Addressing Domain Shifts for Computer Vision Applications via Language

by

Chang Liu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Chang Liu 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following publications form the basis of this thesis. I was the primary author on both papers, responsible for designing and implementing the core methods, leading the overall project direction, conducting the experiments, performing the analysis, and writing the majority of the manuscripts.

Chapter 3 and Chapter 4 of this thesis include content from the workshop paper: **Chang Liu**, Saad Hossain, C. Thomas, Henry Lai, Ravi Vemulapalli, Sirisha Rambhatla, and Alexander Wong; "Language-guided Domain Adaptive Semantic Segmentation" [46]; presented at the Adaptive Foundation Models Workshop in the Conference on Neural Information Processing Systems (NeurIPS) 2024.

In this work, I proposed the main method, developed the codebase, performed the experiments, conducted the analysis, and wrote the paper. Saad Hossain contributed by running baseline experiments for PODA [13] and ULDA [89], assisting in formatting the manuscript, and generating t-SNE visualizations. C. Thomas, Henry Lai, Ravi Vemulapalli, Sirisha Rambhatla, and Alexander Wong participated in discussions and provided feedback throughout the project.

Chapters 2 to 4 include content from a paper currently under review at the International Conference on Computer Vision (ICCV) 2025: **Chang Liu**, Saad Hossain, C. Thomas, Henry Lai, Ravi Vemulapalli, Sirisha Rambhatla, and Alexander Wong; "Building Context-Awareness via Language for Domain Adaptive Semantic Segmentation" [45].

For this paper, I designed the language alignment framework and incorporated the captioning model LLaVA [47], developed the full pipeline, and authored the majority of the manuscript. Saad Hossain conducted part of the experiments on the DarkZurich dataset [65], and Bavesh Balaji ran part of the experiments on the ACDC dataset [67], including ablations on OpenCLIP [27]. Bavesh Balaji also provided writing feedback and contributed to select sections of the paper. Ravi Vemulapalli helped shape the early formulation of the captioning component. C. Thomas, Henry Lai, Ravi Vemulapalli, Sirisha Rambhatla, and Alexander Wong offered feedback, reviewed drafts, and contributed to manuscript editing.

Abstract

Semantic segmentation is used in safety-critical applications such as autonomous driving and cancer diagnosis [41, 24, 5], where accurately identifying small and rare objects is essential. However, pixel-level annotations are expensive and time-consuming, and distribution shifts (e.g. daytime to snowy weather in self-driving, color variations between tumor scans across hospitals) between datasets further degrade model generalization capabilities [79, 21, 23]. Unsupervised domain adaptation for semantic segmentation (DASS) addresses this challenge by training models on labeled source distributions and adapting them to unlabeled target domains [95, 22, 21, 82, 79].

Existing DASS methods rely on either vision-only approaches or language-based techniques. Vision-only frameworks, such as masking [23] and utilizing multi-resolution crops [22], implicitly learn spatial relationships between different image patches but often suffer from noisy pseudo-labels biased toward the source domain [53]. To mitigate noisy predictions, language-based DASS methods leverage generalized representations from large-scale language pre-training. However, those approaches use generic class-level [53] prompts (e.g., "a photo of a {class}") and fail to capture complex spatial relationships between objects, which are key for dense prediction tasks like semantic segmentation.

To address these limitations, we propose LangDA, a language-guided DASS framework that enhances spatial context-awareness by leveraging vision-language models (VLMs). LangDA generates scene-level descriptions (e.g., "a pedestrian is on the sidewalk, and the street is lined with buildings") to encode object relationships. At an image-level, LangDA aligns an image's feature representation with the corresponding scene-level text embedding, improving the model's ability to generalize across domains. LangDA eliminates the need for cumbersome manual prompt tuning and expensive human feedback [15, 26, 89, 13], ensuring consistency and reproducibility.

LangDA achieves state-of-the-art performance on three DASS benchmarks: Synthia $\rightarrow$ Cityscapes, Cityscapes $\rightarrow$ ACDC, and Cityscapes $\rightarrow$ DarkZurich, surpassing existing methods [23, 53] by **2.6%**, **1.4%**, and **3.9%**, respectively. Ablation studies confirm the effectiveness of context-aware image-level alignment over pixel-level alignment. These results demonstrate LangDA's capability to leverage spatial relationships encoded in language to accurately segment objects under domain shift.

Acknowledgements

First, I would like to express my sincere gratitude to my supervisors, Professor Sirisha Rambhatla and Professor Alexander Wong, for providing me with guidance and support throughout my Master's journey. I would especially like to thank Prof. Sirisha for not only supporting me and providing feedback on my research, but also encouraging me to speak up and become a more confident individual.

I would like to thank Professor. Yuhao Chen from the Vision and Image Processing (VIP) lab, for always supporting me and believing in me, and for providing feedback on my writing that strengthened my paper submission.

My appreciation also goes to my thesis committee members, Professor. John Zelek and Professor. Elliot Creager, for taking the time from their busy schedule and providing me with valuable feedback on my thesis.

I would like to express my gratitude to my collaborators– Eddie, Bavesh, Saad, and Daniel– for their contributions to the various projects we've worked on together. I would also like to say thank you to Fernando, Eddie, and Ken, as well as the rest of the members in the Vision and Image Processing (VIP), for their effort and supporting in helping organize the 10th annual Conference on Vision and Intelligent Systems. Their hard work made the conference a success.

Special thank you to my friends, Ksusha, Estelle, Jun, Valerie, and Sophia, for always being there for me and offering their support, both academically and personally.

I am deeply grateful to my parents for their unconditional love and support, regardless of what I decide to do or where I am in life. And to my little sister, Lily, for reminding me what it means to be an older sister, and for the interesting things she brings into my life.

Lastly, I am grateful to C, Henry, Ravi, and Pablo from Apple for their funding support, which made this project possible. I would like to especially thank C for their continued mentorship and for providing insightful feedback on my internal presentations.

Dedication

This thesis is dedicated to my mom, dad, and little sister, Lily.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Dedication vi

List of Figures x

List of Tables xii

List of Abbreviations xiii

1 Introduction 1

1.1 Semantic Segmentation 1

1.2 Domain Shift 3

1.3 Domain Adaptive Semantic Segmentation 4

1.4 Language-guided Domain Adaptive Semantic Segmentation 6

1.5 Thesis Contribution 7

1.6 Overview 9

2 Background **10**
2.1 Vision-only Unsupervised Domain Adaptation . . . 10
2.2 Unsupervised Domain Adaptation with Self-Training . . . 12
2.3 Zero-shot Domain Adaptation with Language . . . 12
2.4 Domain Adaptation with Language . . . 12
2.5 Summary . . . 13

3 Language-guided Domain Adaptive Semantic Segmentation **15**
3.1 Methodology . . . 15
3.1.1 Problem Formulation . . . 15
3.1.2 Overview . . . 15
3.1.3 Source-target Distillation for Visual Domain . . . 17
3.1.4 Context-Aware Caption Generator . . . 19
3.1.5 Image-level Consistency Alignment . . . 20
3.2 Explanation and Analysis on Caption Generation . . . 22
3.2.1 Prompt Generation Process . . . 22
3.2.2 Analysis on Quality of Image-level Descriptors . . . 23
3.3 Experiments . . . 23
3.3.1 Datasets . . . 23
3.3.2 Implementation details . . . 23
3.3.3 Metrics . . . 24
3.3.4 Summary . . . 25

4 Results **26**
4.1 Main Results . . . 26
4.1.1 Quantitative Results on Synthia → CS. . . . 26
4.1.2 Per-class IoU on Synthia → Cityscapes . . . 26
4.1.3 Per-class IoU comparison on other two settings . . . 29

4.2 Qualitative Results . . . 32
4.2.1 Qualitative Results: Cityscapes . . . 32
4.2.2 Qualitative Results: ACDC and DarkZurich . . . 32
4.2.3 Qualitative Results: Captions . . . 35
4.2.4 t-SNE Visualization . . . 35
4.2.5 Confusion Matrix . . . 37
4.3 Ablation Studies . . . 38
4.3.1 Effect of Prompting and Alignment Strategies. . . . 38
4.3.2 Comparison of Source and Target Captions. . . . 38
4.3.3 Ablations on CLIP Text Encoders . . . 39
4.3.4 Hyperparameter Sensitivity . . . 39
4.3.5 Ablation on Caption Refinement Module. . . . 40
4.4 Limitations . . . 40
4.5 Summary . . . 42

5 Conclusion **44**
5.1 Summary of Contributions . . . 44
5.2 Limitations . . . 45
5.3 Future Works . . . 46

References **48**

List of Figures

1.1 Fundamental Modern Computer Vision tasks. 1
1.2 Manual Annotation for Semantic Segmentation: Expensive, Inefficient, and Time-consuming 2
1.3 Typical domain shifts in autonomous driving. 3
1.4 Performance degradation under domain shift. 4
1.5 Progress of DASS over time. 5
1.6 A Comparison: Pixel-level Alignment with Generic Prompt (Existing works) & Image-level Alignment with Scene-aware Captions (Our proposed method) 6
1.7 Misclassifications of Visually Similar Classes. 7
1.8 LangDA: New state-of-the-art in DASS. 8

2.1 Unsupervised Domain Adaptation Methods: An Overview. 11

3.1 LangDA Architecture. 16
3.2 VLM Caption Generation Module and LLM Caption Refinement Module . 17
3.3 Intersection over Union, Illustrated. 18
3.4 Examples: What is a good IoU? 19
3.5 Token Length After Refinement for Synthia and Cityscapes 21
3.6 Token Length Distribution on Synthia and Cityscapes 22
3.7 Domain adaptation dataset examples 24

4.1 Qualitative Results: Synthia→Cityscapes 27

4.2 Qualitative Results: CityScapes→ACDC 30
4.3 Qualitative Results: t-SNE 32
4.4 Qualitative Results: CityScapes→DarkZurich. 33
4.5 Confusion Matrix 37
4.6 Short context length of CLIP 41
4.7 Long-tailed class distribution in DarkZurich [65] 43

List of Tables

4.1 Comparison with state-of-the-art methods in UDA and Zero-shot DA. . . . 28

4.2 Performance under different weather conditions on Clear-to-Adverse-Weather benchmark: Cityscapes→ACDC 29

4.3 Per-class performance on Synthetic-to-Real adaptation benchmark: Synthia to Cityscapes 31

4.4 Per-class performance on Day-to-Night: Cityscapes→DarkZurich 31

4.5 Per-class performance on Clear-to-Adverse-Weather: Cityscapes→ACDC . 31

4.6 Qualitative Results: Captions Generated by Context-Aware Generator for Synthia [63] 34

4.7 Qualitative Results: Captions Generated by Context-Aware Generator for Cityscapes [11] 36

4.8 Effect of prompting and aligning techniques 38

4.9 Ablation: contextual caption applied on source, target, and source + target 39

4.10 Ablation: CLIP-based text encoders. 40

4.11 Ablation: weight of λ_p 40

4.12 Ablation on Language Refinement on Synthia → Cityscapes. 42

List of Abbreviations

ACDC Adverse Conditions Dataset with Correspondences 7, 9

AdamW Adaptive Moment Estimation with decoupled weight decay 24

CLIP Contrastive Language-Image Pre-training 6, 12, 13, 20

DASS Domain Adaptive Semantic Segmentation 4, 7, 8, 12, 13, 15, 20, 21, 23, 40, 46

EMA Exponential Moving Average 16, 21, 24

GPU Graphics Processing Unit 24

IoU Intersection over Union 24

LangDA Language-guided Domain Adaptive Semantic Segmentation 6–9, 14, 25, 40–42

LLaVA Large Language and Vision Assistant 19

LLM Large Language Model 7, 14, 20

mIoU mean Intersection over Union 24, 25

RTX Ray Tracing Texel eXtreme 24

t-SNE t-distributed Stochastic Neighbor Embedding 35

UDA Unsupervised Domain Adaptation 4, 5, 10–13, 21, 24

ViT-B/16 Vision Transformer Base with patch size 16 × 16 24

VLM Vision Language Model 7, 14, 19

Chapter 1

Introduction

Machine learning (ML) has driven technological advancements across diverse fields, including medical diagnostics [41, 24, 5], robotics [33, 48], and autonomous driving [81, 63, 39]. Its core aim is to develop models that extract generalizable patterns from training data, enabling accurate predictions on previously unseen data during testing and deployment.

1.1 Semantic Segmentation

Computer vision is the subfield of machine learning responsible for the perception of the world. Often cited as fundamental tasks [14, 75, 50], computer vision heavily focuses on classification, object detection, and semantic segmentation [8], as seen in Figure 1.1.

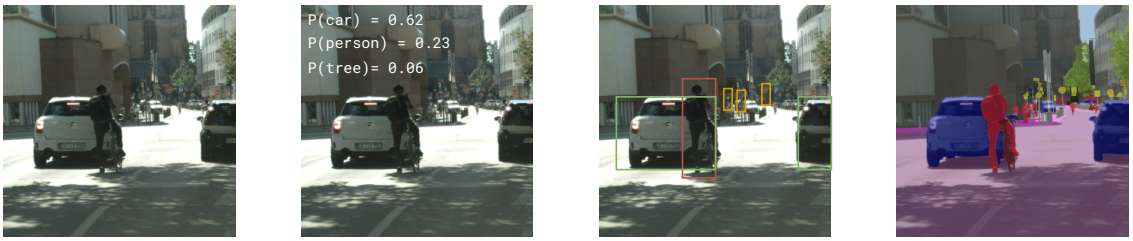


Figure 1.1: **Fundamental Modern Computer Vision tasks.** Classification, Detection, and Semantic Segmentation, on Cityscapes [11] dataset

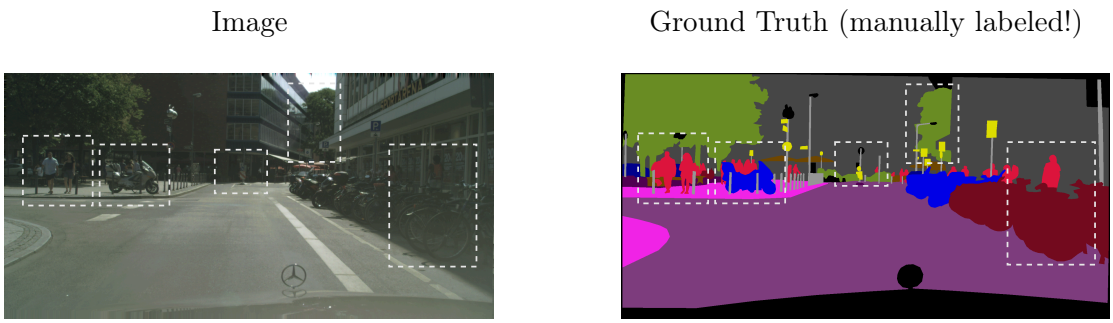


Figure 1.2: **Manual Annotation for Semantic Segmentation**. The white boxes highlight the detailed, dense polygonal boundaries that a human annotator must trace around every object—a costly, inefficient, and time-consuming process. The image is from Cityscapes [11] dataset.

Classification assigns a single label to an entire image, identifying the dominant object or scene. Object detection extends this by localizing and classifying multiple objects within an image using bounding boxes. Among these tasks, semantic segmentation is particularly challenging–it involves partitioning an image into distinct regions and assigning a semantic label to each pixel, producing a dense, pixel-wise annotation of the scene.

Because semantic segmentation enables a fine-grained and detailed understanding of the scene, it is essential in safety-critical applications such as cancer diagnosis [41, 24, 5] and autonomous driving [81, 63, 39], where accurately segmenting small and rare objects is crucial. However, pixel-level annotations required for training segmentation models are expensive and extremely time-consuming, as demonstrated in Figure 1.2. This challenge is further exacerbated by domain shifts–differences between the training (source, e.g., daytime for self-driving) and deployment (target, e.g., snowy weather in winter) environments–which degrade model performance when new data distributions emerge [79, 21, 23]. Collecting sufficient annotated data to cover all potential domain shifts is often impractical and costly. To address this, unsupervised domain adaptation for semantic segmentation (DASS) trains segmentation models on labeled source domains and adapts them to unlabeled target domains– usually based on the student-teacher architecture as illustrated in Figure 2.1 [95, 22, 21, 82, 79].

