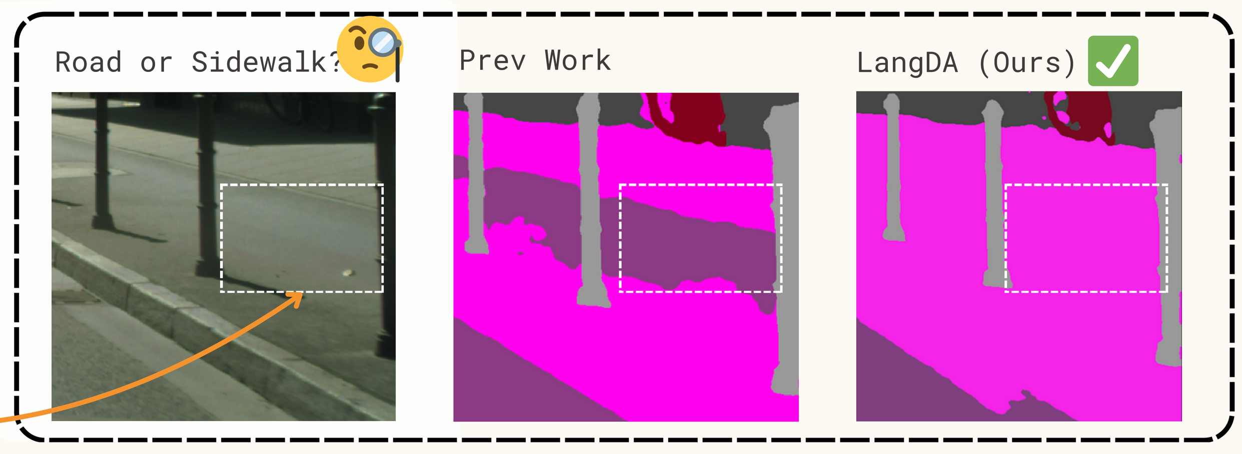
LangDA: Language-guided Domain Adaptive Semantic Segmentation

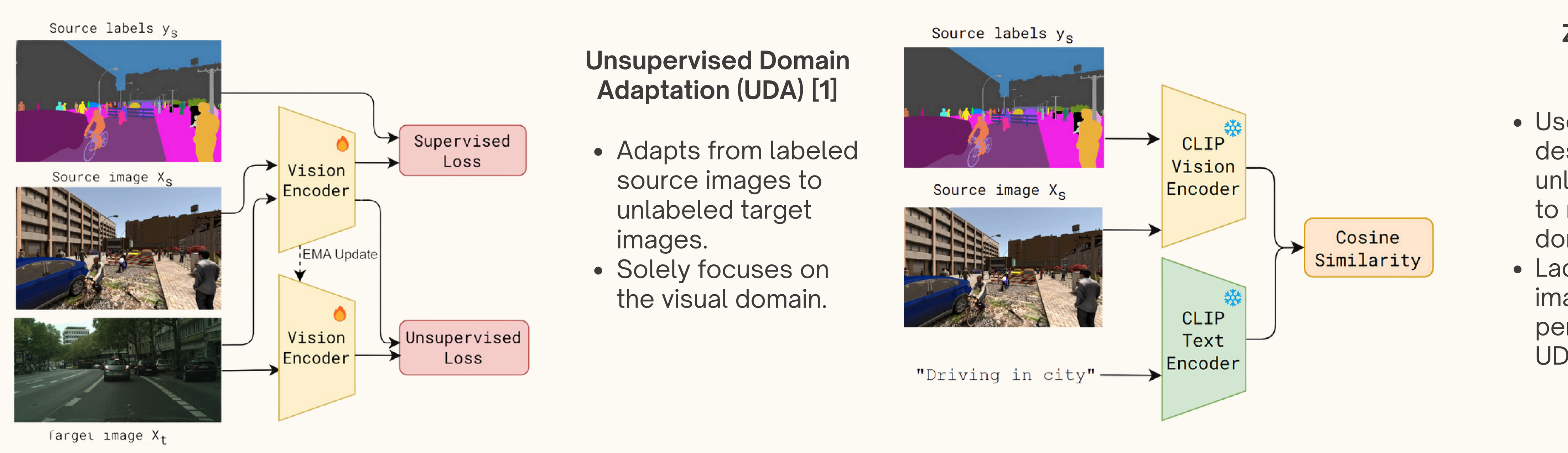
Chang Liu[1], Saad Hossain[1], C Thomas[2], Kwei-Herng Lai[2], Raviteja Vemulapalli[2], Sirisha Rambhatla[1], Alexander Wong[1,2]
University of Waterloo[1], Apple[2]

