Using Language to Mitigate Distribution Shift in Unsupervised Semantic Segmentation

1/21/2025

Chang Liu
MASc in Systems Design Engineering

To fulfill Master's Seminar Milestone at the University of Waterloo

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Modern Deep Learning Tasks

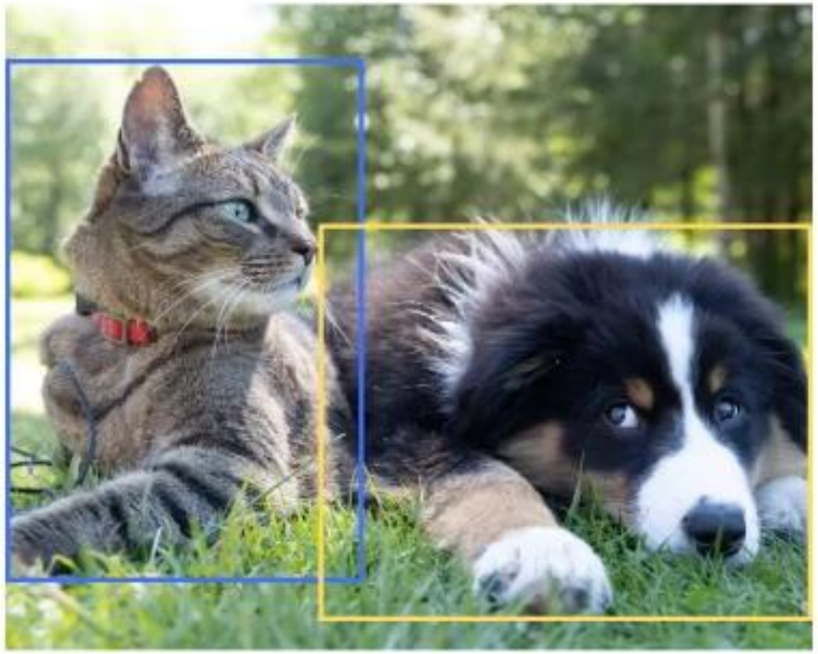
Is this a cat?

"cat"

Modern Deep Learning Tasks

Is this a cat?

What is there in the image
and where?

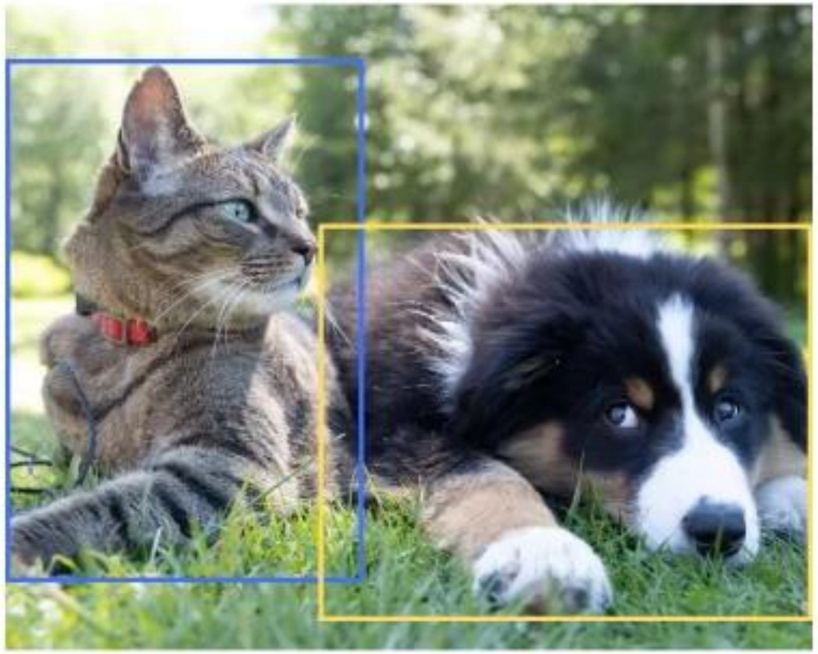


UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Modern Deep Learning Tasks

Is this a cat?

What is there in the image and where?