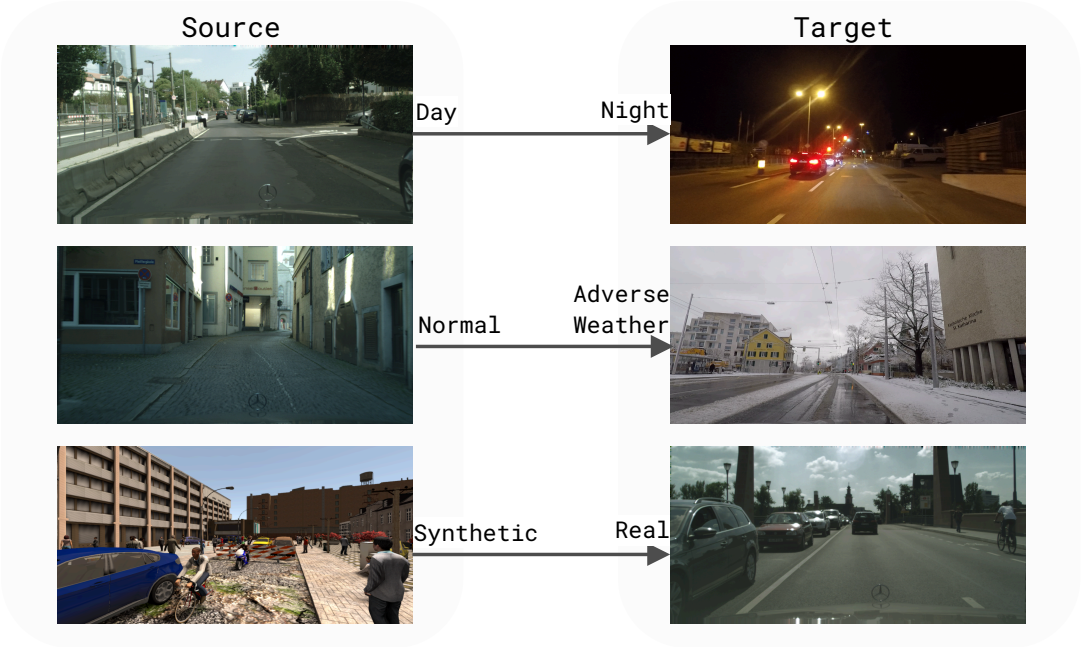


Figure 1.3: **Typical domain shifts in autonomous driving.** *Top:* Training data contains images of well-lit daytime streets [11], whereas the target domain depicts a similar urban context at night. The shift in illumination alters color distribution, introduces motion blur and head-light glare, and obscures object boundaries–factors that drastically change the visual contextual clues segmentation models rely on. [65]. *Center:* A network trained for clear weather struggles when encountering snowy or rainy scenes [67]. Snowfall and ice hide texture cues and change scene geometry (e.g., snowbanks), producing a weather-induced domain shift that confuses models trained on sunny conditions. *Bottom:* Training on synthetic imagery [63] and deploying on real-world scenes [11] introduces a texture and appearance gap that degrades performance.

1.2 Domain Shift

The success of supervised learning, where an image and its ground truth label are used to train the deep neural network, relies on 1) the vastly available labeled training data and 2) the independent and identically distributed (i.i.d.) assumption between training and testing data [49]. However, when deep learning models are deployed in the real world, the

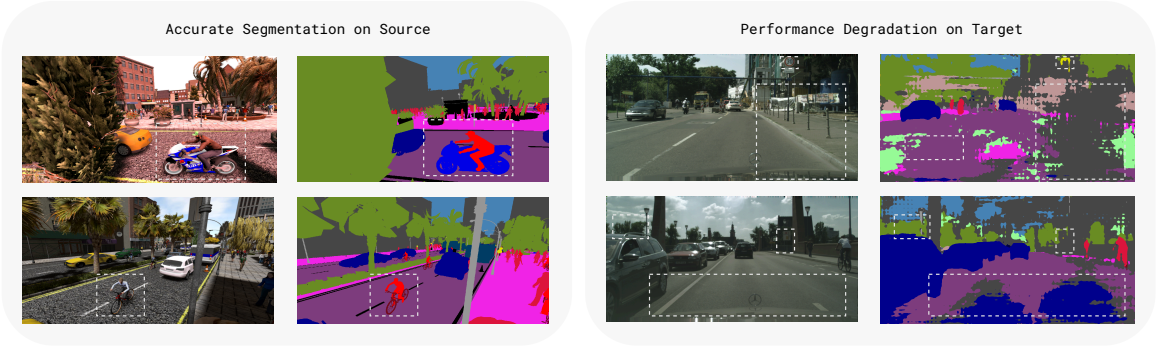


Figure 1.4: **Performance degradation under domain shift.** As shown on the right side of this figure, target domain segmentation results deteriorate when directly applying source-trained models on target data: object boundaries blur (e.g., trees), small classes vanish (e.g,. traffic signs), and misclassifications proliferate (e.,g. cars, roads, vegetation). This drop in accuracy underscores how distribution shifts between source and target data compromise segmentation performance.

i.i.d. assumption rarely holds [83, 49].

Domain shift, also called distribution shift, is common in real-world settings (Section 1.1). Differences in illumination, weather, camera pose, image quality, or visual style between training and test datasets invalidate the learned decision boundaries [83, 49]. In self-driving, a segmentation network trained on bright daytime road scenes, for example, may face nighttime, rain, or snow at deployment, and its accuracy can deteriorate significantly.

Applying a source-trained model directly to the target domain yields substantial degradation (Figure 1.4). Simpler solutions such as fine-tuning with mixed source–target augmentations – e.g., CutMix or ClassMix [90, 54] – offer only limited gains and still demand careful target image selection and hyperparameter tuning. These limitations motivate more principled methods that exploit both richly annotated source data and unlabeled target data, leading to the UDA and DASS techniques investigated in this thesis.

1.3 Domain Adaptive Semantic Segmentation

Unsupervised domain adaptation (UDA) for semantic segmentation aims to bridge the domain gap by learning invariant feature representations that generalize across different domains. A key challenge in UDA is capturing *context*—the spatial relationships between

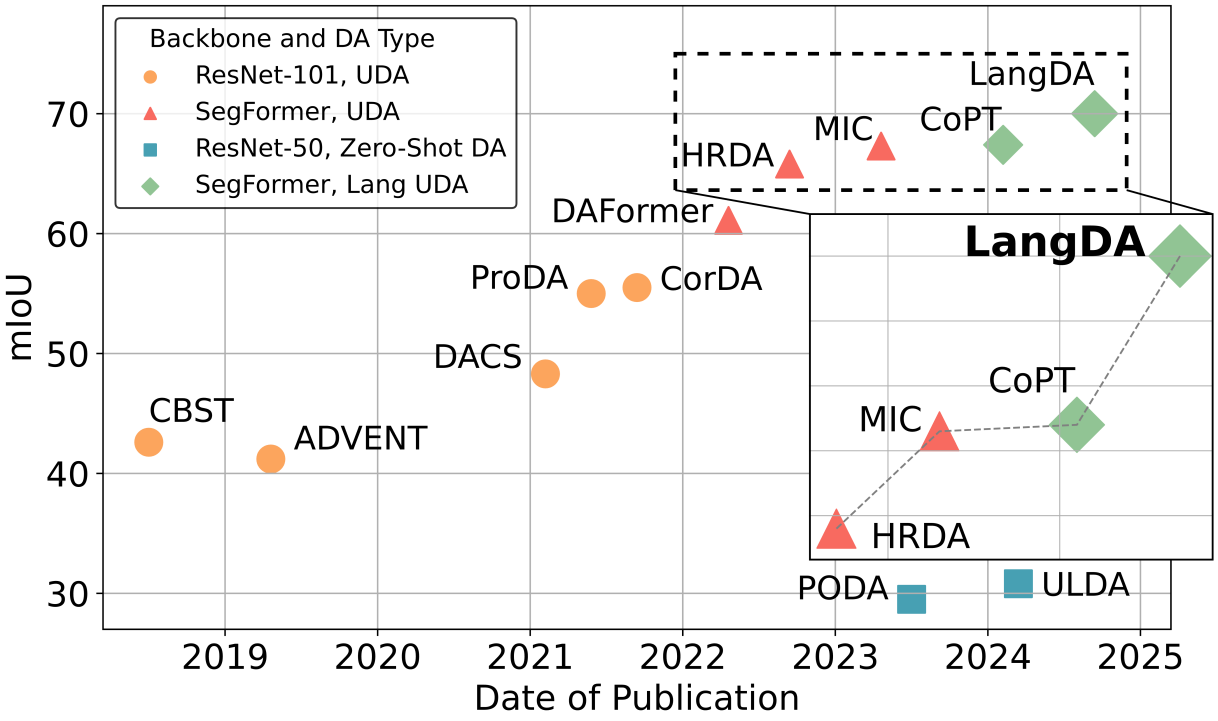


Figure 1.5: **Synthia → Cityscapes: Progress of DASS over time.** Improvements in UDA methods have plateaued in the last two years. Compared to MIC [23], which tries to learn spatial relationships only on vision domains, and CoPT [53], which employs generic language-priors, our proposed method LangDA uses contextual information from descriptive captions, achieving state-of-the-art performance.

objects in each domain–which is critical for accurate segmentation [23].

Existing vision-only UDA methods [22, 23, 21] (Fig. 2.1a) attempt to enhance context awareness by implicitly learning spatial relationships between image patches. For example, HRDA [22] uses a combination of high-resolution crops and low-resolution images to capture both fine-grained and long-range context dependencies. MIC [23] improves context awareness by masking target images, encouraging the model to infer missing information based on the surrounding context. However, these methods are prone to noisy pseudo-labels, which are biased toward the source domain [53, 23].

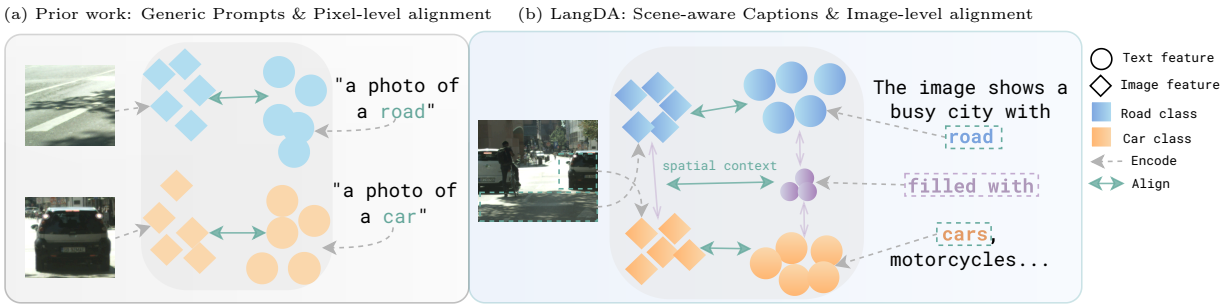


Figure 1.6: (a) Existing language-guided DASS method aligns pixel-level image features with generic prompts. (b) LangDA, our proposed method, explicitly induces spatial context by aligning scene captions containing both relational words (purple) and object names (blue and orange) with *image-level* features to drive context-aware image segmentation.

1.4 Language-guided Domain Adaptive Semantic Segmentation

To mitigate noisy predictions caused by domain gaps, recent self-training methods begin leveraging generalized representations from language models. CoPT [53], in particular, introduces language-based guidance by aligning class-level text embeddings with pixel features. Specifically, CoPT minimizes the cosine similarity between the covariance matrices of text representation (e.g., CLIP [59] embedding of the phrase "a photo of a class") and its corresponding average-pooled pixel features, such as for the class "car." While this improves domain alignment, CoPT and similar approaches focus on class-level alignment and fail to capture the broader context between different objects (e.g., the relationship between a car and a pedestrian). However, capturing these inter-object relationships is crucial for complex scene understanding and robust segmentation.

Motivated by the above, this thesis proposes LangDA, where we explicitly build contextual scene understanding through text. Unlike existing language-guided methods [53, 26, 13, 89] using generic class-level prompts (Figure 1.6a), LangDA aligns image scenes with descriptive captions (e.g., "a {pedestrian} is on the {sidewalk}, and the street is lined with {buildings}..."; see Figure 3.2a). These captions encode semantic context structure using spatial phrases ("on the", "next to", "in front of") and logical connectors ("and", "with"), which are not captured from class names alone (Figure 1.6b). To fully leverage the spatial relationships in the caption, we introduce an image-level consistency module, bringing the entire image's features closer to the corresponding caption. We

demonstrate that leveraging scene relationships in language allows the model to generalize much better to target images.



Figure 1.7: Existing DASS methods struggle with segmenting visually similar classes (e.g., misclassifying the interior of the sidewalk pixels as road). LangDA, our proposed scene-aware approach, mitigates this by leveraging spatial cues in language to improve class separability.

We validate LangDA on three different DASS settings: Synthia → Cityscapes, Cityscapes → ACDC, and Cityscapes → DarkZurich. LangDA achieves state-of-the-art performance in all three UDA settings, surpassing existing methods [23, 53] by **2.6%**, **3.9%**, and **1.4%** respectively, for the challenging semantic segmentation tasks with domain shifts (Figure 1.8). Our ablation studies further highlight the superiority of context-aware image-level alignment over pixel-level alignment. These results confirm LangDA's capacity to extract spatial relationships encoded in language representations for robust domain adaptation.

1.5 Thesis Contribution

This thesis contributes a more accurate and robust language-guided framework for domain adaptive semantic segmentation (DASS), which captures rich object relationships at the scene level and operates without manual prompt engineering or human feedback. Specifically, this thesis introduces LangDA, which aligns a descriptive scene embedding with the corresponding image embedding, thereby encoding rich contextual relationships that existing class-level alignment methods cannot exploit.

Another key contribution of this thesis is the ability of LangDA to adapt to *unknown* domain gaps, in contrast to existing methods that assume knowledge of the domain shift. To achieve this, we propose a context-aware caption generator that automatically produces image-specific descriptions using a Vision–Language Model (VLM) and a Large Language Model (LLM), both initialized with a fixed random seed to ensure reproducibility. This removes the need for predefined textual supervision to describe domain gaps (e.g., "*daytime* photo" vs. "*nighttime* photo"), which proves unreliable in real-world scenarios where domain

shifts are unpredictable, such as tumor scans exhibiting color and spatial variations across hospitals [24, 41, 5].

Furthermore, this thesis contributes standardized caption generation for reproducibility and consistent benchmarking in DASS. We release the first captioning datasets for Synthia [63] and Cityscapes [11], enabling future research in language-guided semantic segmentation with or without domain shifts. By standardizing the captioning process, LangDA eliminates the inconsistencies introduced by handcrafted prompt engineering and supports robust comparative evaluation.

Additionally, this thesis contributes an automated and scalable captioning framework that removes reliance on manual prompt tuning [26, 89, 13, 53] and resource-intensive human feedback [15]. The proposed context-aware caption generator produces captions automatically without requiring manual prompt-specific tuning describing the same scene (e.g., "a daytime photo", "a bright image", or "a sunlit road"), thus ensuring consistency across domain descriptions and eliminating subjective prompting choices.

From a systems perspective, this thesis also contributes an architecture that is as fast at inference as existing vision-only models, while delivering significantly improved segmentation performance. Unlike many cross-modal methods that require language input at both training and inference time, LangDA uses text only during training as an auxiliary learning signal. At inference, it relies solely on the visual model, making it lightweight and practical for real-world deployment. Notably, the method shows substantial improvements on difficult classes such as *person*, *road*, and *sidewalk*, which are often misclassified under domain shift.

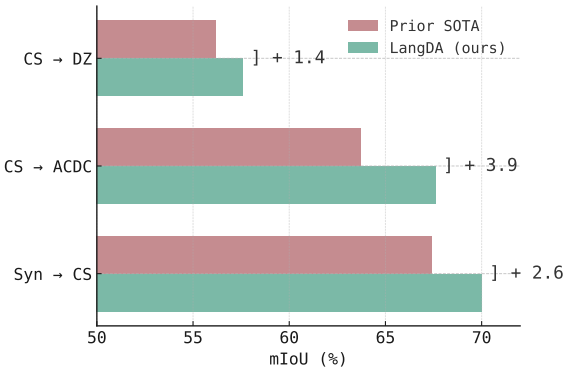


Figure 1.8: LangDA sets the new state-of-the-art in DASS in three settings.

Furthermore, LangDA does not require fine-tuning a VLM or LLM, setting it apart from existing methods that often depend on computationally expensive fine-tuning and learning from human feedback. By leveraging a frozen pre-trained VLM and LLM, our approach reduces training cost while preserving strong adaptation performance.

