on p. 86).

[140] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. „Moment matching for multi-source domain adaptation." In: *ICCV*. 2019 (cit. on pp. 12, 13, 21–24, 36).

[141] Matthew Pitropov, Danson Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. „Canadian Adverse Driving Conditions Dataset." In: *arXiv preprint arXiv:2001.10117* (2020) (cit. on p. 35).

[142] Fabio Pizzati, Raoul de Charette, Michela Zaccaria, and Pietro Cerri. „Domain bridge for unpaired image-to-image translation and unsupervised domain adaptation." In: *WACV*. 2020 (cit. on p. 86).

[143] Kun Qian and Zhou Yu. „Domain adaptive dialog generation via meta learning." In: *arXiv preprint arXiv:1906.03520* (2019) (cit. on p. 37).

[144] Fengchun Qiao, Long Zhao, and Xi Peng. „Learning to learn single domain generalization." In: *CVPR*. 2020 (cit. on p. 36).

[145] Aravind Rajeswaran, Chelsea Finn, Sham M Kakade, and Sergey Levine. „Meta-learning with implicit gradients." In: *NeurIPS*. 2019 (cit. on p. 37).

[146] Sayan Rakshit, Dipesh Tamboli, Pragati Shuddhodhan Meshram, Biplab Banerjee, Gemma Roig, and Subhasis Chaudhuri. „Multi-Source Open-Set Deep Adversarial Domain Adaptation." In: *ECCV*. 2020 (cit. on pp. 12, 13).

[147] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. „Efficient parametrization of multi-domain deep neural networks." In: *CVPR*. 2018 (cit. on p. 13).

[148] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. „Learning multiple visual domains with residual adapters." In: *NeurIPS*. 2017 (cit. on p. 13).

[149] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. „Playing for data: Ground truth from computer games." In: *ECCV*. 2016 (cit. on pp. 25, 47, 71, 73, 93).

[150] Tobias Ringwald and Rainer Stiefelhagen. „UBR ^{2} S: Uncertainty-Based Resampling and Reweighting Strategy for Unsupervised Domain Adaptation.“ In: *arXiv preprint arXiv:2110.11739* (2021) (cit. on p. 88).

[151] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. „U-net: Convolutional networks for biomedical image segmentation.“ In: *MICCAI*. 2015 (cit. on pp. 30, 36).

[152] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M. Lopez. „The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes.“ In: *CVPR*. 2016 (cit. on pp. 25, 71, 73, 93).

[153] Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. „Unsupervised domain adaptation using feature-whitening and consensus loss.“ In: *CVPR*. 2019 (cit. on p. 2).

[154] Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. „Uncertainty-guided source-free domain adaptation.“ In: *ECCV*. 2022 (cit. on p. 2).

[155] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. „Adapting visual category models to new domains.“ In: *ECCV*. 2010 (cit. on p. 36).

[156] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. „Maximum classifier discrepancy for unsupervised domain adaptation.“ In: *CVPR*. 2018 (cit. on p. 36).

[157] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. „Open Set Domain Adaptation by Backpropagation.“ In: *ECCV*. 2018 (cit. on p. 13).

[158] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. „Open Set Domain Adaptation by Backpropagation.“ In: *ECCV*. 2018 (cit. on pp. 60, 61).

[159] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. „Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation.“ In: *ICCV*. 2019 (cit. on p. 97).

[160] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. „ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding.“ In: *ICCV*. 2021 (cit. on pp. 1, 93, 94).

[161] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. „Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation.“ In: *TPAMI* (2020) (cit. on pp. 93, 94, 97).

[162] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. „Unsupervised domain adaptation for semantic segmentation with gans.“ In: *arXiv preprint arXiv: 1711.06969* (2017) (cit. on pp. 36, 47).

[163] Jürgen Schmidhuber. „Evolutionary principles in self-referential learning, or on learning how to learn.“ PhD thesis. Technische Universität München, 1987 (cit. on p. 37).

[164] Tiancheng Shen, Yuechen Zhang, Lu Qi, Jason Kuen, Xingyu Xie, Jianlong Wu, Zhe Lin, and Jiaya Jia. „High Quality Segmentation for Ultra High-resolution Images.“ In: *CVPR*. 2022 (cit. on p. 84).