1 Motivation

- Semantic segmentation is a dense prediction task requiring expensive and time-consuming pixel-level annotations.
- Unsupervised domain adaptation (UDA) aims to transfer knowledge from a label-rich source domain to a target domain with no labels.
- Traditional UDA methods rely solely on image domains for knowledge transfer and struggle to distinguish visually similar classes such as road and sidewalk.
- Can we leverage new modalities to aid semantic segmentation in UDA?
- We propose LangDA—the first approach to leverage language for aligning vision domains to differentiate visually similar classes in domain adaptive semantic segmentation.



2 Prior Works



Unsupervised Domain Adaptation (UDA) [1]

- Adapts from labeled source images to unlabeled target images.
- Solely focuses on the visual domain.

Zero-shot Domain Adaptation [2]

- Uses a prompt description instead of unlabeled target images to mitigate potential domain shifts.
- Lacking unlabeled target images results in reduced performance compared to UDA.

Full Paper

chang.liu@uwaterloo.ca

Figure 1: **Prior Domain Adaptation Methods.** (a) Traditional UDA uses both source images & target images. (b) Zero-shot adaption leverages a prompt description instead of unlabeled target images.

3 LangDA: Language-guided Domain Adaptive Semantic Segmentation (DASS)

We introduce LangDA, the first work that leverages both language and visual domains in UDA for semantic segmentation. LangDA is composed of two modules: **i) Text Generation**, which creates language descriptors, and **ii) Prompt-guided Adaptation**, aligning prior linguistic knowledge with visual features to facilitate adaptation.