Finally, this thesis contributes a new state-of-the-art method in the field of DASS. The method proposed in this thesis, LangDA, achieves 70.0%, 67.6%, and 57.6% mean Jaccard

Index on Synthia → Cityscapes, Cityscapes → ACDC, and Cityscapes → DarkZurich, respectively, surpassing the previous state-of-the-art by 2.6%, 3.9%, and 1.4% , as seen in Figure 1.5. These results demonstrate the effectiveness of LangDA in improving segmentation performance under a range of challenging domain shifts while maintaining efficiency and scalability.

1.6 Overview

Chapter 2 establishes the background and academic context of this thesis, reviews existing literature in domain adaptive semantic segmentation, and positions this work within the broader research landscape.

Chapter 3 describes the methodology of the proposed architecture, LangDA, for language-guided domain adaptive semantic segmentation, as illustrated in Figure 3.1. LangDA consists of source-target distillation for learning domain-invariant visual features, a context-aware caption generator for creating scene descriptions, and image-level consistency alignment for integrating language and vision features into the network.

Chapter 4 details the experimental setup, datasets used, and both quantitative and qualitative results, including ablation studies. Notably, the proposed method achieves state-of-the-art performance on all three standard benchmarks: synthetic-to-real, day-to-night, and normal-to-adverse-weather.

Chapter 5 summarizes the contributions, discusses limitations, and outlines potential future research directions.

Chapter 2

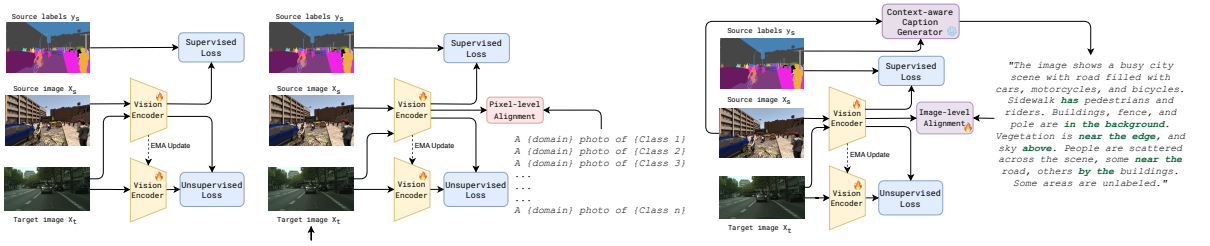
Background

This chapter reviews existing domain adaptation literature relevant to this thesis, highlights gaps in current knowledge, and positions this work within the broader academic context to establish its significance.

Modern domain adaptation algorithms aim to learn representations that generalize across both labeled source and unlabeled target domains, enabling robust performance despite domain gaps. Section 2.1 examines unsupervised domain adaptation (UDA) methods for vision, which primarily use statistical minimization to reduce distribution differences between image domains or adversarial training to learn domain-invariant representations. Section 2.2 discusses recent self-training-based UDA approaches that distill source domain information into the target domain. Section 2.3 explores the zero-shot setting, where no target image is available, but alternative information such as language informs the source domain about the target domain. Finally, Section 2.4 reviews language-guided UDA methods for classification, detection, and semantic segmentation, positioning our work within the existing literature and highlighting its significance.

2.1 Vision-only Unsupervised Domain Adaptation

Before the advent of UDA, models trained solely on a labeled source domain were directly applied to an unlabeled target domain under domain shifts. This naive transfer typically resulted in significant performance degradation, particularly in dense prediction tasks like semantic segmentation.



(a) Vision-only (b) CoPT (ECCV'24) [53] (c) LangDA (Ours)

Figure 2.1: Unsupervised Domain Adaptation Methods: An Overview. **(a) Vision-only UDA** leverages an EMA-updated teacher-student framework with consistency losses to segment unlabeled target data. **(b) CoPT** uses LLM-generated class-wise text prompts and performs pixel-level alignment (aligns pixel features to corresponding class prompts), not focusing on spatial relationships in language. They also require additional supervisory text prompts for target domain. **(c)** Our proposed method **LangDA** utilizes context-aware image captions and performs image-level alignment (aligns image features to the image captions) to facilitate context-aware domain-invariant adaptation. Words providing context are highlighted in **green**.

Early efforts in UDA focused on data-level mixing strategies, such as CutMix [90] and ClassMix [54], which blended source and target images to encourage domain-invariant learning. These approaches laid the groundwork for more principled UDA methods.

Modern UDA techniques adapt a model trained on a labeled source domain to an unlabeled target domain by leveraging one or more of the following approaches: discrepancy minimization [51, 72, 71, 82, 17], adversarial training [82, 16], or self-training [7, 21, 32, 40, 76]. Discrepancy minimization reduces the domain gap using statistical distance functions such as maximum mean discrepancy [51], correlation alignment [72, 71], or entropy minimization [82, 17]. Adversarial training uses a learned domain discriminator within a GAN framework [16] to promote domain-invariant inputs, features, or outputs [82]. Self-training generates pseudo-labels [40] for the target domain based on confidence thresholds [21], with consistency regularization [76, 79] often applied to ensure robustness across different data augmentations [23, 4] and domain-mixup [54, 90, 32, 79]. Other works also apply consistency regularization using prototypes [93] and auxiliary task correlation [84].

2.2 Unsupervised Domain Adaptation with Self-Training

In particular, self-training has shown strong results in segmentation when attempting to learn spatial relations using self-supervised objective. To better leverage ImageNet's real-world high-level semantic classes, DAFormer [21] proposes a transformer-based DASS method and regularizes the bottleneck image features with ImageNet features. HRDA [22] uses high-resolution crops in conjunction with low-res crops to improve fine-grained representation, while MIC [23] leverages masked image modeling to try to implicitly learn spatial relationship in images. Although self-training based methods are more stable and outperform other UDA approaches [21], they remain vulnerable to noisy pseudo-labels dominated by source features, limiting target domain performance [53].

To further bridge the domain gap, recent methods such as CoPT [53] turn to CLIP-based models to leverage its generalized world prior. Although CoPT [53] aligns the covariance matrix of image and text features, it performs this alignment at pixel-level and does not take advantage of the spatial relationship between objects encoded in VLM's representation. Our work addresses this research gap by leveraging language guidance to bridge the source and the target domain in DASS.

2.3 Zero-shot Domain Adaptation with Language

In zero-shot domain adaption for semantic segmentation, recent works adapt to unseen domains using textual descriptions of the unavailable target data. For instance, PODA [13] utilizes text embedding from the CLIP [59] model to approximate the target visual domain. ULDA [89] notes PODA requires separate segmentation heads and adaptation steps for each target domain. Its follow-up work addresses this by enabling a single model to adapt to multiple domain settings. However, ULDA is computationally costly and requires simultaneous training on several domains.

2.4 Domain Adaptation with Language

Recent work in unsupervised domain adaptation (UDA) has largely focused on vision-only strategies, including discrepancy minimization, adversarial training, and self-training, to bridge the gap between labeled source and unlabeled target domains. While effective, these approaches often rely solely on visual cues and may struggle to generalize under

large appearance shifts or in the absence of fine-grained context. In contrast, language-guided domain adaptation introduces an additional modality–natural language–that encodes semantic and contextual knowledge often invariant across domains. By leveraging pretrained vision–language models, language-guided approaches offer a promising alternative to improve robustness and interpretability in UDA, especially for classification, detection, and more recently, dense prediction tasks.

In UDA for classification and detection, Lai *et al.* [38] initially formulated a pseudo-labeling setting and active debiasing of CLIP for unsupervised domain-adaptive classification. Following this, Du *et al.* [12] exploit domain-invariant semantics by mutually aligning visual and text embeddings for domain adaptive classification. Jin *et al.* [31] leverage language to compute fine-grained relationships between different parts of an object for UDA-detection. Lai *et al.* [36] design a domain-aware pseudo-labeling scheme for effective domain disentanglement.

In semantic segmentation, Lim *et al.* [44] uses CLIP for adaptation when source and target domain have different class names. CoDA [15] adapts to adverse weather by generating additional images via text-to-image synthesis and requires costly human feedback for image selection. Despite their success, we do not compare existing methods against [15], as it relies on extra data and manual intervention, making direct comparison to methods trained solely on existing data neither fair nor meaningful. CoPT [53] aligns the covariance matrix of class-wise text embeddings (averaged from source and target domains) with pixel-level image features for DASS. In contrast, no existing DASS method directly aligns spatial relationships in text with visual representations. Our approach fills this research gap by capturing object relationships in language descriptions, and aligning it with image features to mitigate visual domain discrepancies.

2.5 Summary

Existing DASS methods typically follow either vision-only [51, 72, 71, 82, 17, 82, 16, 7, 21, 32, 40, 76] or language-guided approaches [13, 31, 89, 12, 36, 44, 15, 53]. Vision-only methods, such as masking [23] and multi-resolution cropping [22], aim to enhance context-awareness but often suffer from noisy pseudo-labels biased toward the source domain [53]. Language-based methods, by contrast, rely on generic class-level prompts (e.g., "a photo of a class") to introduce semantic priors [53], but fail to capture spatial relationships between objects–information essential for dense prediction tasks like semantic segmentation. While recent work such as CoPT aligns image and text features, it does so only at the pixel level and overlooks scene-level context.

This thesis bridges these two directions by proposing LangDA, a language-guided framework that combines the strengths of both vision-only and language-based UDA. LangDA introduces a context-aware caption generator that produces scene-level descriptions capturing object relationships using frozen VLMs and LLMs. These descriptions are then aligned with global image features in the image-level alignment module, enabling the model to incorporate spatial semantics alongside visual cues. This multimodal approach offers more robust domain adaptation, particularly under large appearance shifts and complex visual layouts, advancing the state of the art in semantic segmentation under domain shift.

Chapter 3

Language-guided Domain Adaptive Semantic Segmentation

3.1 Methodology

3.1.1 Problem Formulation

In DASS task, we learn a model $f_\theta : \mathbb{R}^{H\times W\times 3} \to [0,1]^{H\times W\times K}$ parameterized by θ that performs pixel-wise classification of K classes on an unlabeled target domain dataset $\mathcal{D}_\mathrm{T} = \{x_T^{(i)} \mid x_T^{(i)} \in \mathbb{R}^{H\times W\times 3}\}$, given access to a labeled source domain dataset $\mathcal{D}_\mathrm{S} = \{(x_S^{(i)}, y_S^{(i)}) \mid x_S^{(i)} \in \mathbb{R}^{H\times W\times 3}, y_S^{(i)} \in \{0,1\}^{H\times W\times K}\}$. We utilize a student model $g_\theta : \mathbb{R}^{H\times W\times 3} \to [0,1]^{H\times W\times K}$ parameterized by θ and a teacher model $h_\phi : \mathbb{R}^{H\times W\times 3} \to [0,1]^{H\times W\times K}$ parameterized by ϕ.

3.1.2 Overview

Our proposed method, LangDA, leverages an EMA-updated one-stage knowledge distillation framework for online self-training (Sec. 3.1.3) to adapt across visual domains. The knowledge distillation architecture consists of two models: a student model and a teacher model. The student network is trained with two segmentation losses simultaneously: a supervised loss on the labeled source dataset D_S, and an unsupervised loss on the unlabeled target dataset D_T. For the unlabeled target data, the teacher network generates pseudo-labels that are used as predictions to inform the student network on D_T.

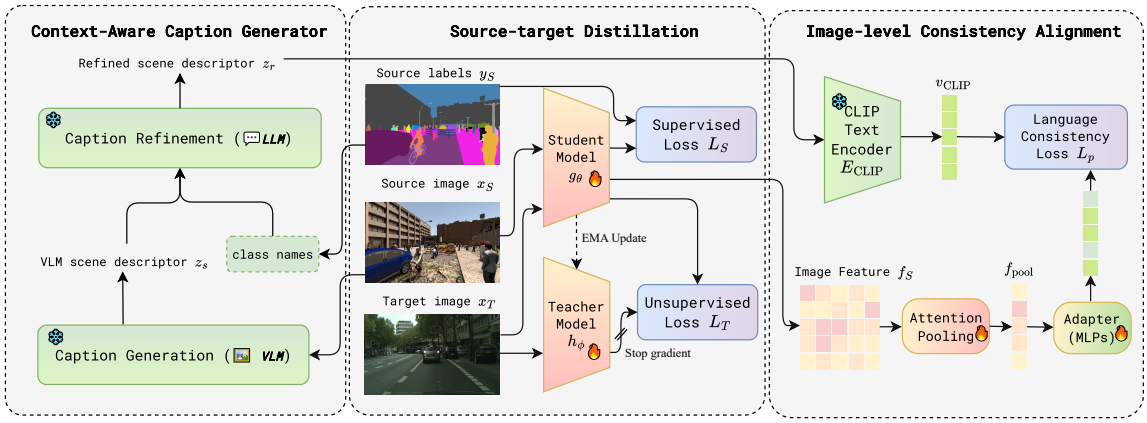


Figure 3.1: **LangDA Architecture**. LangDA is a prompt-driven UDA framework that leverages contextual language descriptions to bridge domain gaps between labeled source images and unlabeled target images. LangDA includes two modules: context-aware caption generation and language-consistency alignment. *Left:* Context-aware generation is a two step process. First, a captioning model generates captions that encode context relationships for the source image (e.g. "there is a sidewalk *on one side of* the street"). Then, the captions are improved by passing class names from ground truth labels into an LLM. *Right:* In the image-level consistency alignment module, an adapter (explained in Section 3.1.5) projects the image features from the trained network onto the same latent space as text embeddings. The LangDA image encoder is trained from scratch because the CLIP image encoder performs poorly on semantic segmentation tasks.

In addition to the supervised and unsupervised objectives for aligning image domains, we propose a Contextual Language Consistency objective in Sec. 3.1.5. This objective utilizes a context-aware caption generator (Sec. 3.1.4) that processes the source image to generate scene text embeddings, which are then aligned with the visual features extracted from the student network via image-level alignment (Sec. 3.1.5). This alignment reinforces context awareness and domain-invariance within the student model.

Finally, the teacher network's parameters are updated through an exponential moving average (EMA) of the student network's weights, ensuring gradual and stable adaptation throughout training. Figure 3.1 summarizes LangDA's architecture. The individual modules are explained in the following subsections.

Caption Generation: VLM

Query (Image): 

Query (Text):
Describe the image in detail for semantic segmentation tasks. Be sure to include the class names {CLASS_NAMES} and their pixel locations.

VLM: {VLM_CAPTION}
The image depicts a busy city street with a mix of vehicles and pedestrians. There are several cars, including a blue car parked on the side of the road, and a motorcycle. A bicycle is also present in the scene. A person is riding a bicycle, while another person is riding a motorcycle. There are numerous people walking along the sidewalk, some of them carrying handbags. A few pedestrians are also riding bicycles. The street is lined with buildings, and there is a traffic light visible in the scene. The sky is visible in the background, adding to the urban atmosphere.

(a) We generate scene descriptions for source images using a VLM [47]. Class names are acquired from ground-truth labels $y_S^{(i)}$. We can see the VLM provides contextual relationships, such as "street is lined with buildings" and "numerous people walking along the sidewalk".

Caption Refinement: LLM

System: *You are a helpful assistant for refining and condensing detailed image caption descriptions for semantic segmentation.*

Query: Shorten the description to less than 77 tokens. Do not use quotation marks or parentheses. Be sure to include the class name ['road', 'sidewalk', 'building', 'wall', 'fence', 'pole', 'vegetation', 'sky', 'person', 'rider', 'car', 'motorcycle', 'bicycle', 'unlabeled'] and their pixel locations. The description is {VLM_CAPTION}

LLM:
The image shows a busy city scene with road filled with cars, motorcycles, and bicycles. Sidewalk has pedestrians and riders. Buildings, fence, and pole are in the background. Vegetation is near the edge, and sky above. People are scattered across the scene, some near the road, others by the buildings. Some areas are unlabeled.

(b) We summarize generated captions with an LLM. We appended a system-level prompt to inform the LLM of our semantic segmentation objective. We can see that LLM preserves spatial relationships from the VLM captions, as in "Sidewalk has pedestrians and riders".

Figure 3.2: **Left**: VLM Caption Generation. **Right**: LLM Caption Refinement Module.