[165] Karen Simonyan and Andrew Zisserman. „Very Deep Convolutional Networks for Large-Scale Image Recognition.“ In: *ICLR*. 2015 (cit. on p. 1).

[166] Karen Simonyan and Andrew Zisserman. „Very deep convolutional networks for large-scale image recognition.“ In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on p. 47).

[167] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. „Implicit neural representations with periodic activation functions.“ In: *NeurIPS*. 2020 (cit. on pp. 87, 90).

[168] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. „Scene representation networks: Continuous 3d-structure-aware neural scene representations.“ In: *NeurIPS*. 2019 (cit. on pp. 86, 90).

[169] Han Sun, Rui Gong, Konrad Schindler, and Luc Van Gool. „SF-FSDA: Source-Free Few-Shot Domain Adaptive Object Detection with Efficient Labeled Data Factory.“ In: *Conference on Lifelong Learning Agents (CoLLAs)*. 2023 (cit. on pp. x, 108).

[170] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson WH Lau. „Night-time scene parsing with a large real dataset." In: *TIP* 30 (2021), pp. 9085–9098 (cit. on p. 94).

[171] Antti Tarvainen and Harri Valpola. „Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results." In: *NeurIPS*. 2017 (cit. on p. 66).

[172] Takeshi Teshima, Issei Sato, and Masashi Sugiyama. „Few-shot domain adaptation by causal mechanism transfer." In: *ICML*. 2020 (cit. on p. 61).

[173] Antonio Torralba and Alexei A Efros. „Unbiased look at dataset bias." In: *CVPR*. 2011 (cit. on p. 36).

[174] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. „Dacs: Domain adaptation via cross-domain mixed sampling." In: *WACV*. 2021 (cit. on pp. 2, 3, 57, 61, 66, 68, 72, 74–77, 79, 83, 93, 94, 96, 97).

[175] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. „Learning to adapt structured output space for semantic segmentation." In: *CVPR*. 2018 (cit. on pp. 1, 2, 18, 21, 26–29, 41, 47–51, 53–55, 57, 60, 61, 69, 83, 86, 93, 97).

[176] Yi-Hsuan Tsai, Kihyuk Sohn, Samuel Schulter, and Manmohan Chandraker. „Domain adaptation for structured output via discriminative patch representations." In: *ICCV*. 2019 (cit. on p. 83).

[177] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. „Adversarial discriminative domain adaptation." In: *CVPR*. 2017 (cit. on pp. 11, 36).

[178] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. „Deep domain confusion: Maximizing for domain invariance." In: *arXiv preprint arXiv:1412.3474* (2014) (cit. on p. 11).

[179] Laurens Van der Maaten and Geoffrey Hinton. „Visualizing data using t-SNE." In: *JMLR* 9.11 (2008) (cit. on p. 78).

[180] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. „Scan: Learning to classify images without labels." In: *ECCV*. 2020 (cit. on p. 62).

[181] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. „IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments.“ In: *WACV*. 2019 (cit. on p. 2).

[182] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. „Deep hashing network for unsupervised domain adaptation.“ In: *CVPR*. 2017 (cit. on p. 36).

[183] Riccardo Volpi, Hongseok Namkoong, Ozan Sener, John C Duchi, Vittorio Murino, and Silvio Savarese. „Generalizing to unseen domains via adversarial data augmentation.“ In: *NeurIPS*. 2018 (cit. on p. 36).

[184] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. „ADVENT: Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation.“ In: *CVPR*. 2019 (cit. on pp. 2, 11, 26–29, 31, 36, 44, 53, 54, 57, 61, 72, 75–77, 79, 83, 93, 97).

[185] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. „Dada: Depth-aware domain adaptation in semantic segmentation.“ In: *ICCV*. 2019 (cit. on p. 86).

[186] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. „Domain adaptive semantic segmentation with self-supervised depth estimation.“ In: *ICCV*. 2021 (cit. on p. 96).

[187] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. „Continual Test-Time Domain Adaptation.“ In: *CVPR*. 2022 (cit. on p. 88).