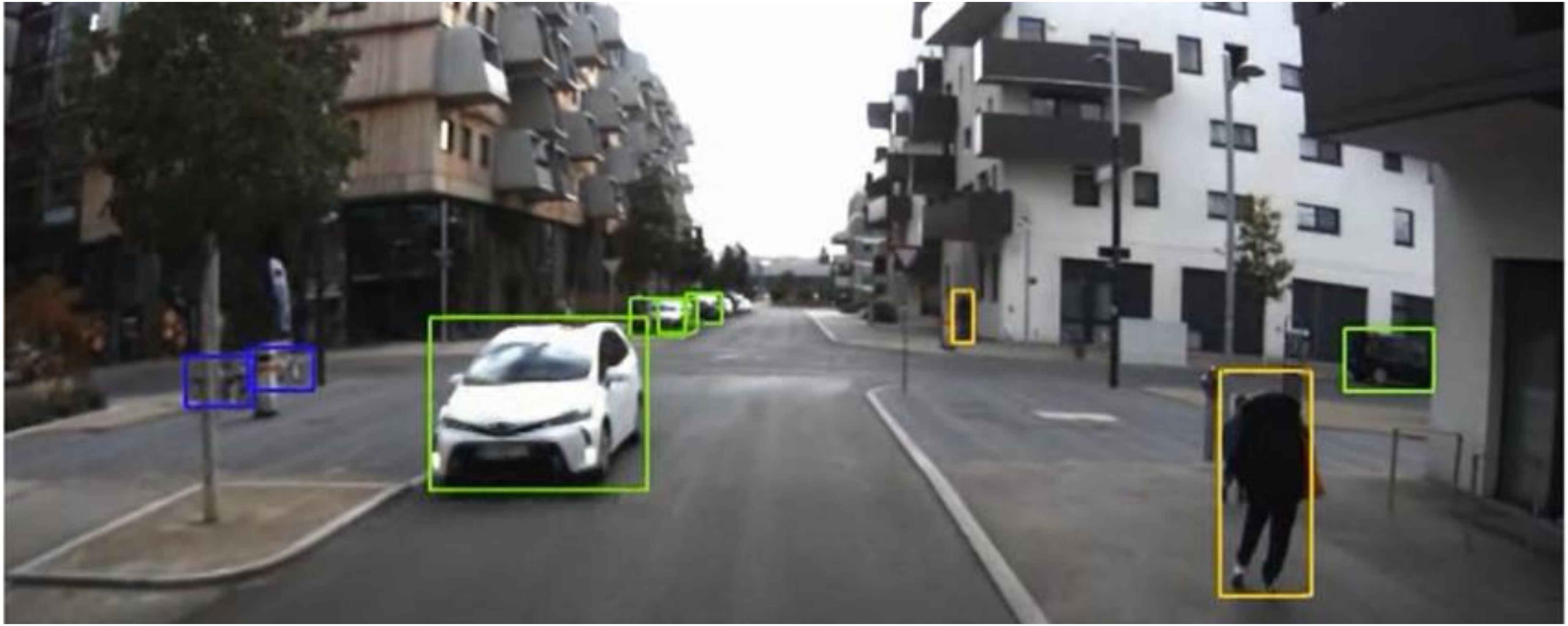
Which pixels belong to which object

COMPLEX SCENE: HOW SHOULD WE UNDERSTAND IT?

Classification Detection Semantic Segmentation

Classification: This is... a road?

Detection



Semantic Segmentation

However... Manual Annotation 😭

Deep learning methods

Semi-supervised learning

Weakly-supervised learning

Transfer learning

Unsupervised domain adaptation

Learning from Synthetic Data

Zero-shot learning and few-shot learning

Active learning

Deep learning methods

Semi-supervised learning

Weakly-supervised learning

Transfer learning

Unsupervised domain adaptation

Learning from Synthetic Data

Zero-shot learning and few-shot learning

Active learning

Transfer and Adaptation

- Learn on one task, transfer to another
- Learn on one labelled distribution, test on another distribution

Transfer and Adaptation

- Learn on one task, transfer to another
- Learn on one labelled distribution, test on another distribution