3.1.3 Source-target Distillation for Visual Domain

To adapt from a labeled source image domain $\mathcal{D}_\text{S}$ to an unlabeled target image domain $\mathcal{D}_\text{T}$, we use single-stage online self-training (ST) [79, 21] to distill visual knowledge from the source to the target. This involves simultaneously training a student model $g_\theta(\cdot)$ and

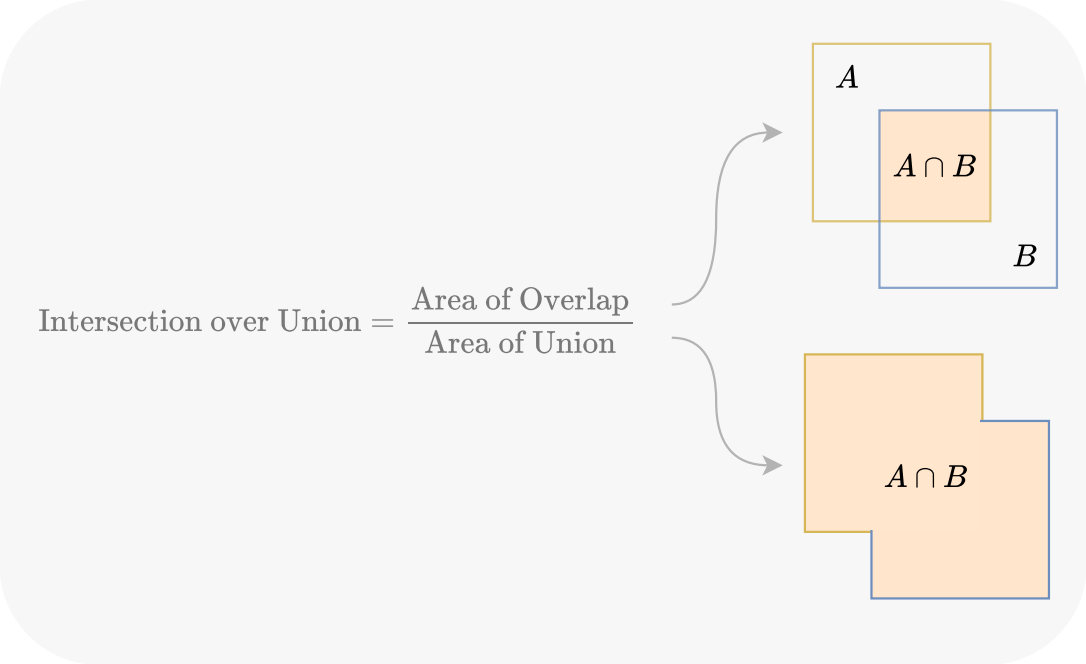


Figure 3.3: Intersection over Union, Illustrated.

a teacher model $h_\phi(\cdot)$. The student segmentation network g_θ is trained on $\mathcal{D}_\mathrm{S}$ using a categorical cross-entropy loss:

$$\mathcal{L}_S^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} y_S^{(i,j,c)} \log g_\theta(x_S^{(i)})^{(j,c)} \tag{3.1}$$

For target data $\mathcal{D}_\mathrm{T}$ with no labels, we generate pseudo-labels (p_T) via teacher network $h_\phi(\cdot)$, using the argmax of the softmax output (Eq. (3.2), sg denotes the stopping gradient).

$$p_T^{(i,j,c)} = \mathbb{I}[c = \arg\max_{c'} \ \mathrm{sg}(h_\phi(x_T^{(i)})^{(j,c')})] \tag{3.2}$$

A quality estimate for pseudo-labels is provided based on the ratio of pixels exceeding a confidence threshold τ in the softmax probability.

$$q_T^{(i)} = \frac{1}{H\cdot W}\sum_{j=1}^{H\times W} \mathbb{I}[\max_{c'} h_\phi(x_T^{(i)})^{(j,c')} > \tau] \tag{3.3}$$

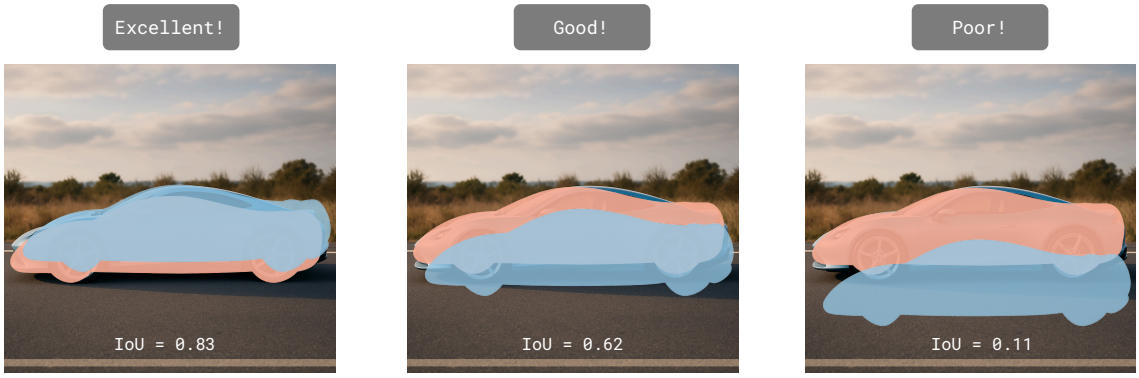


Figure 3.4: Examples: What is a good IoU?

These pseudo-labels and their confidence estimates are used to train g_θ on the target domain to compute the unsupervised loss for the teacher model.

$$\mathcal{L}_T^{(i)} = -\sum_{j=1}^{H\times W}\sum_{c=1}^{C} q_T^{(i)} p_T^{(i,j,c)} \log g_\theta(x_T^{(i)})^{(j,c)} \,. \tag{3.4}$$

Pseudo-labels can either be generated online or offline. Following [21], we opt for online self-training due to its simplicity and single training stage, which is crucial for comparing and ablating network architectures. In online self-training, the teacher network is updated as the exponentially moving averages of the student network after each training step.

$$\phi_{t+1} \leftarrow \alpha\phi_t + (1-\alpha)\theta_t \tag{3.5}$$

3.1.4 Context-Aware Caption Generator

Capturing spatial context relations between objects is critical for dense prediction tasks [23]. To capture contextual information in our text prompts, we generate detailed captions describing each scene using multi-modal foundational models and utilize a Vision Language Model (VLM) to generate image-level text prompts. Specifically, we choose the open-source captioning model LLaVA [47], which has GPT-4V level capabilities and is trained in a unified vision-language embedding space. LLaVA utilizes visual prompt instruction tuning and provides rich visual and linguistic context, making it ideal for our task at hand.

We develop our captions once at model initialization and store them in a memory bank. For each image, we obtain a caption $z_S^{(i)}$ from LLaVA [47] (Fig. 3.2a), forming the caption bank $\mathcal{C}_\mathrm{S} = z_S^{(i)} \mid z_S^{(i)} \in \mathbb{R}^\ell$, where ℓ is the number of tokens.

To refine $\mathcal{C}_S$ and obtain $\mathcal{C}_r = z_r^{(i)} \mid z_r^{(i)} \in \mathbb{R}^\ell$, $\ell \leq 77$, we design a system-level prompt [92] (see Fig. 3.2b) to precisely guide the Large Language Model's response style and expertise, ensuring the caption is tailored for segmentation tasks. Next, we refine each $z_S^{(i)} \in \mathcal{C}_S$ using a Large Language Model (LLM) [30] to retain only the classes present in its ground truth segmentation mask $y_S^{(i)}$, as shown in Fig. 3.2b. This avoids hallucination [3, 35], since each source image $x_S^{(i)}$ may contain only a subset of all possible classes. Since CLIP [59] accepts a maximum token length of 77, the refinement process condenses the VLM caption from 140 tokens to approximately 70 tokens while preserving key contextual information. See 3.2.2 for an analysis of token lengths in context descriptions.

To obtain text features for image-level alignment, we pass the refined captions $\mathcal{C}_r$ to the frozen CLIP encoder E_{CLIP} [59] to obtain the set of text feature vectors,

$$\mathcal{V}_{\text{CLIP}} = \{v_{\text{CLIP}}^{(i)} \mid v_{\text{CLIP}}^{(i)} = E_{\text{CLIP}}(z_r^{(i)}),\ v_{\text{CLIP}}^{(i)} \in \mathbb{R}^C\}$$

where C is the CLIP multi-modal embedding space dimension.

3.1.5 Image-level Consistency Alignment

To align source image features with generalized context-aware text features, we impose an image-level minimization objective on the distance between CLIP textual features $\mathcal{V}_{\text{CLIP}}$ and the source image feature $\mathcal{F}_S$, $\mathcal{F}_S = \{f_S^{(i)} \mid f_S^{(i)} = g_\theta(x_S^{(i)}), f_S^{(i)} \in \mathbb{R}^{C \times H \times W}\}$, where C is the dimension of the CLIP multimodal embedding space, H and W are the height and width of the source image features.

To enforce image features to share the same dimensions as the CLIP embeddings, we apply attention pooling [59] and transform $f_S^{(i)}$ into $f_{\text{pool}}^{(i)}$, where $f_{\text{pool}}^{(i)} \in \mathbb{R}^C$. Since the CLIP image encoder is trained on sparse prediction tasks such as classification, it is known to perform poorly on segmentation tasks [9, 94]. Therefore, we train a dedicated image encoder specifically optimized for segmentation tasks.

To ensure image and text features are in a shared latent space, we project frozen CLIP text features and image features into the same space using a trainable multilayer perception layer (or an "adapter"). Parameter-efficient fine-tuning works have used adapters in NLP and multi-modal feature fusion without distribution shift [19, 69, 73]. We empirically demonstrate its effectiveness in DASS (Sec. 4.1).

To bring text features closer to image features, we define a language consistency objective function:

$$\mathcal{L}_p^{(i)}(f_{\text{pool}}^{(i)}, v_{\text{CLIP}}^{(i)}) = 1 - \frac{f_{\text{pool}}^{(i)} \cdot v_{\text{CLIP}}^{(i)}}{\|f_{\text{pool}}^{(i)}\| \, \|v_{\text{CLIP}}^{(i)}\|}. \tag{3.6}$$

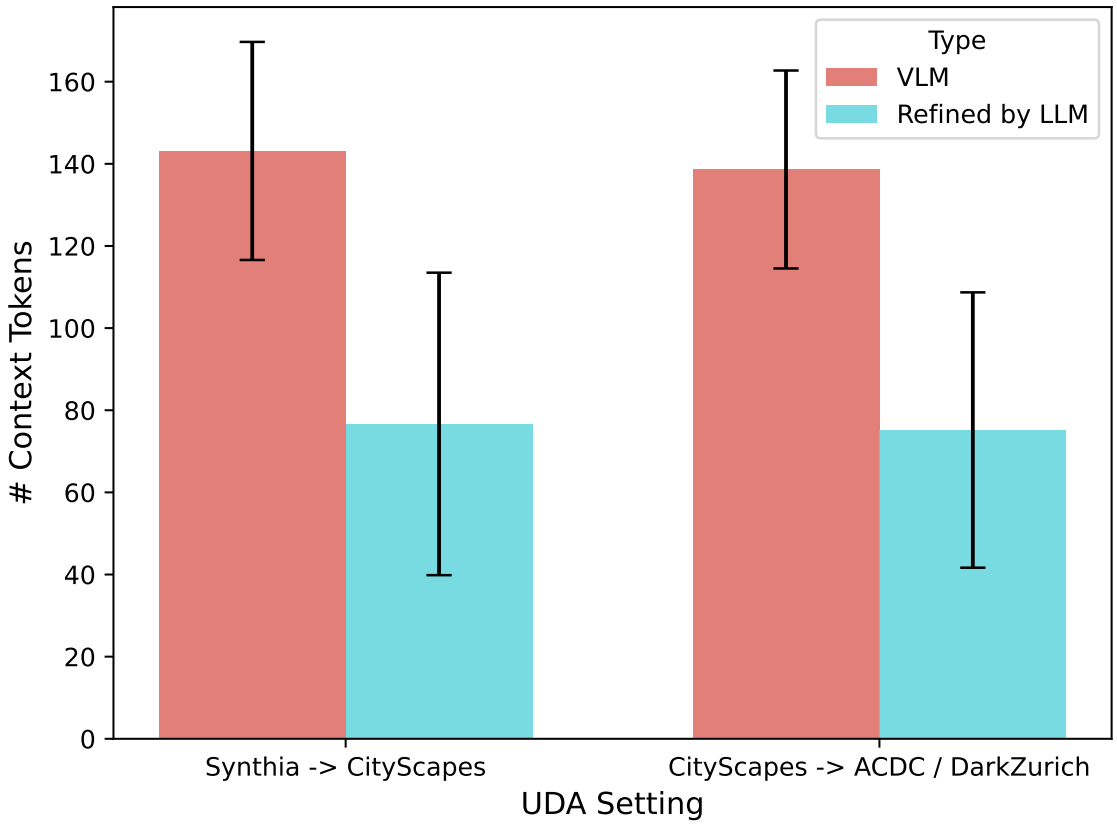


Figure 3.5: **Token Length After Refinement on Synthia [63] and Cityscapes [11].** Token lengths are shorter after refinement, averaging around 77 tokens – the maximum context length accepted by CLIP.

To the best of our knowledge, we are the first work to employ this exact language consistency loss in DASS. Our loss leverages the cosine similarity distance metric, used by CLIP [59] for aligning multimodal embeddings, to explicitly align source image features with text embeddings. The language consistency objective guides the source image features toward the text embedding. Additionally, the global target image features are implicitly steered towards the text feature space through the EMA model update in Eq.(3.5), further reinforcing the learned multimodal consistency.

The overall UDA loss $\mathcal{L}$ is the weighted sum of the supervised loss, unsupervised loss and language consistency loss $\mathcal{L} = \mathcal{L}_S + \mathcal{L}_T + \lambda_{\text{p}}\mathcal{L}_{\text{p}}$. This new loss establishes a stronger link between visual and textual domains, driving state-of-the-art performance in DASS.

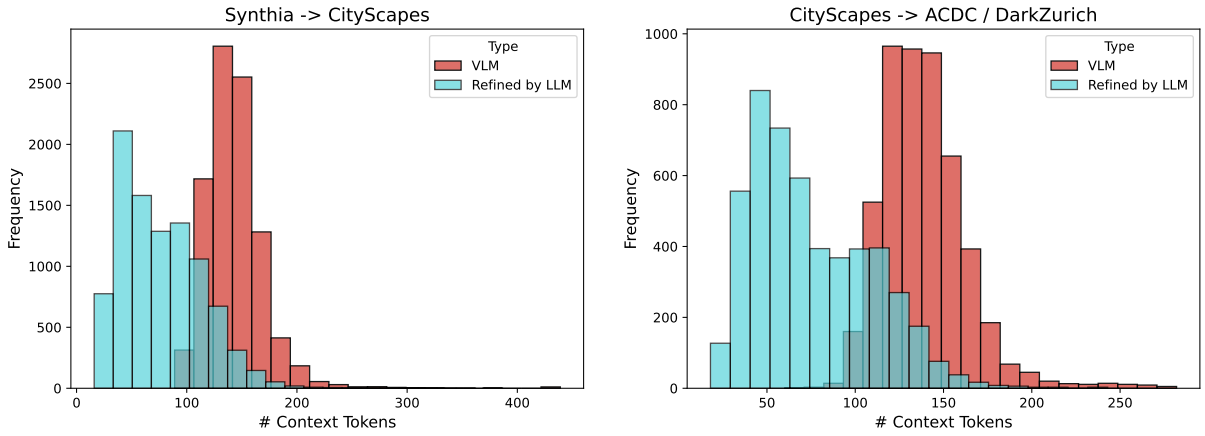


Figure 3.6: **Left**: Token Length Distribution on Synthia. Token length for VLM is much longer than token length for LLM after refinement. Most token length centered around 150 before refinement, whereas token length centered around 70 after refinement. **Right**: Token Length Distribution on Cityscapes. Note the distribution of token length is much longer and less uniform compared to Synthia. This is because Cityscapes is real-world data, which contains varied scenes and requires more detailed description.

3.2 Explanation and Analysis on Caption Generation

3.2.1 Prompt Generation Process

Figure 3.2a illustrates the process of generating a contextual scene description using a VLM [47] for a given source image $x_S^{(i)}$. We prompt the VLM with a source image and a prompt including class names (shown in green) obtained from ground truth segmentation masks. We can see that the VLM clearly characterizes the contextual relationship in the given image.

The scene description generated by the VLM is much longer than 77 tokens, the maximum context length supported by CLIP [59]. We further use an LLM [30] to refine the generated captions. Figure 3.2a shows the LLM [30] refinement module of the corresponding source image $x_S^{(i)}$, note that the token length is less than 77 after refinement and the contextual relationship is preserved.

3.2.2 Analysis on Quality of Image-level Descriptors

Figure 3.5 shows the tokens before refinement have average context lengths of around 140 tokens on both Synthia and Cityscapes, which is too long for the CLIP [59] context length that only accepts a maximum of 77 tokens. After refinement, we can see the average context length drops down to around 77 tokens for both datasets, showcasing the necessity of our LLM-based caption refinement process.