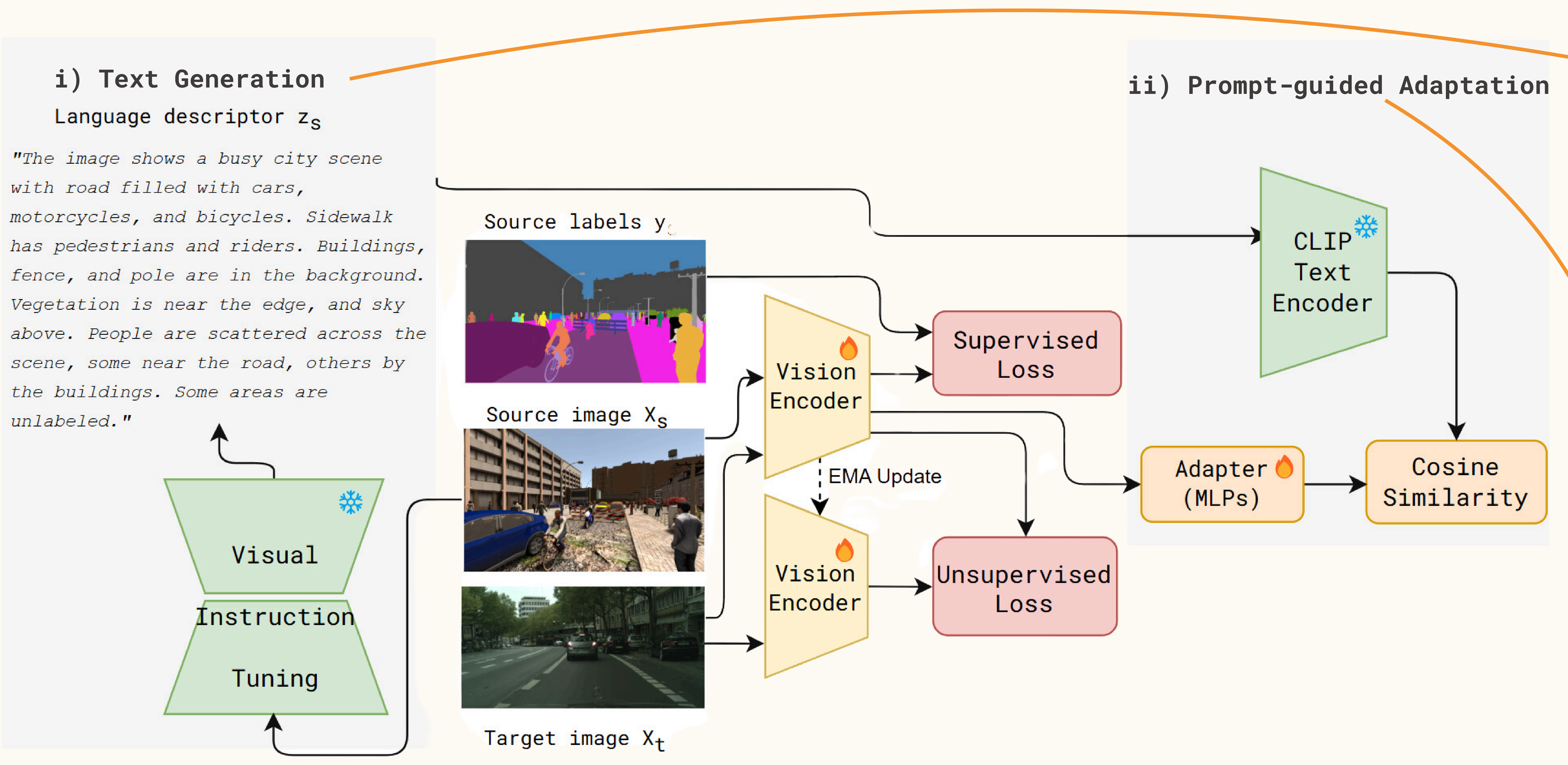


Figure 2: **Model Architecture.** Our proposed LangDA method generates image-level language descriptors to facilitate adaptation while introducing very few learnable parameters (namely adapters)

i) Text Generation

We first generate the language descriptors using Visual Instruction Tuning, and obtain

$$\mathcal{C}_S = \{z_S^{(i)} \mid z_S^{(i)} \in \mathbb{R}^l\}$$

We then refine the generated descriptors using a Large Langauge Model (LLM).

$$\mathcal{C}_r = \{z_r^{(i)} \mid z_r^{(i)} \in \mathbb{R}^l, l \leq 77\}$$

We pass the refined captions $\mathcal{C}_r$ to the frozen CLIP encoder to obtain the set of language feature vectors.

$$\mathcal{V}_{\text{CLIP}} = \{v_{\text{CLIP}}^{(i)} \mid v_{\text{CLIP}}^{(i)} = E_{\text{CLIP}}(z_r^{(i)}), v_{\text{CLIP}}^{(i)} \in \mathbb{R}^{512}\}$$

ii) Prompt-guided Adaptation

To align source image features with domain-invariant textual features, we introduce an image-level minimization objective on the distance between CLIP's text features and source image feature,

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

where $f_S^{(i)}$ is the pooled visual feature for each image.

4 Results & Discussions

Takeway: LangDA effectively distinguishes visually similar classes, like road and sidewalk, where previous methods struggle. LangDA achieves a notable +0.9% improvement in mIoU, a significant result for dense prediction tasks like semantic segmentation. These promising results highlights the potential of language-guided DASS.

Future work: We are currently extending LangDA to multi-resolution adaptation frameworks and evaluating LangDA on diverse adaptation scenarios, including Normal → Adverse Weather and Day → Night transitions. Preliminary results demonstrate significant performance gains.