Unsupervised Domain Adaptation

Domain Gap

Different, but related data distributions

Source domain -> Target domain

- Different weather, lighting, locations
- Synthetic vs. real

Domain Gap

Different, but related data distributions

Source domain -> Target domain

- Different weather, lighting, locations
- Synthetic vs. real

UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

Domain Gap

Different, but related data distributions

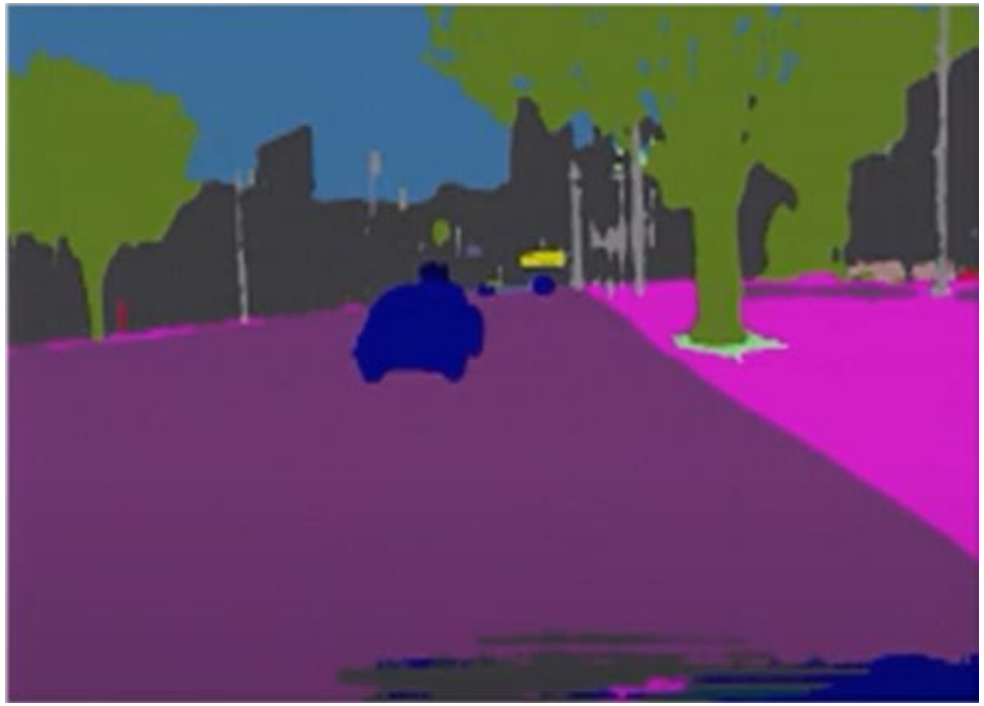
Source domain -> Target domain

- Different weather, lighting, locations
- Synthetic vs. real

Domain Gap

Different, but related data distributions

Source domain -> Target domain

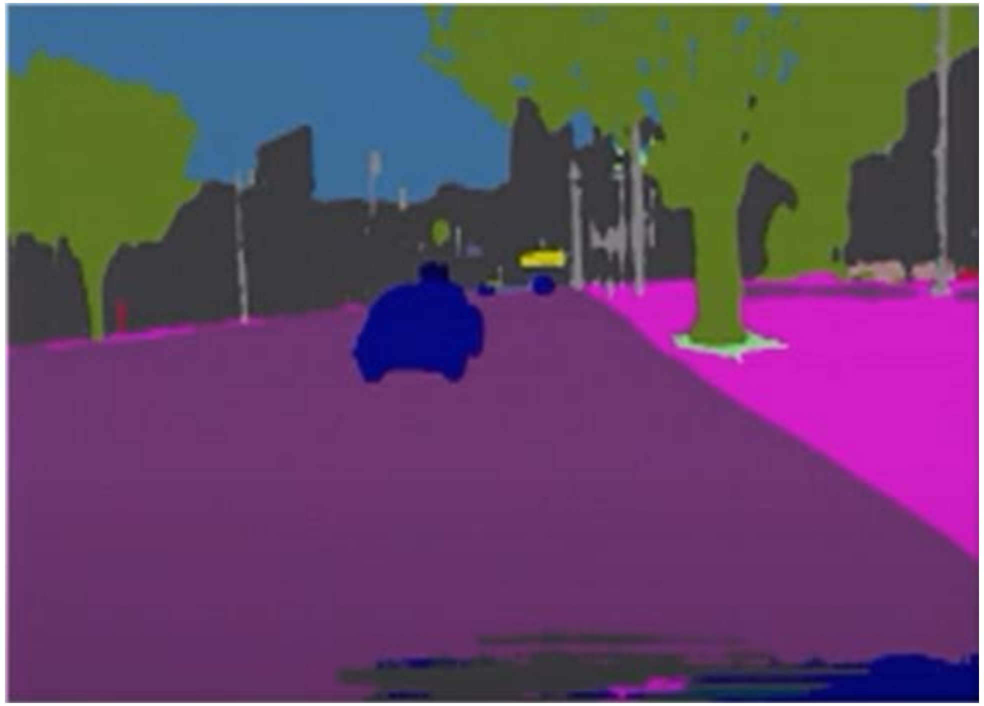


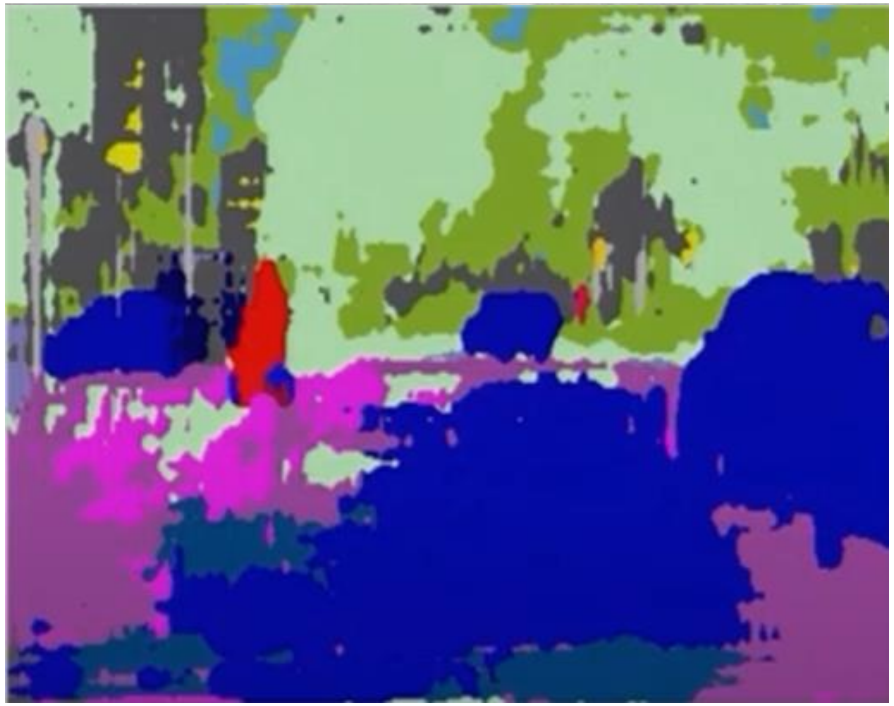
- Different weather, lighting, locations
- Synthetic vs. real