This is further elucidated in Figure 3.6, where the distributions of context lengths of raw and refined captions are visualized. From this figure, we can clearly see that the distributions of context lengths of the refined captions for both datasets are centered around 70. These results, combined with the example refined prompt in Figure 3.2b, demonstrate that the LLM refinement module reduced the number of tokens significantly while preserving the context relationships.

3.3 Experiments

3.3.1 Datasets

Following standard practices in DASS, we explore three self-driving adaptation scenarios: synthetic-to-real, clear-to-adverse-weather, day-to-night; using Synthia $\rightarrow$ Cityscapes, Cityscapes $\rightarrow$ ACDC, and Cityscapes $\rightarrow$ DarkZurich, respectively (Figure 3.7). Synthia [63] serves as the synthetic source, containing 9,400 images at 1280×760 resolution. Real dataset benchmarks include Cityscapes [11] with 2,975 training and 500 validation images at 2048×1024 resolution for clear weather, DarkZurich [65] with 2,416 training and 151 test images for nighttime, and ACDC [67] with 1,600 training and 2,000 test images for adverse weather (fog, night, rain, and snow).

3.3.2 Implementation details

To evaluate LangDA's robustness, we combine LangDA with several vision-only DASS methods (DAFormer [21], HRDA [22], and MIC [23]) and evaluate on Synthia$\rightarrow$Cityscapes.

For main experiments, we build on MIC, the state-of-the-art vision-only DASS method. LangDA uses LLaVA [47] to generate spatially aware scene captions, refined with Mistral-Large-2 [30], which supports a 128k token context window and has an Apache 2.0 license.



Figure 3.7: **Domain adaptation dataset examples.** Each column shows three representative images from Synthia [63], Cityscapes [11], DarkZurich [65], and ACDC [67]. The ACDC examples correspond to fog (top), snow (middle), and rain (bottom). All images are uniformly resized for consistency.

Refined captions are encoded using a frozen ViT-B/16 CLIP text encoder [59]. All backbones are initialized with ImageNet [64] pretrained.

For the default UDA setting, we follow the training scheme used by HRDA [22]. Specifically, we utilize the AdamW [52] with a learning rate of 6×10^{-5} for the encoder and 6×10^{-4} for the decoder, a batch size of 2, linear learning rate warmup, a loss weight λ=1, an EMA factor α=0.999, quality threshold $\tau = 0.968$, DACS [79] data augmentation, Rare Class Sampling [21], and ImageNet Feature Distance [21]. All experiments were conducted on one NVIDIA RTX A6000 GPU. All our implementations were carried out using the MMSegmentation framework [10].

3.3.3 Metrics

We evaluate our method using the standard semantic segmentation metric, the mean Jaccard Index, also known as mean Intersection over Union (mIoU).

Intersection over Union (IoU), illustrated in Figure 3.3, is the primary evaluation metric in semantic segmentation, measuring the overlap between predicted segmentation masks and ground truth. For multi-class semantic segmentation, mIoU is computed as the average IoU across all classes, providing a comprehensive measure of segmentation performance.

Mathematically, IoU quantifies the similarity between two arbitrary shapes ($A, B \subseteq \mathbb{S} \in \mathbb{R}^n$), calculated as:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{3.7}$$

.

IoU is scale-invariant, meaning that its value remains consistent regardless of the size of the space $\mathbb{S}$ in which A and B exists [61]. This property makes IoU a widely used metric in computer vision tasks, ensuring robust evaluation across varying spatial scales.

Generally, an mIoU score above 60%, as seen in Figure 3.4, is considered indicative of good segmentation performance, reflecting a significant overlap between predicted and ground truth segmentation masks.

3.3.4 Summary

This chapter presents the proposed methodology, LangDA (Section 3.1), provides additional discussion and analysis on context-aware caption generator (Section 3.2), and outlines the experimental setup (Section 3.3).

Section 3.1 describes the proposed methodology, which consists of three key components: (1) source–target visual distillation (Section 3.1.3), (2) a context-aware caption generator (Section 3.1.4), and (3) image-level alignment via a Contextual Language Consistency objective (Section 3.1.5). Source–target distillation (Section 3.1.3) adapts across visual domains by employing an online self-training framework, where the models are trained using supervised loss on labeled source images and unsupervised loss on unlabeled target images. In addition to visual supervision, we introduce an image-level alignment module (Section 3.1.5) that promotes domain-invariant feature learning by aligning global image features with scene-level text embeddings. These embeddings are derived from descriptive captions generated by a context-aware caption generator, which captures rich spatial relationships between objects. Together, these components combine visual and language-based supervision for robust domain adaptation, as illustrated in Figure 3.1.

Section 3.2 provides additional details on the prompt generation process, including visualizations (Figure 3.5, Figure 3.6) and preliminary analysis that justifies the necessity of the caption refinement module. Section 3.3 outlines the four domain adaptation datasets used (Section 3.3.1), describes the implementation details and training configurations (Section 3.3.2), and discusses the evaluation metric–mean Intersection over Union (mIoU)–used for quantitative comparison (Section 3.3.3).

Chapter 4

Results

4.1 Main Results

4.1.1 Quantitative Results on Synthia → CS.

To assess the proposed method LangDA, we quantitatively compare LangDA with existing methods on synthetic-to-real adaptation benchmark, Synthia → Cityscapes. We evaluate our method using the standard semantic segmentation metric, mean Jaccard Index (also termed mean Intersection over Union, or, mIoU). We report the results in Table 4.1, where it is evident that LangDA significantly outperforms the existing state-of-the-art by 2.6%. Further, we can see that LangDA, when combined with current methods, consistently improves their performance across the board, ranging from 0.9% to 2.7%. Moreover, LangDA not only benefits transformer architectures like DAFormer [21] but also boosts performance when combined with complex approaches involving additional regularization [23] or multi-resolution processing [22]. Our results demonstrate the importance of explicitly inducing context awareness via text.

4.1.2 Per-class IoU on Synthia → Cityscapes

We provide a detailed class-wise comparison of existing methods with LangDA in Table 4.3. From Table 4.3, LangDA consistently achieves new SOTA on almost all classes. Specifically, LangDA significantly outperforms existing works on the *sidewalk* and *road* classes by 14% and 8% respectively. As seen in Figure 4.1, vision-only methods that try to build context-awareness [23] still struggle to differentiate classes with similar visual appearances (such as

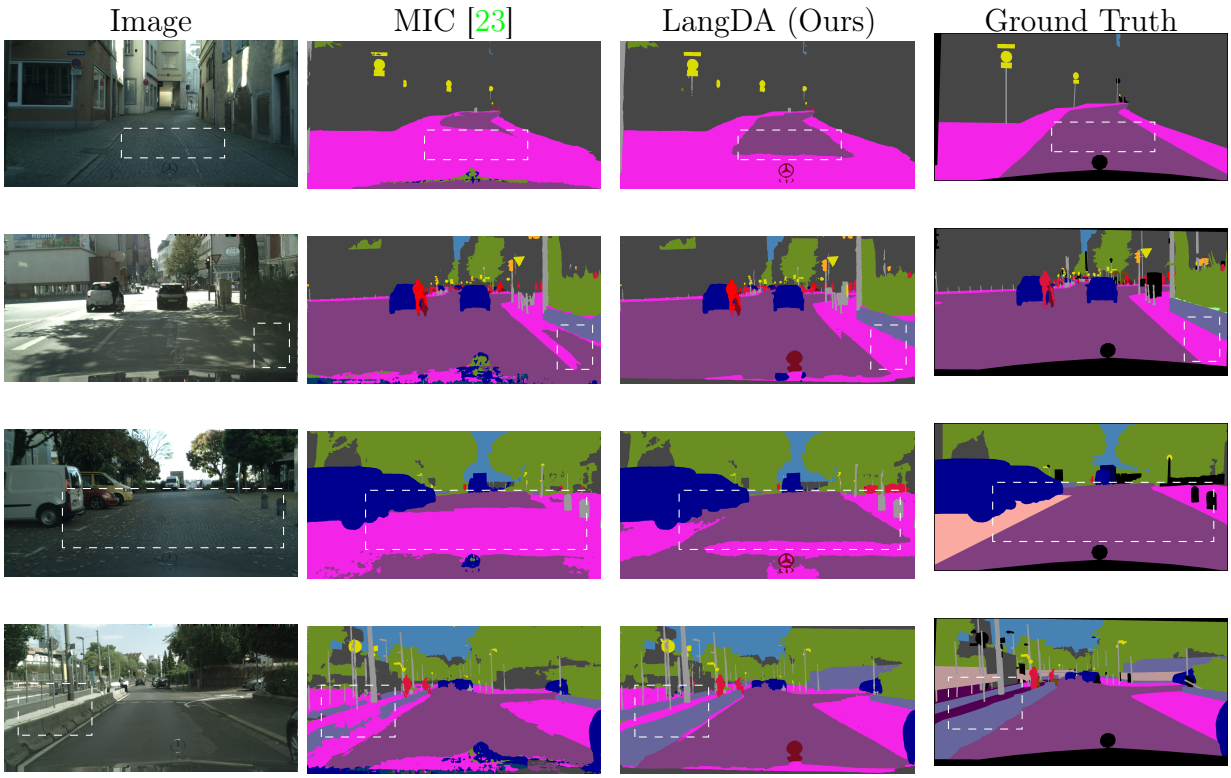


Figure 4.1: **Qualitative Results: Synthia→Cityscapes.** Existing vision-only DASS approaches face difficulty discerning visually similar classes (e.g. *road* and *sidewalk*). Our proposed method, LangDA, discerns classes with extremely similar pixels while vision-only method struggles to do so.

Table 4.1: **Comparison with state-of-the-art methods in UDA and Zero-shot DA.** We performed our experiments on standard adaptation benchmark Synthia → Cityscapes. Source only refers to lower bound DA baselines with no adaptation (i.e., training on source and evaluation on target). All methods' results are taken from the published paper except for those labeled with †, which indicates the method was reproduced. Our method, when plugged into 3 existing UDA frameworks, attains state-of-the-art performance.

Method	**Backbone**	**Unlabeled Target Data**	**Text Prompts**	**mIoU ↑**
Source only	ResNet-50			29.3
PODA† [13]	ResNet-50		✓	29.5
ULDA† [89]	ResNet-50		✓	30.8
Source only	ResNet-101			29.4
ADVENT [82]	ResNet-101	✓		41.2
CBST [96]	ResNet-101	✓		42.6
DACS [79]	ResNet-101	✓		48.3
CorDA [84]	ResNet-101	✓		55.0
ProDA [93]	ResNet-101	✓		55.5
DAFormer† [21]	MiT-B5	✓		61.1
LangDA (Ours) + DAFormer	MiT-B5	✓	✓	**62.0**
HRDA [22]	MiT-B5	✓		65.8
LangDA (Ours) + HRDA	MiT-B5	✓	✓	**66.3**
MIC [23]	MiT-B5	✓		67.3
CoPT [53]	MiT-B5	✓	✓	67.4
LangDA (Ours) + MIC	MiT-B5	✓	✓	**70.0**

road and *sidewalk*). With the proposed image-level alignment module, the image features are aligned with textual features of the entire scene, facilitating contextual understanding of the difference between classes via language. These results reinforce the benefit of proposed image-level alignment module and building context through language.

Additionally, LangDA outperforms existing methods in classes that are more reliant on context from 0.5% to 2.87%. For instance, the class *rider* usually appears alongside bicycles, *fence* is almost always in front of buildings, and *pole* is often beside the road [23]. This result provides evidence that visual-only contextual alignment is insufficient in bridging domain gaps. Our proposed LangDA method captures additional contextual cues embedded

Table 4.2: Performance under different weather conditions on Clear-to-Adverse-Weather benchmark: Cityscapes→ACDC

Method	Rain	Snow	Fog	Night	**All**
MIC [23]	**74.2**	60.7	56.8	**57.6**	63.3
CoPT [53]	74.0	60.6	60.4	56.8	63.7
LangDA (Ours)	73.3	**68.3**	**68.9**	55.7	**67.6**

in the latent prior of VLMs. Moreover, our method occasionally benefits classes such as *building* and *vegetation* where context might not play a big role [23]. This showcases the significance of language world priors in our representations in improving such classes. See Section 4.2.4 for additional t-SNE visualization.

4.1.3 Per-class IoU comparison on other two settings

To evaluate LangDA's generalizability, we test it on two challenging scenarios: normal-to-adverse weather (Cityscapes → ACDC) in Table 4.5 and day-to-night (Cityscapes → DarkZurich) in Table 4.4. LangDA outperforms existing methods by 3.9% and 1.4% mIoU, respectively, setting a new state-of-the-art. However, similar to CoPT and MIC, LangDA struggles with the *sky* and *vegetation* classes, where the model sometimes misclassifies the sky as trees for each other. This is likely due to improved segmentation from the source-trained domain, where trees frequently obscure the sky in the Cityscapes dataset. Additionally, ground truth labels for sky and trees are likely unreliable in DarkZurich due to the inherent difficulty for humans to distinguish trees from sky in low-light conditions. Despite this, LangDA consistently attain top one or two performances across most classes in both datasets. Notably, on CS → DZ, it achieves substantial improvements of 8.16% and 8.74% mIoU on the *bus* and *motorbike* classes, respectively, outperforming CoPT and MIC. This highlights the importance of context relations, especially for smaller classes that occupy less of the image, where understanding contextual positioning is crucial.

For CS → ACDC, LangDA outperforms existing SOTA by 3.9% in mIoU as seen in Table 4.2. LangDA substantially outperforms MIC and CoPT on the fog and snow conditions (8.5% and 7.5% respectively) while maintaining a similar performance on the other two conditions. This is mainly attributed to the fact that fog and snow introduce uniform

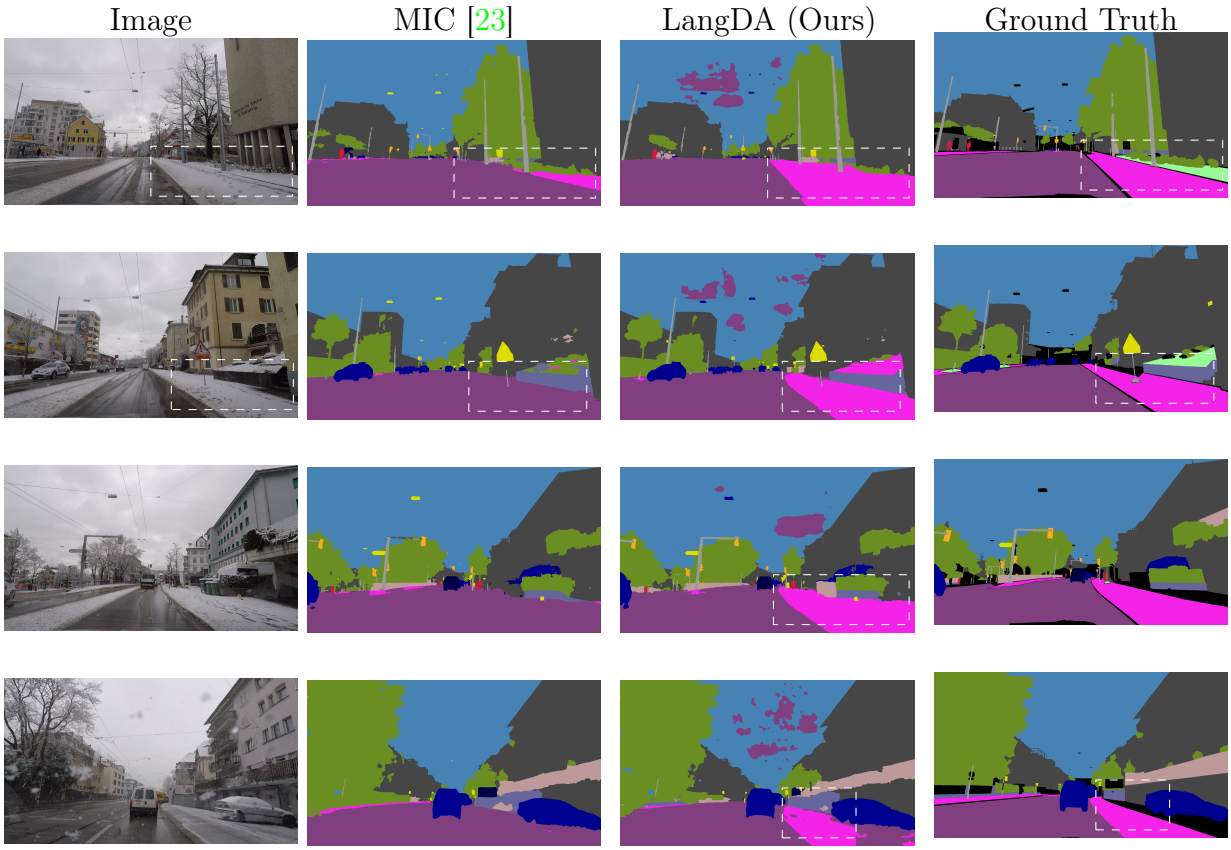


Figure 4.2: **Qualitative Results: CityScapes→ACDC.** Existing methods struggle to segment sidewalks when adapting from normal to adverse weather, primarily due to visual changes (e.g., sidewalks are covered in snow during winter). By incorporating the semantic meaning of sidewalks in language, LangDA achieves accurate segmentation even in adverse weather conditions.