[188] Wenguan Wang, Tianfei Zhou, Fisher Yu, Jifeng Dai, Ender Konukoglu, and Luc Van Gool. „Exploring Cross-Image Pixel Contrast for Semantic Segmentation.“ In: *arXiv preprint arXiv: 2101.11939* (2021) (cit. on pp. 36, 62, 69).

[189] Xudong Wang, Zhaowei Cai, Dashan Gao, and Nuno Vasconcelos. „Towards universal object detection by domain attention.“ In: *CVPR*. 2019 (cit. on p. 13).

[190] Yuxi Wang, Junran Peng, and ZhaoXiang Zhang. „Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation.“ In: *ICCV*. 2021 (cit. on pp. 84, 86, 88).

[191] Zhonghao Wang, Yunchao Wei, Rogerio Feris, Jinjun Xiong, Wen-Mei Hwu, Thomas S Huang, and Honghui Shi. „Alleviating semantic-level shift: A semi-supervised domain adaptation method for semantic segmentation.“ In: *CVPR Workshops*. 2020 (cit. on pp. 76, 77).

[192] Magnus Wrenninge and Jonas Unger. „Synscapes: A photorealistic synthetic dataset for street scene parsing.“ In: *arXiv preprint arXiv:1810.08705* (2018) (cit. on pp. 71, 73).

[193] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. „Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation.“ In: *CVPR*. 2021 (cit. on pp. 97, 98).

[194] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. „A One-Stage Domain Adaptation Network with Image Alignment for Unsupervised Nighttime Semantic Segmentation.“ In: *TPAMI* (2021) (cit. on pp. 97, 98).

[195] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Zhenguo Li, and Ping Luo. „DetCo: Unsupervised Contrastive Learning for Object Detection.“ In: *arXiv preprint arXiv:2102.04803* (2021) (cit. on p. 62).

[196] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. „SegFormer: Simple and efficient design for semantic segmentation with transformers.“ In: *NeurIPS*. 2021 (cit. on pp. 1, 94).

[197] Qi Xu, Yinan Ma, Jing Wu, Chengnian Long, and Xiaolin Huang. „Cdada: A curriculum domain adaptation for nighttime semantic segmentation.“ In: *ICCV*. 2021 (cit. on p. 97).

[198] Ruijia Xu, Ziliang Chen, Wangmeng Zuo, Junjie Yan, and Liang Lin. „Deep cocktail network: Multi-source unsupervised domain adaptation with category shift.“ In: *CVPR*. 2018 (cit. on pp. 12, 13, 22–24).

[199] Xingqian Xu, Zhangyang Wang, and Humphrey Shi. „Ultrasr: Spatial encoding is a missing key for implicit image function-based arbitrary-scale super-resolution.“ In: *arXiv preprint arXiv:2103.12716* (2021) (cit. on pp. 87, 90).

[200] Yanchao Yang and Stefano Soatto. „Fda: Fourier domain adaptation for semantic segmentation.“ In: *CVPR*. 2020 (cit. on pp. 72, 74–77, 79, 83).

[201] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. „Volume rendering of neural implicit surfaces.“ In: *NeurIPS*. 2021 (cit. on p. 86).

[202] Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. „Universal domain adaptation.“ In: *CVPR*. 2019 (cit. on pp. 58–61).

[203] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. „BDD100K: A diverse driving dataset for heterogeneous multitask learning.“ In: *CVPR*. 2020 (cit. on p. 47).

[204] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. „MonoSDF: Exploring Monocular Geometric Cues for Neural Implicit Surface Reconstruction.“ In: *NeurIPS*. 2022 (cit. on p. 86).

[205] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. „Cutmix: Regularization strategy to train strong classifiers with localizable features.“ In: *ICCV*. 2019 (cit. on p. 68).

[206] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. „Central moment discrepancy (cmd) for domain-invariant representation learning.“ In: *ICLR*. 2017 (cit. on p. 11).

[207] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. „WildDash-Creating Hazard-Aware Benchmarks.“ In: *ECCV*. 2018 (cit. on pp. 47, 48).