Domain Gap

Different, but related data distributions

Source domain -> Target domain



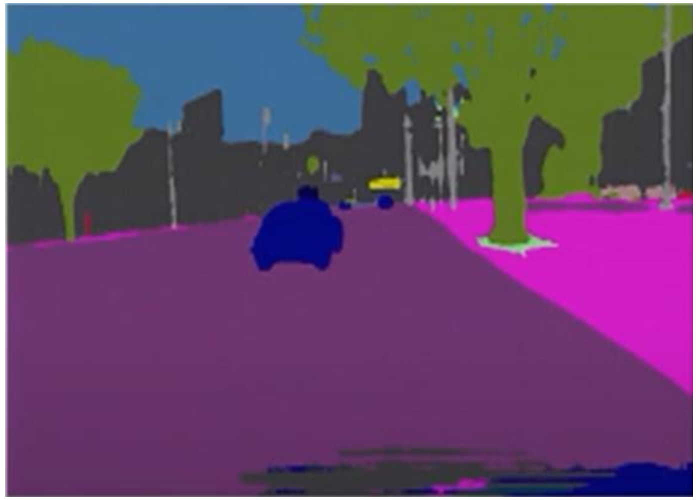


- Different weather, lighting, locations
- Synthetic vs. real

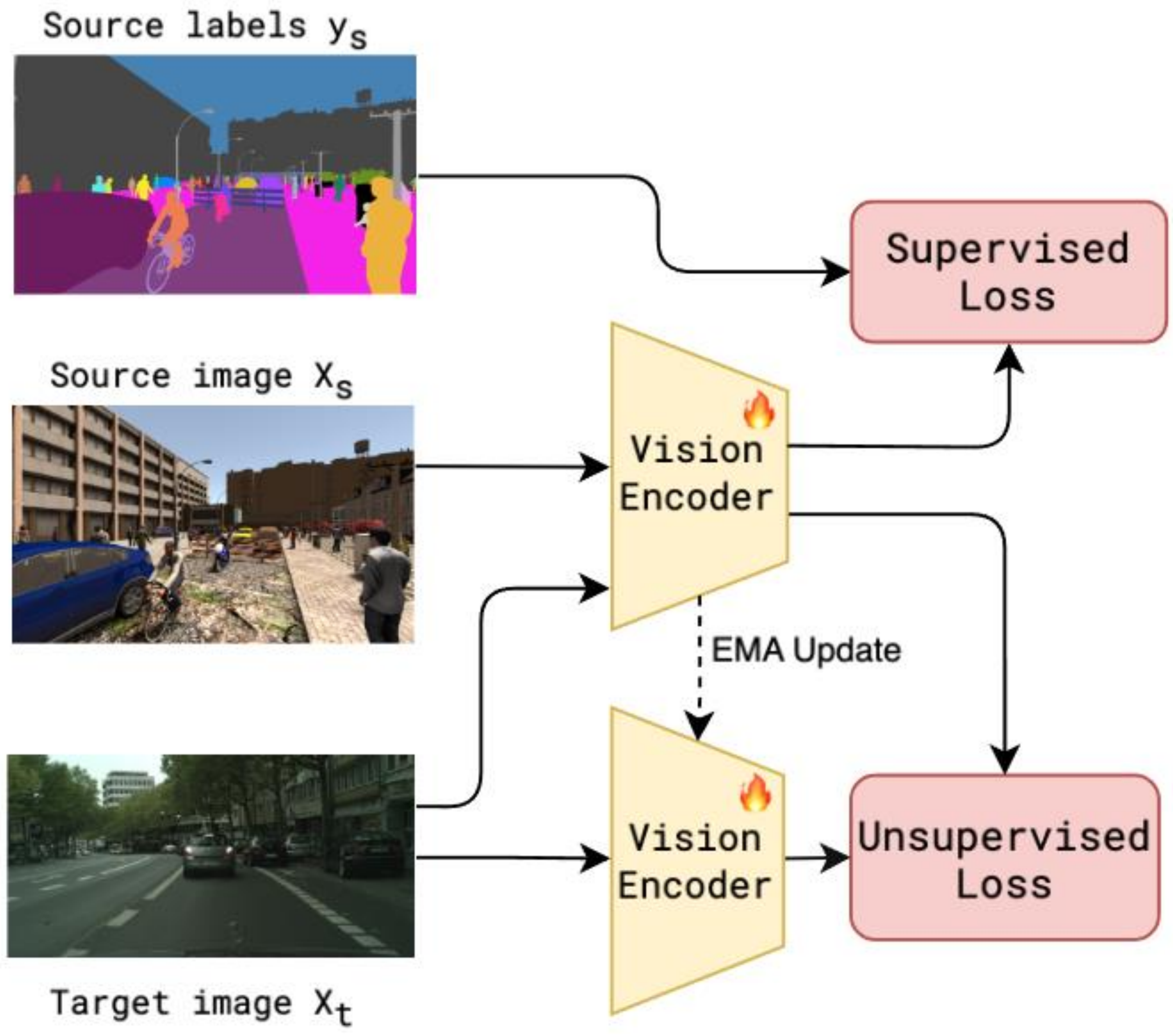
Unsupervised Domain Adaptation (UDA)