Table 4.3: Per-class performance on Synthetic-to-Real adaptation benchmark: Synthia to Cityscapes. **Bold** indicates the best score. Underline indicates second-best score. All methods' results are taken from publications unless otherwise indicated.

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Sky	Person	Rider	Car	Bus	M.bike	Bike	**% mIoU ↑**
ADVENT [82]	85.6	42.2	79.7	8.7	0.4	25.9	5.4	8.1	80.4	84.1	57.9	23.8	73.3	36.4	14.2	33.0	41.2
DACS [79]	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
ProDA [93]	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
DAFormer [21]	86.2	42.3	88.2	38.4	8.6	49.9	55.6	54.1	86.9	89.3	73.4	47.1	87.8	57.3	53.1	60.2	61.1
HRDA [22]	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
MIC [22]	86.6	50.5	89.3	47.9	7.8	59.4	66.7	63.4	87.1	94.6	81.0	58.9	90.1	61.9	67.1	64.3	67.3
CoPT [53]	83.4	44.3	90.0	50.4	8.0	60.0	67.0	63.0	87.5	**94.8**	**81.1**	58.6	89.7	66.5	**68.9**	**65.0**	67.4
LangDA (Ours)	**92.0**	**58.6**	**90.8**	**57.3**	**9.7**	**62.9**	**69.9**	**64.2**	**88.3**	94.4	80.6	**59.5**	**91.2**	**70.6**	66.4	63.8	**70.0**

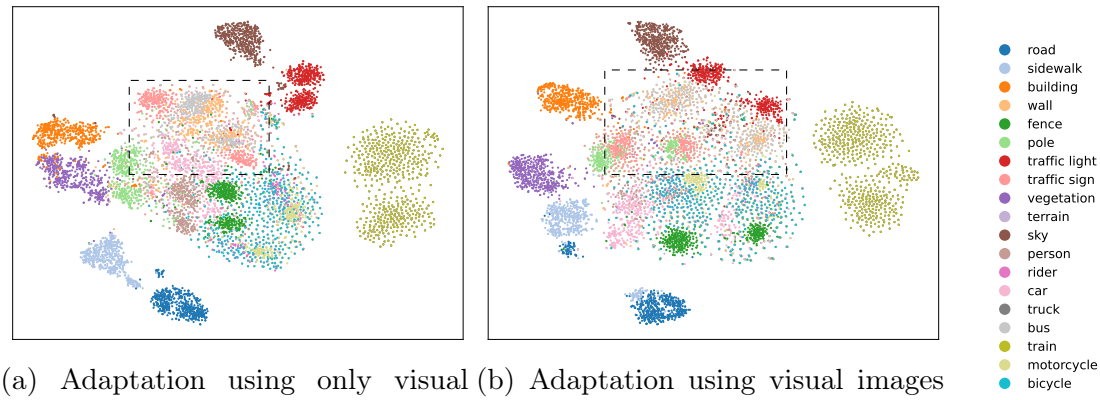
Table 4.4: Per-class performance on Day-to-Night: Cityscapes→DarkZurich. **Bold** indicates the best score. Underline indicates second-best score. All methods' results are taken from publications except for those labeled with †, which indicates the method was retrained.

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	**% mIoU ↑**
ADVENT [82]	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
MGCDA [66]	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
DANNet [87]	90.0	54.0	74.8	41.0	21.1	25.0	26.8	30.2	**72.0**	26.2	**84.0**	47.0	33.9	68.2	19.0	0.3	66.4	38.3	23.6	44.3
DAFormer [21]	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 [22]	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
MIC† [23]	86.2	57.9	**81.0**	51.6	**21.4**	61.2	23.7	**55.1**	57.3	42.0	59.0	62.2	55.4	65.2	78.6	5.22	90.0	53.4	42.3	55.2
CoPT [53]	**92.7**	66.1	80.0	49.0	19.3	**63.2**	**51.7**	52.0	50.9	43.1	55.9	61.7	56.5	59.6	**79.4**	2.96	90.8	50.0	42.5	56.2
LangDA (Ours)	92.0	**68.4**	79.6	**53.4**	19.9	61.7	32.1	54.7	44.7	**44.0**	50.9	**62.7**	**60.0**	**84.1**	78.2	**15.4**	**92.1**	**58.7**	**43.8**	**57.6**

Table 4.5: Per-class performance on Clear-to-Adverse-Weather: Cityscapes→ACDC. **Bold** indicates the best score. Underline indicates second-best score. All methods' results are taken from publications unless otherwise indicated.

Method	Road	S.walk	Build.	Wall	Fence	Pole	Tr.Light	Sign	Veget.	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	M.bike	Bike	% mIoU ↑
ADVENT [82]	72.9	14.3	40.5	16.6	21.2	9.3	17.4	21.2	63.8	23.8	18.3	32.6	19.5	69.5	36.2	34.5	46.2	26.9	36.1	32.7
MGCDA [66]	73.4	28.7	69.9	19.3	26.3	36.8	**53.0**	53.3	**75.4**	32.0	84.6	51.0	26.1	77.6	43.2	45.9	53.9	32.7	41.5	48.7
DANNet [87]	84.3	54.2	77.6	38.0	30.0	18.9	41.6	35.2	71.3	39.4	**86.6**	48.7	29.2	76.2	41.6	43.0	58.6	32.6	43.9	50.0
DAFormer [21]	58.4	51.3	84.0	42.7	35.1	50.7	30.0	57.0	74.8	52.8	51.3	58.3	32.6	82.7	58.3	54.9	82.4	44.1	50.7	55.4
CoPT [53]	49.1	**70.3**	83.6	**59.4**	**42.4**	**58.5**	48.3	**67.2**	73.5	**60.7**	45.0	69.3	45.2	83.4	**76.3**	74.5	**88.2**	54.4	61.4	63.7
LangDA (Ours)	**84.7**	63.0	**88.2**	56.4	40.9	57.5	42.9	64.7	**75.4**	59.4	82.8	**69.8**	**46.6**	**89.1**	74.8	**83.2**	**88.2**	**55.9**	**61.8**	**67.6**

domain shifts primarily in terms of texture and visibility, rather than drastic and uneven changes in lighting or color. However, a setting like night introduces uneven sources of lighting (streetlights) when compared to a daytime urban dataset like Cityscapes. Since CoPT uses domain knowledge of the target data, they have slightly better performance in such settings. This shows that our domain-agnostic objective may be more useful for uniform domain shifts.



(a) Adaptation using only visual images[21].

(b) Adaptation using visual images and contextual language descriptions.

Figure 4.3: **t-SNE of DAFormer (Left) and LangDA (Ours, Right)** After aligning language and visual features, we observe more well-defined boundaries and improved class clustering.

4.2 Qualitative Results

4.2.1 Qualitative Results: Cityscapes

Figure 4.1 presents the qualitative results of LangDA in comparison to existing vision-only DASS methods. LangDA demonstrates superior performance, particularly in distinguishing visually similar classes, such as road and sidewalk, by leveraging contextual language scene descriptions. This is evident in the areas highlighted by the dashed white box.

4.2.2 Qualitative Results: ACDC and DarkZurich

Figure 4.2 and Figure 4.4 depict qualitative comparisons of LangDA with existing SOTA methods on CS→ACDC and CS→DarkZurich respectively. Specifically, for snowy scenes in ACDC, existing works [23] are less robust and cannot distinguish the snow-covered *sidewalks* from *road*, whereas LangDA can accurately segment *sidewalks* even when the *sidewalks* are covered by snow. This showcases the significance of language-guided context relations in comparison to implicitly learned visual context relations.

Furthermore, for DarkZurich in day-to-night adaptation, current works fail in extremely dark scenarios with uneven sources of lighting and misclassify dark background pixels as *cars*. This results in a significant number of false positives for the *car* class. However, LangDA

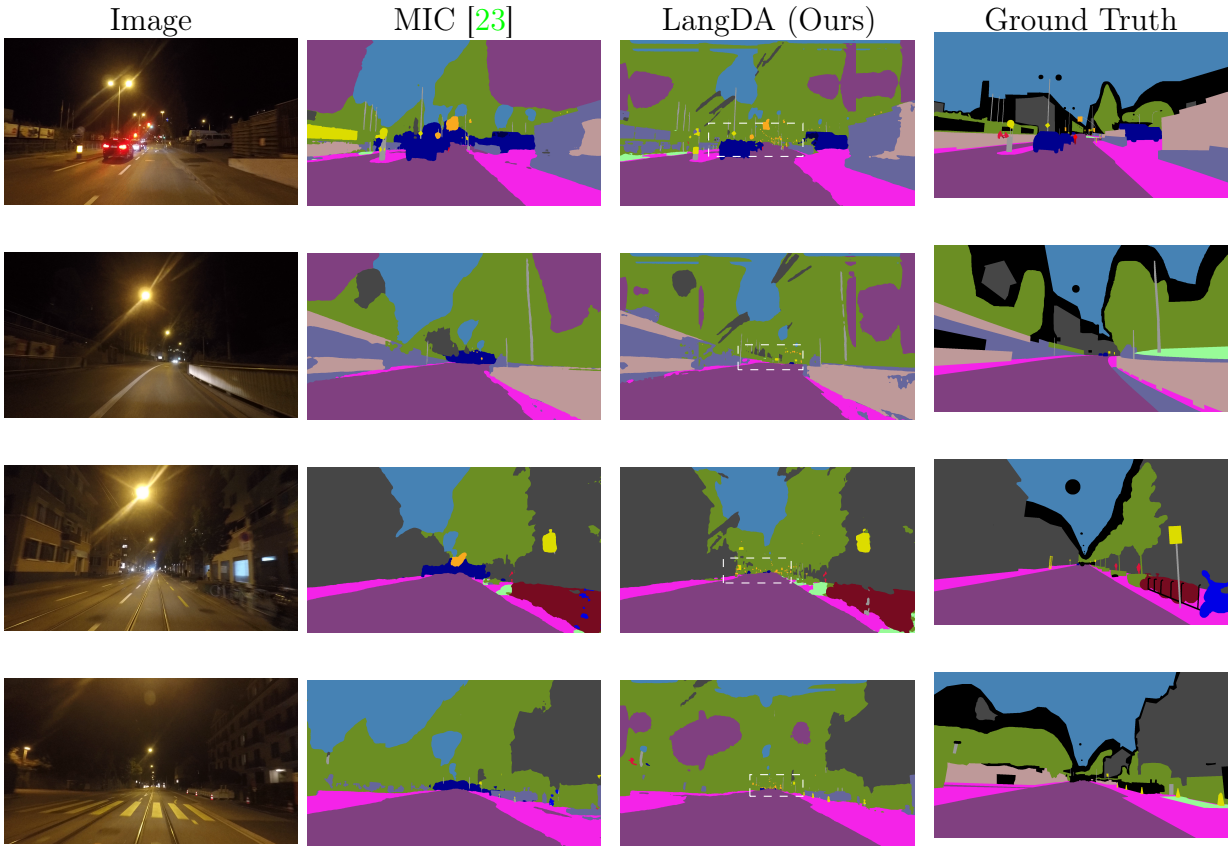
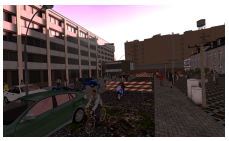


Figure 4.4: **Qualitative Results: CityScapes→DarkZurich.** Prior methods often mistaken dark pixels (row 3) and reflected lights (row 4) as cars due to the challenge of visually differentiating car lights from streetlights in low-light conditions. By leveraging additional contextual information through language, the proposed LangDA method accurately segments cars even in the dark.

Table 4.6: Qualitative Results: Captions Generated by Context-Aware Generator for Synthia [63]. Each row shows an image (left), its raw VLM-generated caption (middle), and the LLM-refined caption (right). The refined captions are more concise while preserving key spatial and semantic information.

Image	VLM Generated Description	LLM Refined Caption
	The image depicts a busy city street with a large number of people walking and riding bicycles. There are several cars and a motorcycle on the street, and a bicycle is also visible. The street is lined with buildings, and there is a fence nearby. There are numerous people in the scene, some walking and others riding bicycles. A person on a bicycle is located near the center of the image, while another person on a bicycle is closer to the right side. A person walking is also present in the scene, near the center. In addition to the people, there are several cars and a motorcycle on the street. A car is located on the left side of the image, while another car is further back on the right side. A motorcycle is parked on the left side of the street. The scene also includes a few objects such as a bench, a traffic light, and a handbag. A bench is located on the right side of the street, while a traffic light is situated near the center. A handbag is placed on the ground, close to the center of the image.	The image shows a city street with people, bicycles, cars, and a motorcycle. People are near the center and right, with bicycles. Cars are on the left and right, with a motorcycle parked left. Buildings line the street, and a fence is nearby. Also, a bench is right, a traffic light center, and a handbag near the center. Classes and locations: person (center, right), bicycle (center, right), car (left, right), motorcycle (left), building (street edge), fence (nearby), bench (right), traffic light (center), handbag (center).
	The image depicts a busy city street with a large crowd of people walking on the sidewalk. There are several cars and a motorcycle parked or driving along the street. A bicycle can also be seen among the crowd. The scene is set in a town square with a mix of buildings, including a tall building in the background. There are numerous people in the scene, with some walking closer to the camera and others further away. A few individuals are carrying handbags, and a person is holding a cell phone. The street is lined with a fence, and there are multiple traffic lights visible in the area. The sky is visible in the background, and the overall atmosphere of the scene is bustling and lively.	City street with people on sidewalk. Cars, motorcycle, bicycle on road. Buildings, fence, pole, sky visible.

can accurately segment dark pixels and *cars* even in low-light settings, demonstrating the significance of context-aware language representations.

4.2.3 Qualitative Results: Captions

Table 4.6 and Table 4.7 present example captions for Synthia [63] and Cityscapes [11], respectively. The VLM-generated captions provide detailed and accurate descriptions, while the LLM-refined versions offer concise summaries that retain essential scene understanding.

In Table 4.6, both captions in Row 1 correctly identify key objects such as the car on the left and pedestrians near the center and right. While some hallucinations (e.g., mention of a handbag) appear in the raw captions, they have limited effect on segmentation since the model relies primarily on vision, with text used only as auxiliary supervision in a closed-set setting (i.e. unknown classes are not segmented).

In Table 4.7, captions for real-world images are more grounded. For example, Row 1 correctly identifies a bus on the road and parked bicycles, while Row 2 describes the walking person, motorbikes, and a white van on the left, demonstrating accurate spatial and semantic understanding.

4.2.4 t-SNE Visualization

t-SNE [80] is a non-linear dimensionality reduction method that embeds high-dimensional data into a low-dimensional space by minimizing the Kullback–Leibler divergence between pairwise similarity distributions. t-SNE is widely used to visualize high-dimensional data.

Figure 4.3 shows the t-SNE visualizations of feature distributions of vision-only method DAFormer and LangDA built on top of it. After integrating language-driven feature alignment, our method LangDA shows improved per-class clustering. For instance, in DAFormer's [21] t-SNE, the feature representations of walls (light orange) and traffic signs (rose pink) overlap in the image domain, likely due to traffic signs often visually appear in front of walls from driver's first-person view. On the other hand, walls and traffic signs are semantically distinguishable and do not appear together in language contextual descriptions, contributing to improved mIoU segmentation for LangDA in Table 4.3. These findings demonstrate the importance of inducing contextual relationships in language and highlight the advantages of LangDA for segmenting classes in UDA settings.

Table 4.7: Qualitative Results: Captions Generated by Context-Aware Generator for Cityscapes [11]. Each row shows an image (left), its raw VLM-generated caption (middle), and the LLM-refined caption (right).