[208] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. „mixup: Beyond empirical risk minimization.“ In: *ICLR*. 2018 (cit. on p. 68).

[209] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. „Generalizable Semantic Segmentation via Model-agnostic Learning and Target-specific Normalization.“ In: *arXiv preprint arXiv:2003.12296* (2020) (cit. on p. 37).

[210] Jing Zhang, Zewei Ding, Wanqing Li, and Philip Ogunbona. „Importance weighted adversarial nets for partial domain adaptation.“ In: *CVPR*. 2018 (cit. on p. 13).

[211] Junyi Zhang, Ziliang Chen, Junying Huang, Liang Lin, and Dongyu Zhang. „Few-shot structured domain adaptation for virtual-to-real scene parsing.“ In: *ICCV Workshops*. 2019 (cit. on pp. 61, 71, 78).

[212] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. „Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation.“ In: *CVPR*. 2021 (cit. on pp. 3, 83, 84, 86, 88, 93, 96).

[213] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. „Category anchor-guided unsupervised domain adaptation for semantic segmentation.“ In: *NeurIPS*. 2019 (cit. on p. 36).

[214] Yang Zhang, Philip David, and Boqing Gong. „Curriculum Domain Adaptation for Semantic Segmentation of Urban Scenes.“ In: *ICCV*. 2017 (cit. on pp. 36, 60).

[215] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. „Datasetgan: Efficient labeled data factory with minimal human effort.“ In: *CVPR*. 2021 (cit. on p. 108).

[216] Han Zhao, Shanghang Zhang, Guanhang Wu, José M. F. Moura, Joao P Costeira, and Geoffrey J Gordon. „Adversarial Multiple Source Domain Adaptation.“ In: *NeurIPS*. 2018 (cit. on p. 13).

[217] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. „Pyramid scene parsing network.“ In: *CVPR*. 2017 (cit. on p. 98).

[218] Sicheng Zhao, Bo Li, Xiangyu Yue, Yang Gu, Pengfei Xu, Runbo Hu, Hua Chai, and Kurt Keutzer. „Multi-source domain adaptation for semantic segmentation.“ In: *NeurIPS*. 2019 (cit. on pp. 12, 13, 27, 28, 36).

[219] Xiangyun Zhao, Samuel Schulter, Gaurav Sharma, Yi-Hsuan Tsai, Manmohan Chandraker, and Ying Wu. „Object Detection with a Unified Label Space from Multiple Datasets.“ In: *ECCV*. 2020 (cit. on pp. 13, 22–24).

[220] Zhedong Zheng and Yi Yang. „Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation.“ In: *IJCV* 129.4 (2021), pp. 1106–1120 (cit. on pp. 84, 86, 88, 89, 92, 93, 95, 96, 100, 102, 103).

[221] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. „In-place scene labelling and understanding with implicit scene representation.“ In: *ICCV*. 2021 (cit. on p. 86).

[222] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. „Semantic understanding of scenes through the ade20k dataset.“ In: *International Journal of Computer Vision* 127 (2019), pp. 302–321 (cit. on p. 1).

[223] Brady Zhou, Nimit Kalra, and Philipp Krähenbühl. „Domain Adaptation Through Task Distillation.“ In: *ECCV*. 2020 (cit. on p. 53).

[224] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. „Domain adaptive ensemble learning.“ In: *TIP* 30 (2021), pp. 8008–8018 (cit. on p. 88).

[225] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. „Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networkss.“ In: *ICCV*. 2017 (cit. on p. 26).

[226] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. „Confidence regularized self-training.“ In: *ICCV*. 2019 (cit. on pp. 3, 83, 84, 86).

[227] Yang Zou, Zhiding Yu, BVK Vijaya Kumar, and Jinsong Wang. „Unsupervised domain adaptation for semantic segmentation via class-balanced self-training.“ In: *ECCV*. 2018 (cit. on pp. 1, 3, 36, 50, 60, 83, 84, 86, 93, 96).

COLOPHON

This document was typeset in LaTeX using the typographical look-and-feel `classicthesis`. Most of the graphics in this thesis are generated using `pgfplots` and `pgf/tikz`. The bibliography is typeset using `biblatex`.