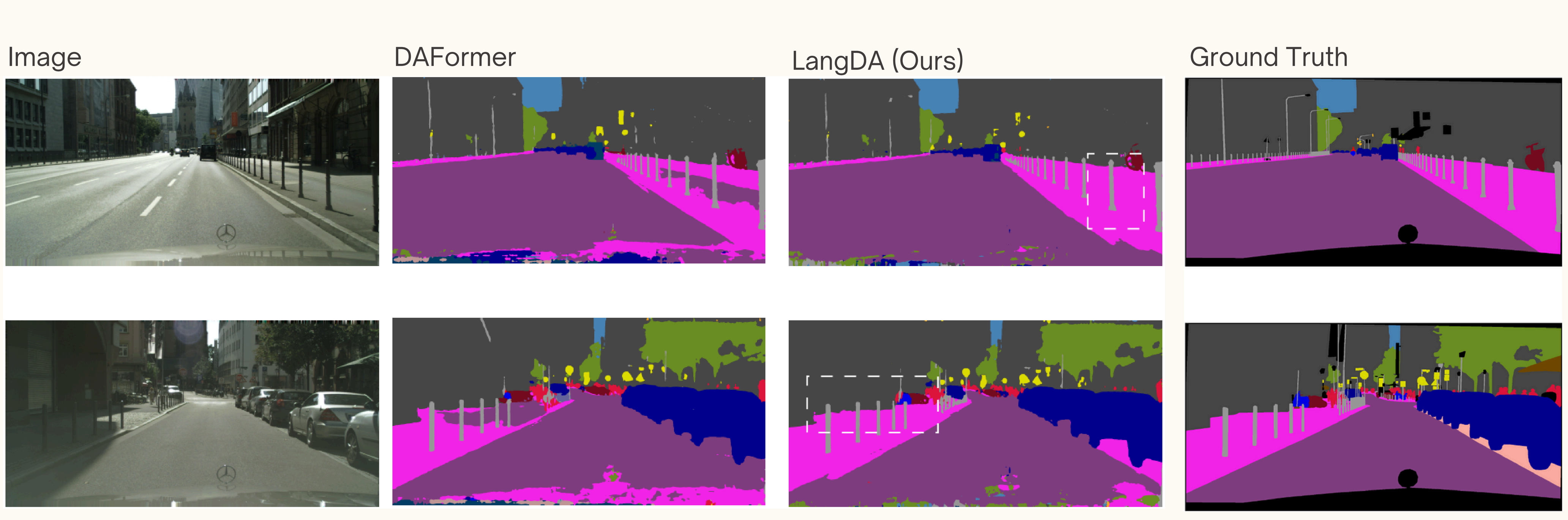
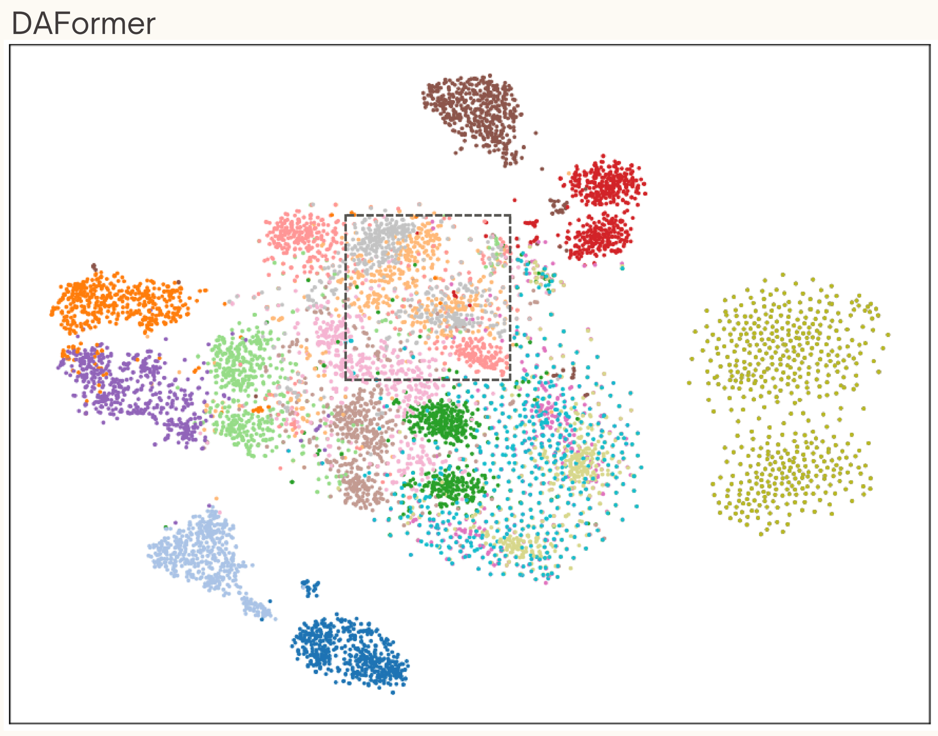


Figure 3: **Qualitative Results: Synthia→Cityscapes.** Existing DASS approaches face difficulty discerning visually similar classes (e.g. road and sidewalk). Our proposed method, LangDA, successfully segments visually similar pixels under language guidance.