Labeled Source Domain

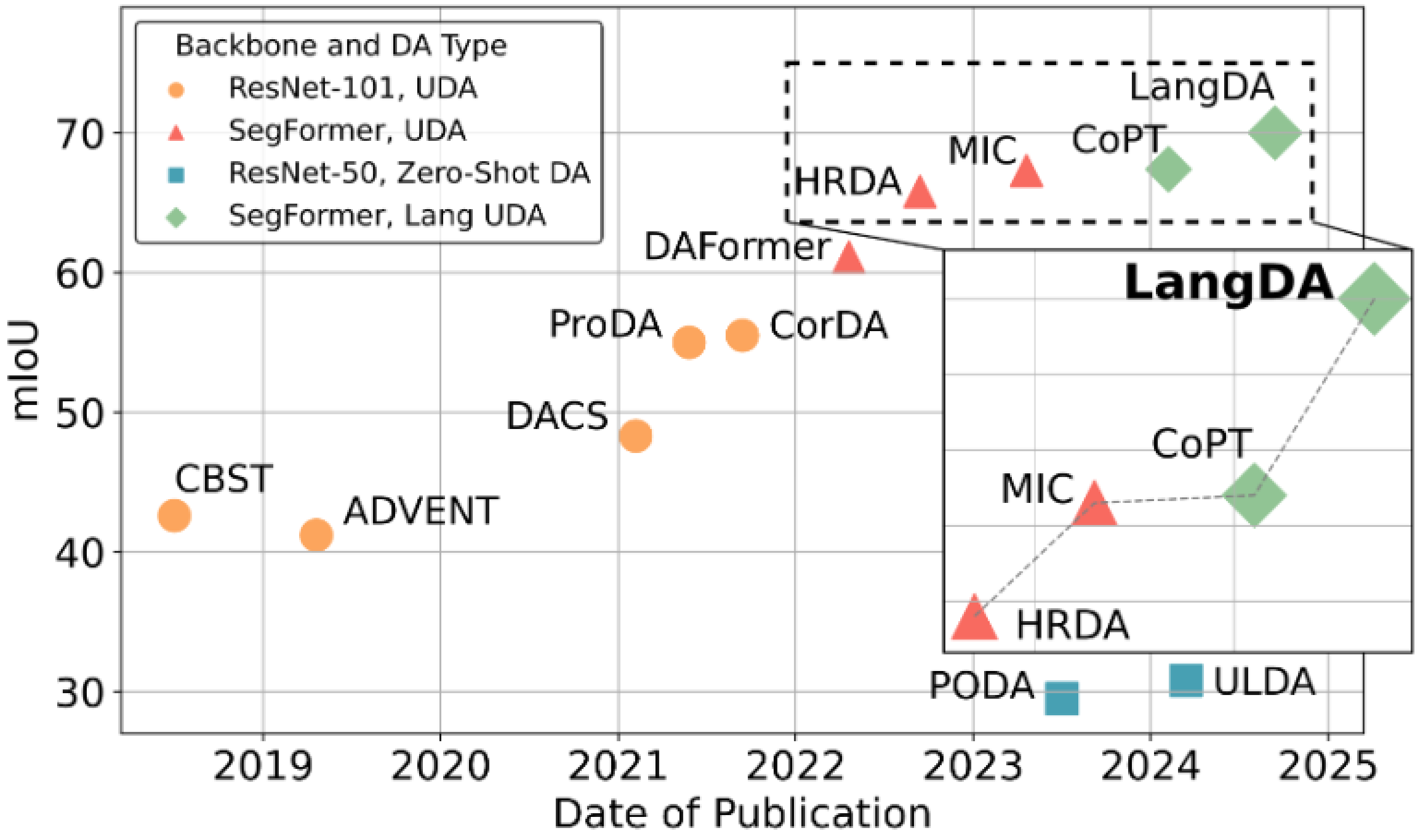
Unlabeled Target Domain



Traditional UDA Method



Traditional UDA Method has plateau-ed in the last 2 yrs



Give the model more information?

This image depicts a snowy urban street scene. Key details include:

1. **Buildings**: On the left, a mix of modern apartment complexes and a bright yellow building with the text "Restaurant Piaton" is visible. On the right, there is a church building with "Katholische Kirche St. Katharina" written on its wall.

2. **Street**: The road appears wet with patches of snow and slush. Sidewalks are snow-covered, with footprints visible.

3. **Traffic**: Traffic lights show green, and overhead power lines suggest tram or trolleybus infrastructure.

4. **Weather**: Overcast sky, snow on trees and rooftops, indicating recent or ongoing cold weather.