Image	VLM Generated Description	LLM Refined Caption
	The image depicts a busy city street with a red bus driving down the road, surrounded by various vehicles and pedestrians. There are multiple cars on the street, including a white car in front of the bus and a few others scattered throughout the scene. A bicycle can also be seen parked on the sidewalk. There are several traffic lights and traffic signs visible in the image, with one traffic light located near the left side of the street and another towards the right side. A person is walking on the sidewalk, and a few more pedestrians can be seen in the background. The street is lined with buildings, and a few parked cars are visible on the side of the road. The overall scene is bustling with activity, showcasing the typical hustle and bustle of a city street.	Bus on road, surrounded by cars. Bicycle parked on sidewalk. Traffic lights on left and right. Person walking on sidewalk. Buildings line the street.
	The image depicts a city street with a man walking down the sidewalk, and a white van parked on the side of the road. There are several cars parked along the street, including one on the left side, one in the middle, and two on the right side. A motorcycle is also parked on the sidewalk near the man. There are multiple traffic signs in the scene, with one located near the center of the image, another on the left side, and a third one on the right side. A fire hydrant can be seen on the sidewalk near the center of the image. The man is walking on the sidewalk, and there is a person standing near the middle of the scene. The street appears to be a mix of brick and cobblestone, giving it a charming atmosphere.	A city street scene with a person walking on the sidewalk and another standing nearby. A white van and several cars are parked on the road: left, middle, and right sides. A motorcycle is parked on the sidewalk. Traffic signs are located at the center, left, and right.

4.2.5 Confusion Matrix

To further investigate the degradation in minority classes during day-to-night adaptation, we examine sky and vegetation as representative classes and construct a confusion matrix to investigate pixel-level misclassifications.

As shown in Fig. 4.5, LangDA demonstrates strong contextual awareness, with significant more true positives correctly predicted along the diagonal compared to vision-only methods. However, we observe a notable confusion between vegetation (trees) and sky in low-light conditions. This suggests that while the model effectively learns spatial relationships—such as vegetation appearing adjacent to the sky—it struggles to distinguish between them in darkness, leading to misclassifications.

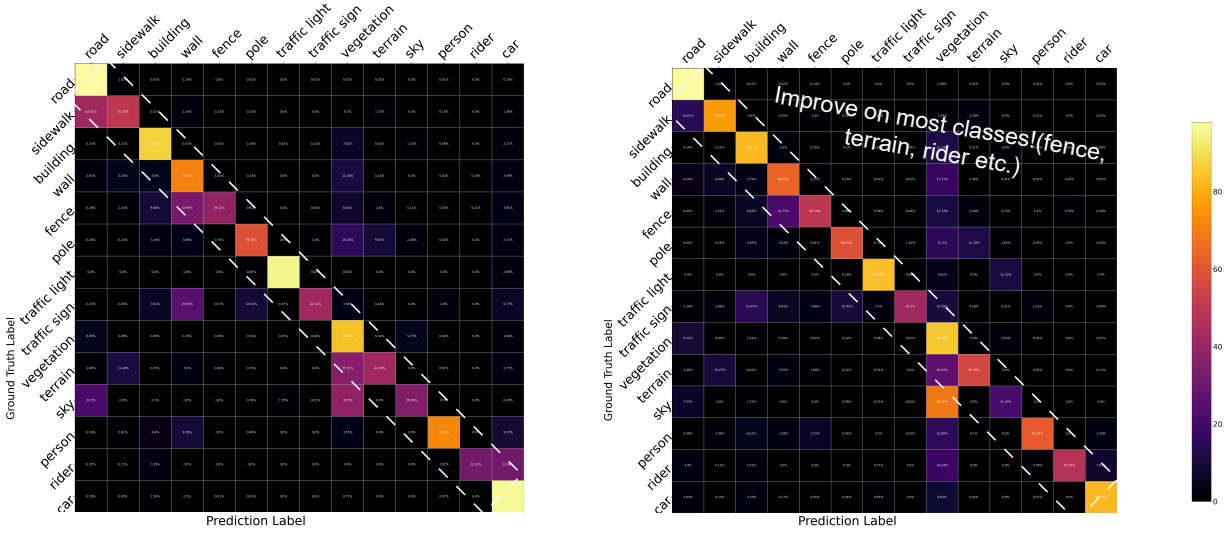


Figure 4.5: Confusion Matrix: MIC [23] (left) and LangDA (Ours, right). We can see the method proposed in our thesis has higher true positive rate, indicated by lighter colored diagonal values. The confusion matrix is computed on the validation set rather than the test set, as the ground truth for the test set is unavailable– test set mIoU is obtained by submitting predictions to a privately hosted server [65].

Table 4.8: Effect of prompting and aligning techniques on Synthetic-to-Real adaptation benchmark: Synthia → Cityscapes.

	Caption Generation		Alignment Technique		% mIoU↑
	Context-aware Captions	Class-level Prompts	Image-level Alignment	Pixel-level Alignment	
1	✓	✗	✓	✗	**70.0**
2	✗	✓	✓	✗	68.9
3	✗	✓	✗	✓	68.7
4	✗	✗	✗	✗	67.3

4.3 Ablation Studies

4.3.1 Effect of Prompting and Alignment Strategies.

To evaluate the impact of the proposed modules, we replace context-aware caption generator with generic prompts and swap image-level alignment with pixel-level alignment in Tab. 4.8. To assess the benefit of context-aware caption generator (row 2), we replace our contextual caption with text embedding of generic class-level prompts aligned to the entire image. This results in a 1.1% drop in mIoU, demonstrating the effectiveness of the caption generator. To evaluate image-level alignment (row 3), we align generic class prompts (e.g. "A photo of a {class}") to the pixel features, leading to a 1.3% mIoU reduction, showing the effectiveness of image-level alignment. While the gain from image-level alignment alone may seem modest (row 2 and 3), combining image-level alignment with context-aware caption generator (row 1 and 4) produces a substantial boost, demonstrating the exceptional synergy of our proposed modules.

4.3.2 Comparison of Source and Target Captions.

In our main experiments, we use image-level captions on source data for text-based supervision. Table 4.9 examines whether applying our objective to target image captions can provide further improvement. For unsupervised target captions, we omit ground-truth masks, as these are considered unavailable. Instead, we prompt the LLM to summarize the scene description without further class name refinement, while the VLM receives text query not containing class names. The table shows that applying LangDA on either source or target captions achieves state-of-the-art (SOTA) performance on Synthia → CS, supporting our hypothesis that *explicit language-induced context is advantageous*. The slight performance drop for target-only supervision may be attributed to the hallucinations of VLMs and

Table 4.9: Ablation: contextual caption applied on source, target, and source + target

Method	Image Captions	% mIoU↑
MIC [23]	✗	67.3
CoPT [53]	Source only	67.4
LangDA (Ours)	Source only	**70.0**
LangDA (Ours)	Target only	69.1
LangDA (Ours)	Source + Target	68.0

LLMs without class name refinement. Furthermore, similar to CoPT [53], when applying our objective to both source and target data, we backpropagated once after applying it only on source data to avoid memory issues, which alters the optimization process and may lead to lower performance.

4.3.3 Ablations on CLIP Text Encoders

To study LangDA's dependence on text encoder, we generate text embeddings by passing our refined context-aware captions through different text encoders [59, 27] and retrain LangDA using each of these text embeddings individually for Synthia → Cityscapes (CS). The results are reported in Table 4.10. From the table, it is evident that LangDA outperforms MIC [23] (67.3%) and CoPT [53] (67.4%) irrespective of the chosen text encoder, highlighting the significance of the proposed context-aware caption generator.

4.3.4 Hyperparameter Sensitivity

LangDA exhibits low sensitivity to λ_p, weight of the language consistency loss, eliminating the need for extensive hyperparameter tuning. As shown in Table 4.11, performance remains strong as long as λ_p remains within a reasonable range and is weighted less than the combined unsupervised and supervised losses. LangDA's robustness to hyperparameter changes makes it simple and practical to deploy.

4.3.5 Ablation on Caption Refinement Module.

Table 4.12 compares direct truncation of VLM-generated captions to LLM-based refinement. We observe improved segmentation mIoU with LLM refinement–an encouraging result in dense prediction tasks, where many recent methods achieve less than a 1% overall gains. However, the improvement is smaller than expected.

A potential explanation lies in the limited effective context length of the CLIP text encoder. Although CLIP supports 77 tokens, recent work suggests its effective context length is closer to 20 tokens for image retrieval tasks using detailed captions (Figure 4.6 [91]). As a result, the refined and unrefined captions may appear similarly truncated to CLIP at merely 20 tokens, reducing the impact of refinement. We hypothesize that using text encoders with longer context windows may further enhance the benefits of caption refinement.

4.4 Limitations

While the proposed method achieves state-of-the-art performance in the field of DASS, with mIoU scores of 70.0%, 57.6%, and 67.6% on synthetic-to-real, day-to-night, and normal-to-adverse weather benchmarks respectively, these results still fall short of the 85% mIoU typically achieved under fully supervised settings. This gap reflects a fundamental limitation of unsupervised methods: the lack of ground truth labels leads to reliance on pseudo-labels, which are inherently noisy and often biased toward the source domain. Although LangDA already promotes more generalized feature representations via contexts in language, pseudo-label quality remains a bottleneck. In the normal-to-adverse weather

Table 4.10: Ablation: CLIP-based text encoders.

Method	Text Encoder	% mIoU↑
MIC [23]	✗	67.3
CoPT [53]	CLIP	67.4
LangDA (Ours)	OpenCLIP	68.9
LangDA (Ours)	LongCLIP	69.0
LangDA (Ours)	CLIP	**70.0**

Table 4.11: λ_p

λ_p	% mIoU ↑
2	67.9
1	69.6
0.1	**70.0**
0.01	69.7
✗	67.3

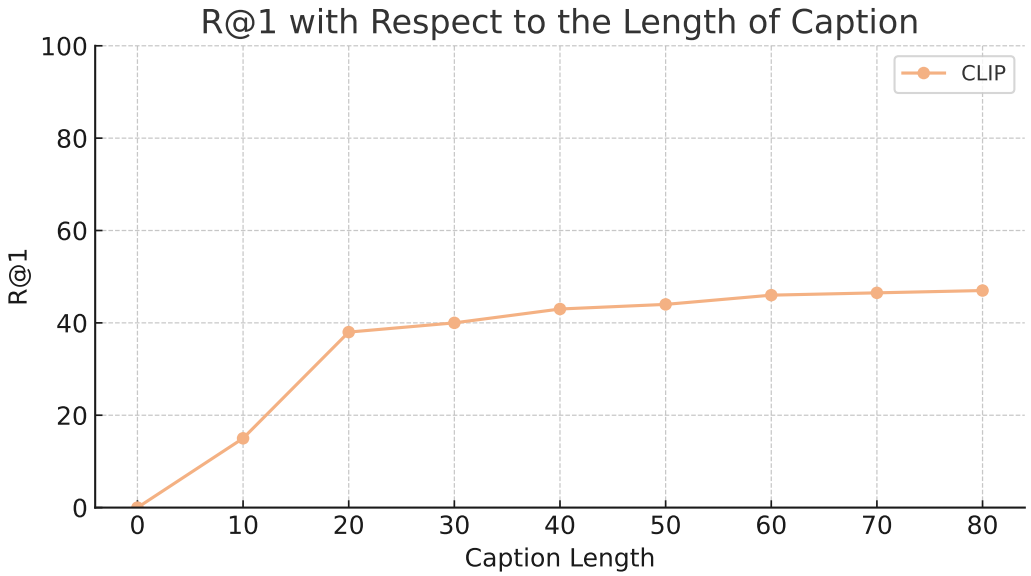


Figure 4.6: **Short context length of CLIP**. Image retrieval accuracy with respect to caption length, illustrating that CLIP's effective context length is significantly shorter than its nominal limit of 77 tokens. Figure is generated using data from [91].

setting, all class-level mIoUs exceed 40%—indicating that LangDA produces generalized and consistent pseudo-labels across classes (Table 4.5). However, overall performance remains limited by pseudo-label accuracy, suggesting that all classes could benefit uniformly from improved pseudo-label refinement. Future work may explore strategies such as adaptive confidence thresholds or external guidance to enhance pseudo-label quality.

Additionally, under-performing classes tend to be small or infrequent in the dataset. While LangDA improves segmentation for smaller classes overall, we observe that some categories still perform poorly: *sidewalk* and *fence* in synthetic-to-real (Table 4.3), and *bus*, *traffic light*, and *fence* in day-to-night (Table 4.4). These issues are due to the long-tailed class frequency distribution (Figure 4.7), which is not explicitly addressed in this work. Future methods should consider incorporating domain adaptation techniques tailored to imbalanced class distributions [58, 18, 28].

Table 4.12: Ablation on Language Refinement on Synthia → Cityscapes.

Caption Generation	mIoU ↑
VLM + LLM	**70.0**
VLM	69.4

4.5 Summary

This chapter presents quantitative and qualitative results demonstrating the effectiveness of the proposed LangDA framework across three standard domain adaptation benchmarks: Synthia → Cityscapes (synthetic-to-real), Cityscapes → ACDC (normal-to-adverse weather), and Cityscapes → DarkZurich (day-to-night). LangDA consistently outperforms prior vision-only and language-guided DASS methods, achieving state-of-the-art mIoU of 70.0%, 67.6%, and 57.6%, respectively (Section 4.1). Notably, LangDA yields significant improvements in context-dependent and underrepresented classes (e.g., *sidewalk*, *bus*, *motorbike*) and provides robust gains under diverse weather and lighting conditions, as evidenced by both per-class IoU (Tables 4.3 to 4.5) and qualitative visualizations (Figures 4.1, 4.2 and 4.4).

Extensive ablation studies confirm that LangDA's core components–context-aware caption generation and image-level alignment–contribute complementary gains. The method remains robust across different CLIP-based text encoders (Table 4.10) and is insensitive to hyperparameter variations Table 4.11. Further, qualitative results (Figures 4.1, 4.2 and 4.4) and t-SNE visualizations (Figure 4.3) show enhanced class separability and improved segmentation boundaries. While noisy pseudo-label and long-tailed class distributions remain limitations, LangDA provides a scalable and efficient framework that advances the state of the art in unsupervised domain adaptation for semantic segmentation.

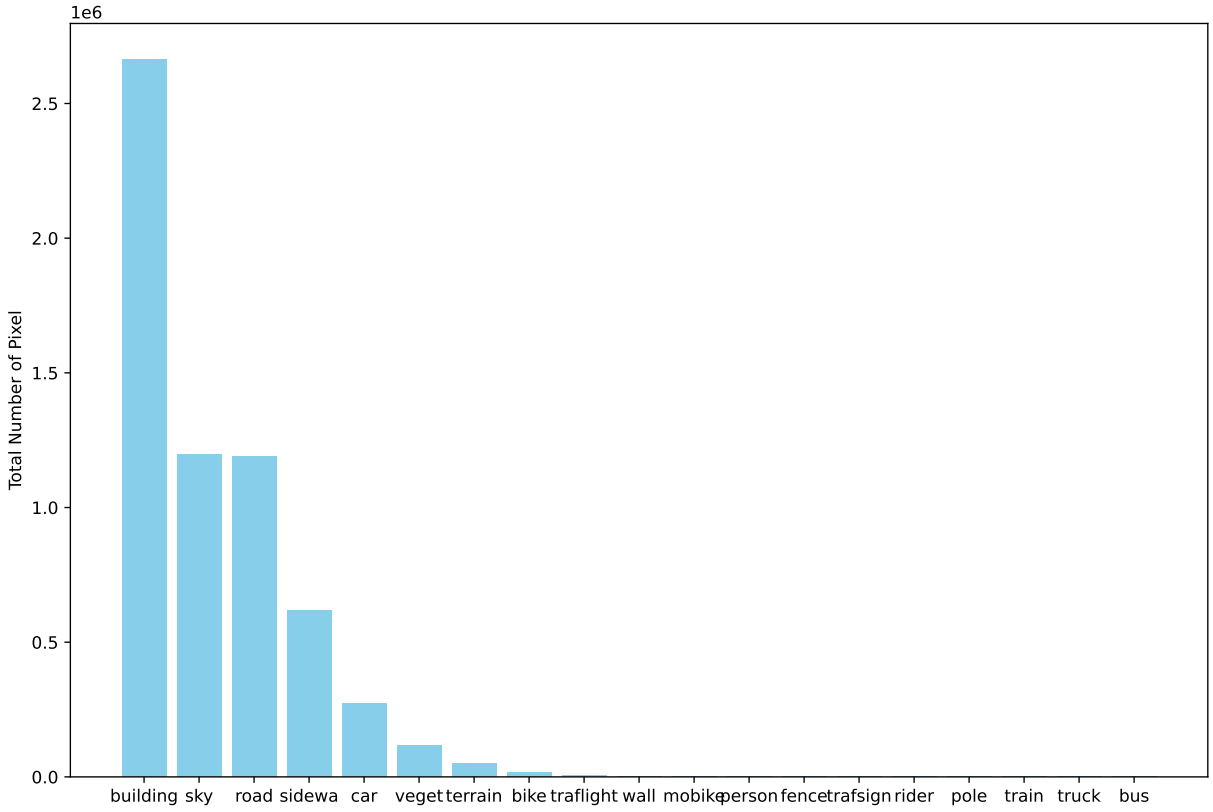


Figure 4.7: Long-tailed class distribution in DarkZurich [65]. A few frequent classes (e.g., road, building) dominate the pixel count, while many classes (e.g., bus, pole, traffic light) are underrepresented, posing challenges for balanced domain adaptation.