Method	Backbone	Unlabeled Target Data	Text Prompts	% mIoU ↑
Source only	ResNet-50			29.3
PODA† [8]	ResNet-50		✓	29.5
ULDA† [40]	ResNet-50		✓	30.8
Source only	ResNet-101			29.4
ADVENT [37]	ResNet-101	✓		41.2
CBST [43]	ResNet-101	✓		42.6
DACS [36]	ResNet-101	✓		48.3
CorDA [38]	ResNet-101	✓		55.0
ProDA [42]	ResNet-101	✓		55.5
DAFormer† [11]	SegFormer	✓		61.1
LangDA (Ours)	SegFormer	✓	✓	**62.0**

Table 1: **Comparison with state-of-the-art methods in UDA and Zero-shot DA.** We performed our experiments on standard synthetic-to-real adaptation benchmark Synthia → Cityscapes. "Source only" refers to lower bound DA baselines with no adaptation (i.e. training on source and evaluation on target).



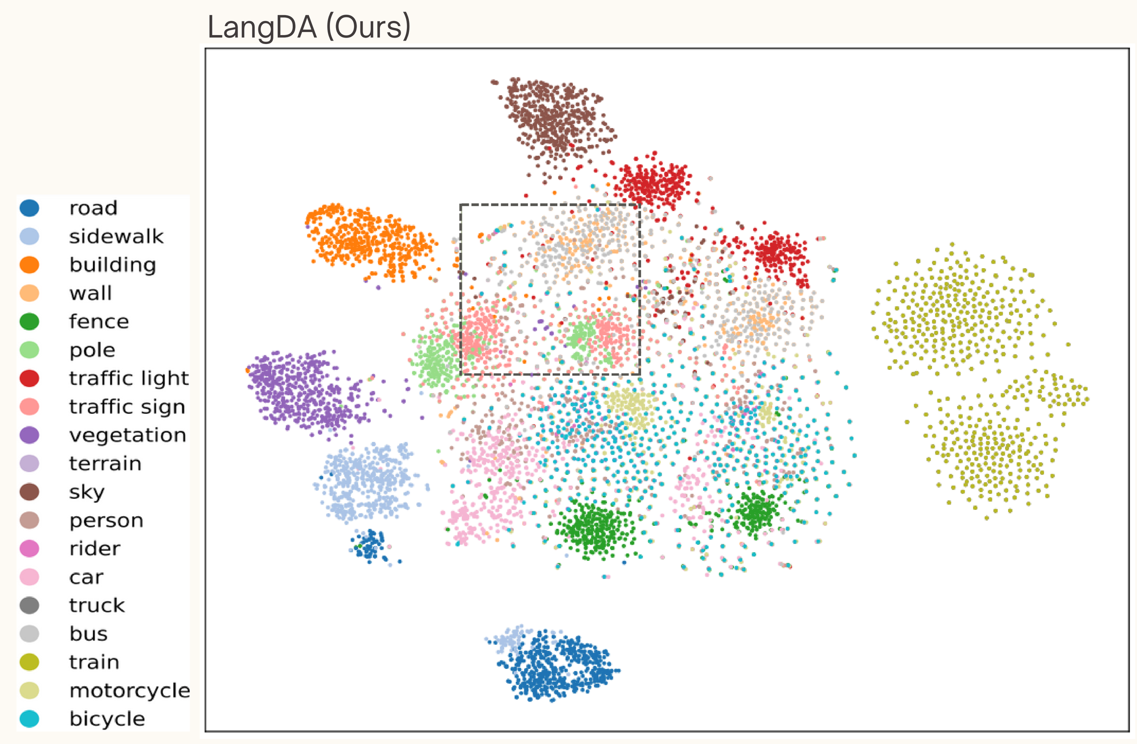


Figure 3: **Qualitative Results: t-SNE.** LangDA shows improved feature clustering. In DAFormer's t-SNE, the feature representations of walls (light orange) and traffic signs (rose pink) overlap in the image domain, likely due to traffic signs often visually appearing in front of walls from driver's first-person view. On the other hand, walls and traffic signs are semantically distinguishable in language, contributing to LangDA's enhanced segmentation result.

References

[1] L. Hoyer, D. Dai, and L. 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, 2022, pp. 9924–9935.
[2] S. Yang, Z. Tian, L. Jiang, and J. Jia, "Unified language-driven zero-shot domain adaptation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 23 407–23 415.

Sirisha Rambhatla to acknowledge support of the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance Grant, ALLRP 587270 - 23.