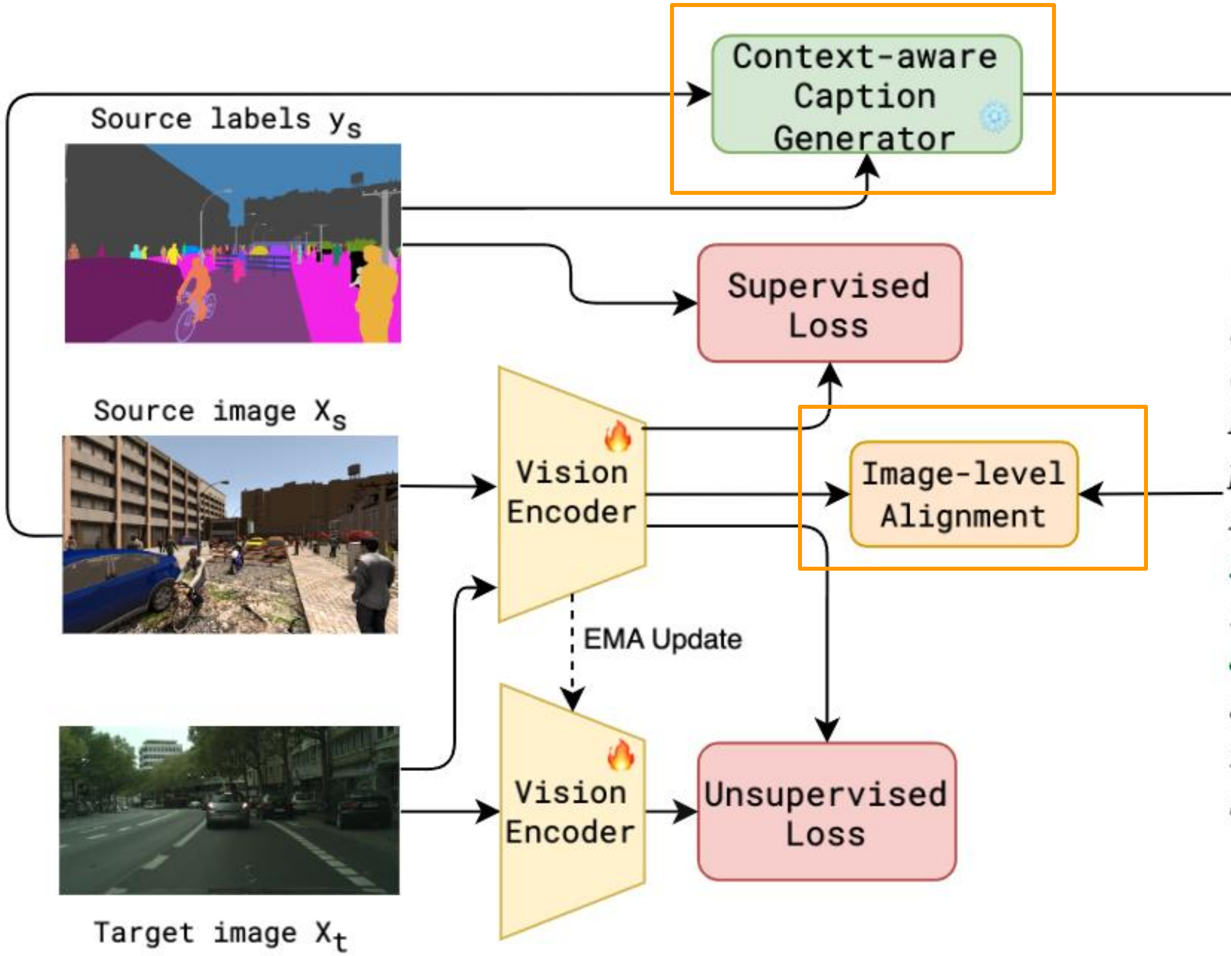
Vision Language Models understands the road scene (world priors)

Please describe this image.

This image depicts a snowy urban street scene. Key details include:

1. **Buildings:** On the left, a mix of modern apartment complexes and a bright yellow building with the text "Restaurant Piaton" is visible. On the right, there is a church building with "Katholische Kirche St. Katharina" written on its wall.
2. **Street:** The road appears wet with patches of snow and slush. Sidewalks are snow-covered, with footprints visible.
3. **Traffic:** Traffic lights show green, and overhead power lines suggest tram or trolleybus infrastructure.
4. **Weather:** Overcast sky, snow on trees and rooftops, indicating recent or ongoing cold weather.

Bringing in Language Information – LangDA (Ours)



"*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.*"

Caption Generation with VLM

Caption Generation: VLM

Query (Image):

Query (Text):

Describe the image in detail for semantic segmentation tasks. Be sure to include the class names ['road', 'sidewalk', 'building', 'wall', 'fence', 'pole', 'vegetation', 'sky', 'person', 'rider', 'car', 'motorcycle', 'bicycle', 'unlabeled'] 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.

Caption Refinement with LLM

Caption Refinement

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 {CLASS_NAMES} and their pixel locations. The description is {VLM_CAPTION}