Chapter 5

Conclusion

In this paper, we introduce LangDA, the first work to explicitly induce spatial relationships through language to learn rich context relations and generalized representations for DASS. LangDA achieves this by developing a) a context-aware caption generator that expresses spatial context in images using language, and b) an image-level alignment module that promotes context awareness via alignment between the scene caption and the entire image. Unlike prior methods relying on pixel-level alignment, LangDA's image-level consistency aligns with the entire image, promoting a holistic understanding of class relationships. Notably, LangDA is also the first language-guided DASS method that requires no supervisory text for describing the target domain (e.g. "a {daytime} photo" to "a {snowy} photo"). As shown in extensive experiments and ablations, LangDA establishes a new SOTA on three key DASS benchmarks, achieving significant performance improvements over existing methods and consistently improving performance when integrated with other UDA methods. In doing so, LangDA establishes that extracting key context relationships from language is a promising avenue for DASS. Moreover, it also motivates future work in automatically tailoring captions to draw critical information for a particular task.

5.1 Summary of Contributions

This thesis analyzes ways of incorporating language to address visual domain shift. It is the first work to directly encode rich contextual object relationships in an image through text for DASS. Unlike prior methods that align individual class embeddings (e.g., text embedding of "car" with its pixel embedding), LangDA aligns a descriptive scene embedding

with the corresponding image embedding (image-level alignment), capturing higher-order object relationships.

Moreover, the proposed method eliminates the need for manually created text prompts by automatically generating spatially aware scene descriptions using a frozen Vision-Language Model (VLM) and Large Language Model (LLM). Manual prompt generation is often inconsistent and difficult to reproduce, especially for complex scenes (e.g., snow-covered roads), where slight variations in phrasing can affect results. By using a seeded VLM, LangDA ensures standardized captions, improving reproducibility for benchmarking.

Furthermore, existing methods rely on textual domain gap supervision (e.g., aligning "a {daytime} photo of {class}" with "a {nighttime} photo of {class}"), which fails when domain shifts are unpredictable, such as tumor scans with varying color and location distributions across hospitals [24, 41, 5]. LangDA overcomes this limitation by employing a context-aware caption generator that automatically creates image-specific captions, making it effective even in cases where the domain gap is unknown.

In addition, this work removes the need for manual prompt tuning [26, 89, 13, 53] and human feedback citegong2024coda, which are often inconsistent and resource-intensive. This standardizes captioning for benchmarking, eliminating variability introduced by manual tuning and ensuring more reliable evaluations.

To further advance language-guided semantic segmentation research, this thesis introduces the first captioning dataset for Synthia and Cityscapes, providing a valuable resource for future studies, whether in the presence or absence of domain shifts.

Another contribution is that the proposed method is designed for efficient deployment, requiring only the target-trained model at inference. Unlike many language-guided approaches, it does not require captions or language input at test time – text serves purely as a learning aid during training. Despite its lightweight nature, LangDA significantly improves segmentation performance, particularly for challenging classes such as person, road, and sidewalk, without increasing model complexity.

Overall, the proposed method achieves 70.0%, 67.6%, and 57.6% mean Jaccard index on Synthia $\rightarrow$ Cityscapes, Cityscapes $\rightarrow$ ACDC, and Cityscapes $\rightarrow$ DarkZurich, surpassing state-of-the-art methods by 2.6%, 3.9%, and 1.4%, respectively.

5.2 Limitations

Although extensive ablation studies have been conducted, several areas remain open for further exploration.

One key aspect is the choice of image encoder. In this work, we trained our own image encoder to enhance segmentation performance for dense tasks. However, an alternative approach could involve using a frozen CLIP image encoder with a trained decoder. A comparative study of these two strategies could provide deeper insights into the trade-offs between leveraging a shared vision-language space with a frozen CLIP encoder and optimizing for visual segmentation tasks with a trained encoder.

Additionally, LangDA relies on CLIP and other vision–language models (VLMs) to provide generalized representations that facilitate adaptation. The effectiveness of our language-guided approach depends on the ability of CLIP embeddings to capture relevant contextual information. However, recent studies highlight several limitations of CLIP. For instance, its effective context length is often shorter than the nominal maximum [91], and it can struggle to differentiate spatial relations such as left versus right in images. CLIP is also susceptible to spurious correlations–e.g., associating objects with typical backgrounds–leading to reduced robustness under distribution shifts [85]. Moreover, it under-performs in multi-object reasoning, often overemphasizing the first or largest object mentioned in the prompt, which compromises scene understanding in complex contexts [1].

If CLIP fails to encode relevant semantics for a domain-specific dataset, our proposed method will be less effective. In such cases, future work should explore extracting richer embeddings from alternative VLM models (e.g. BLIP [43], ALIGN [29], GPT-4 [2], LLaMA [78], Gemini [77], Mistral [30]) to improve adaptation in scenarios where CLIP is insufficient.

Furthermore, while per-class results indicate improvements for the majority of classes, performance degradation is observed for specific classes in certain datasets, such as ACDC and DarkZurich. Further research is necessary to understand the underlying factors contributing to these inconsistencies to enhance the overall robustness of the proposed architecture.

5.3 Future Works

First, the feature alignment technique between visual and language features could be extended beyond domain adaptation to standard semantic segmentation tasks without distribution shifts. It may also be applicable to classification and detection tasks. Investigating whether this alignment improves performance in broader vision tasks would be valuable.

Furthermore, the network architecture proposed in this thesis could be applied to zero-shot domain adaptation, where instead of available unlabeled target data for DASS, no target data is available during training. In this scenario, aligning source features to

target descriptions using our method may provide improvements, making the approach more generalizable to settings without access to unlabeled target data.

Future work may also extend this thesis by exploring more effective alignment of image and text features to better leverage multi-modal representations in language-guided DASS. Broadly, this can be approached in three ways: (1) enhancing caption quality, (2) improving multi-modal representations, and (3) refining alignment techniques. Among these, improving the underlying multi-modal representations appears more promising than focusing on caption refinement. Our observations show that captions summarized by different LLMs (e.g., ChatGPT-4o vs. Mistral) tend to converge to similar content, limiting the impact of further prompt-tuning. Moreover, since CLIP is trained for classification, small variations in captions may not significantly affect its representation. This is supported by our ablation results (Section 4.3.3), where replacing CLIP with stronger alternatives like OpenCLIP [27] or LongCLIP [91] led to larger performance gains than refining captions.

Beyond cross-modal adaptation, where language enhances vision tasks, this approach could be extended to open-vocabulary or open-set scenarios under distribution shift. We seek to explore this direction next.

Finally, by providing a captioning dataset for Synthia and Cityscapes, we encourage future work to leverage contextual scene descriptions for dense semantic segmentation tasks. Future directions include integrating referring expression segmentation, where models use text prompts to segment specific objects or regions; exploring pretraining strategies that jointly model visual and textual data for dense prediction; and designing architectures that dynamically adapt language guidance based on scene complexity. Additionally, there is potential to extend this framework to instance and panoptic segmentation, where understanding object-level context and relationships is critical.

References

[1] Reza Abbasi, Ali Nazari, Aminreza Sefid, Mohammadali Banayeeanzade, Mohammad Hossein Rohban, and Mahdieh Soleymani Baghshah. Clip under the microscope: A fine-grained analysis of multi-object representation. arXiv preprint arXiv:2502.19842, 2025.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[4] Nikita Araslanov and Stefan Roth. Self-supervised augmentation consistency for adapting semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15384–15394, 2021.

[5] Hritam Basak and Zhaozheng Yin. Quest for clone: Test-time domain adaptation for medical image segmentation by searching the closest clone in latent space. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 555–566. Springer, 2024.

[6] Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio Calmon, and Himabindu Lakkaraju. Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

[7] Lin Chen, Zhixiang Wei, Xin Jin, Huaian Chen, Miao Zheng, Kai Chen, and Yi Jin. Deliberated domain bridging for domain adaptive semantic segmentation. *Advances in Neural Information Processing Systems*, 35:15105–15118, 2022.

[8] Jian Cheng, Yang-Feng Hu, Yu Dai, Xue Qiao, Li Yao, and Jun-Yan Yang. Nighttime semantic segmentation with unsupervised learning and cross attention. In *Asian Conference on Machine Learning*. PMLR, 2023.

[9] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4123, June 2024.

[10] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020.

[11] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[12] Zhekai Du, Xinyao Li, Fengling Li, Ke Lu, Lei Zhu, and Jingjing Li. Domain-agnostic mutual prompting for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23375–23384, 2024.

[13] Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, and Raoul De Charette. Poda: Prompt-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18623–18633, 2023.

[14] Xin Feng, Youni Jiang, Xuejiao Yang, Ming Du, and Xin Li. Computer vision algorithms and hardware implementations: A survey. *Integration, the VLSI Journal*, 69:309–320, 2019.

[15] Ziyang Gong, Fuhao Li, Yupeng Deng, Deblina Bhattacharjee, Xianzheng Ma, Xiangwei Zhu, and Zhenming Ji. Coda: Instructive chain-of-domain adaptation with severity-aware visual prompt tuning. In *European Conference on Computer Vision*, pages 130–148. Springer, 2024.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. volume 27, 2014.

[17] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

[18] Xiao Gu, Yao Guo, Zeju Li, Jianing Qiu, Qi Dou, Yuxuan Liu, Benny Lo, and Guang-Zhong Yang. Tackling long-tailed category distribution under domain shifts. In *European Conference on Computer Vision*, pages 727–743. Springer, 2022.

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[20] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

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

[22] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *European conference on computer vision*, pages 372–391. Springer, 2022.

[23] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023.

[24] Zhaoyu Hu, Yuhao Sun, Liuguan Bian, Chun Luo, Junle Zhu, Jin Zhu, Shiting Li, Zheng Zhao, Yuanyuan Wang, Huidong Shi, et al. Uda-gs: A cross-center multimodal unsupervised domain adaptation framework for glioma segmentation. *Computers in Biology and Medicine*, 185:109472, 2025.

[25] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

[26] Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, and Yong Jae Lee. A sentence speaks a thousand images: Domain generalization through distilling clip with

language guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11685–11695, 2023.

[27] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[28] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7610–7619, 2020.

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[30] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[31] Sheng Jin, Xueying Jiang, Jiaxing Huang, Lewei Lu, and Shijian Lu. LLMs meet VLMs: Boost open vocabulary object detection with fine-grained descriptors. In *The Twelfth International Conference on Learning Representations*, 2024.

[32] Daehan Kim, Minseok Seo, Kwanyong Park, Inkyu Shin, Sanghyun Woo, In So Kweon, and Dong-Geol Choi. Bidirectional domain mixup for domain adaptive semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1114–1123, 2023.

[33] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[34] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022.

[35] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9579–9589, 2024.

[36] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2691–2701, January 2024.

[37] Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, and Meng Cao. Empowering unsupervised domain adaptation with large-scale pre-trained vision-language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2691–2701, 2024.

[38] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. Padclip: Pseudo-labeling with adaptive debiasing in clip for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16155–16165, October 2023.

[39] Mengcheng Lan, Xinjiang Wang, Yiping Ke, Jiaxing Xu, Litong Feng, and Wayne Zhang. Smooseg: smoothness prior for unsupervised semantic segmentation. *Advances in Neural Information Processing Systems*, 36:11353–11373, 2023.

[40] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896. Atlanta, 2013.

[41] Kyungsu Lee, Haeyun Lee, Georges El Fakhri, Jonghye Woo, and Jae Youn Hwang. Self-supervised domain adaptive segmentation of breast cancer via test-time fine-tuning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 539–550. Springer, 2023.

[42] Attila Lengyel, Sourav Garg, Michael Milford, and Jan C van Gemert. Zero-shot day-night domain adaptation with a physics prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4399–4409, 2021.

[43] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[44] Jeongkee Lim and Yusung Kim. Cross-domain semantic segmentation on inconsistent taxonomy using vlms. In *European Conference on Computer Vision*, pages 18–35. Springer, 2024.

[45] Chang Liu, Bavesh Balaji, Saad Hossain, C Thomas, Kwei-Herng Lai, Raviteja Vemulapalli, Alexander Wong, and Sirisha Rambhatla. Langda: Building context-awareness via language for domain adaptive semantic segmentation. *arXiv preprint arXiv:2503.12780*, 2025.

[46] Chang Liu, Saad Hossain, C Thomas, Kwei-Herng Lai, Raviteja Vemulapalli, Sirisha Rambhatla, and Alexander Wong. Language-guided domain adaptive semantic segmentation. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2024.

[47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[48] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Pengju An, Xiaoqi Li, Kaichen Zhou, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. *Advances in Neural Information Processing Systems*, 37:40085–40110, 2024.

[49] Xiaofeng Liu, Chaehwa Yoo, Fangxu Xing, Hyejin Oh, Georges El Fakhri, Je-Won Kang, Jonghye Woo, et al. Deep unsupervised domain adaptation: A review of recent advances and perspectives. *APSIPA Transactions on Signal and Information Processing*, 11(1), 2022.

[50] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021.

[51] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[52] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[53] Cristina Mata, Kanchana Ranasinghe, and Michael S Ryoo. Copt: Unsupervised domain adaptive segmentation using domain-agnostic text embeddings. In *European conference on computer vision*, page 424–440, 2024.

[54] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1369–1378, 2021.

[55] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3764–3773, 2020.

[56] Jinman Park, Francois Barnard, Saad Hossain, and Sirisha Rambhatla. Implicit stylization for domain adaptation. In *NeurIPS Workshop on Domain Generalization*, 2023.

[57] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2085–2094, 2021.

[58] Zhiliang Peng, Wei Huang, Zonghao Guo, Xiaosong Zhang, Jianbin Jiao, and Qixiang Ye. Long-tailed distribution adaptation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3275–3282, 2021.

[59] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[60] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[61] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[62] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[63] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[64] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[65] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7374–7383, 2019.

[66] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3139–3153, 2020.

[67] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.

[68] Deval Shah. Intersection over Union (IoU). https://www.v7labs.com/blog/intersection-over-union-guide. [Accessed 31-03-2025].

[69] Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36:55361–55374, 2023.

[70] Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36:55361–55374, 2023.

[71] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, pages 2058–2065, 2016.

[72] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pages 443–450. Springer, 2016.

[73] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022.

[74] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5227–5237, 2022.

[75] Richard Szeliski. *Computer Vision: Algorithms and Applications*. Texts in Computer Science. Springer, 2nd edition, 2022.

[76] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 1195–1204. Curran Associates, Inc., 2017.

[77] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[78] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[79] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1379–1389, 2021.

[80] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[81] Antonin Vobecky, David Hurych, Oriane Siméoni, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Drive&segment: Unsupervised semantic segmentation of urban scenes via cross-modal distillation. In *European conference on computer vision*, pages 478–495. Springer, 2022.

[82] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019.

[83] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[84] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021.

[85] Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. A sober look at the robustness of clips to spurious features. In *Advances in Neural Information Processing Systems*, 2024.

[86] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[87] Xinyi Wu, Zhenyao Wu, Hao Guo, Lili Ju, and Song Wang. Dannet: A one-stage domain adaptation network for unsupervised nighttime semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15769–15778, 2021.

[88] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[89] Senqiao Yang, Zhuotao Tian, Li Jiang, and Jiaya Jia. Unified language-driven zero-shot domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23407–23415, 2024.

[90] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[91] Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325. Springer, 2024.

[92] Lechen Zhang, Tolga Ergen, Lajanugen Logeswaran, Moontae Lee, and David Jurgens. Sprig: Improving large language model performance by system prompt optimization. *arXiv preprint arXiv:2410.14826*, 2024.

[93] Pan Zhang, Bo Zhang, Ting Zhang, Dong Chen, Yong Wang, and Fang Wen. Prototypical pseudo label denoising and target structure learning for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12414–12424, 2021.

[94] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022.

[95] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.

[96] Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pages 289–305, 2018.