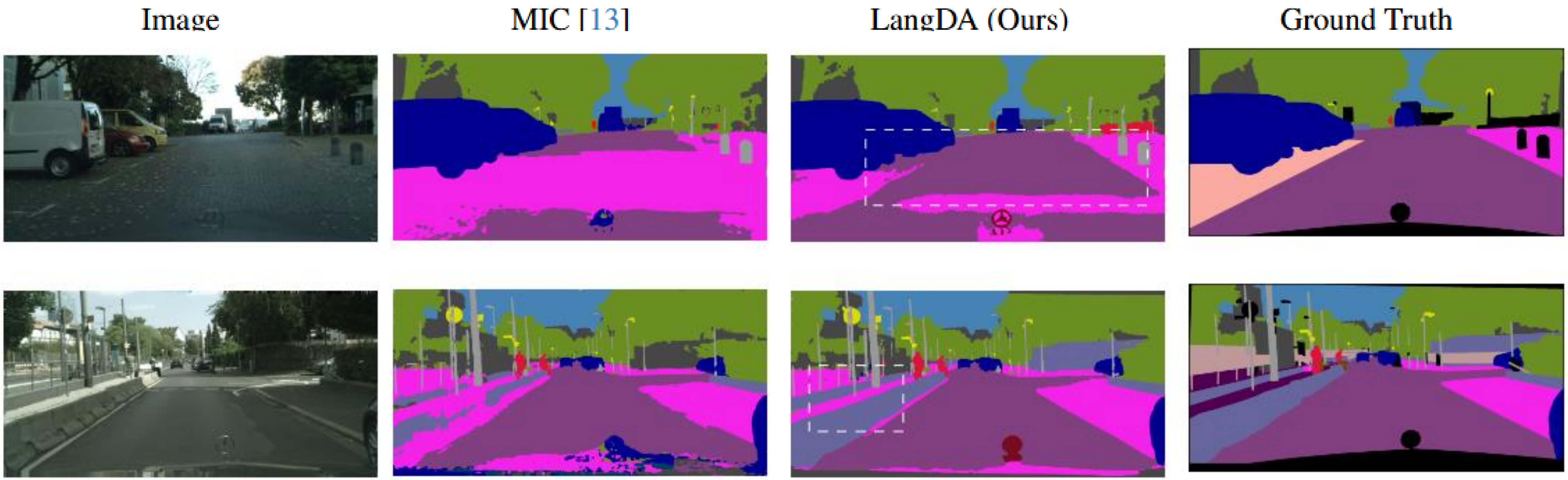
LLM:
The image shows a busy city scene with road filled with cars, motorcycles, and bicycles. Sidewalk has pedestrians and riders. Buildings, fence, and pole are in the background. Vegetation is near the edge, and sky above. People are scattered across the scene, some near the road, others by the buildings.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

Align Image Text Features - Consistency Objective

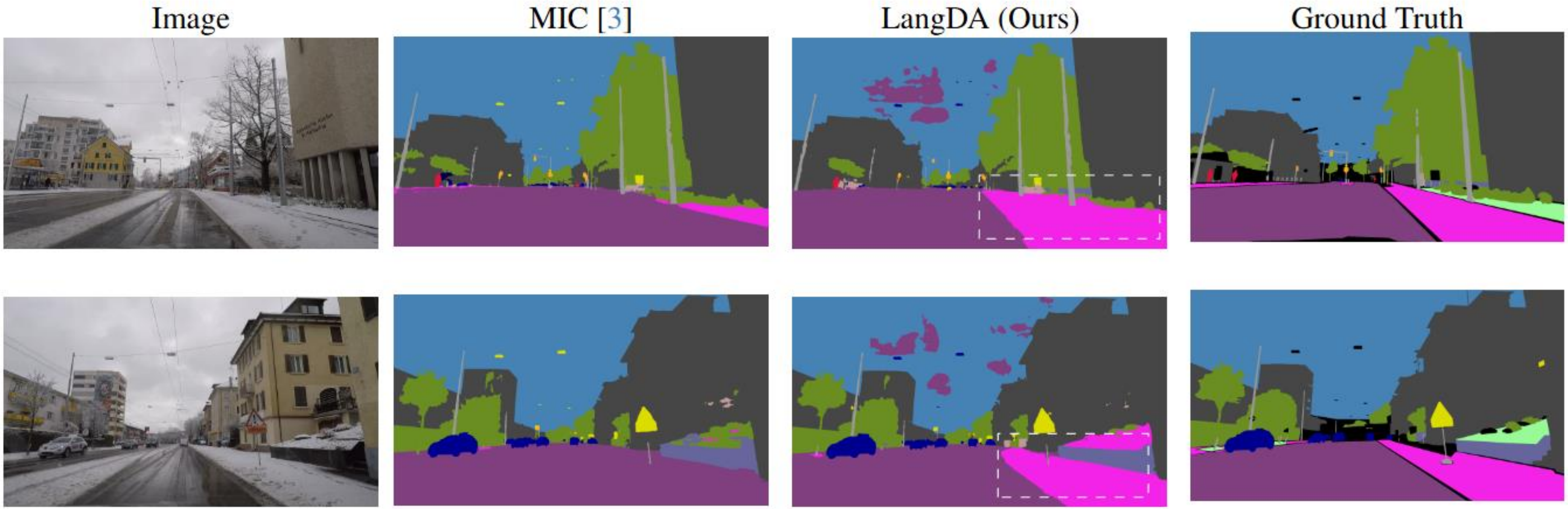
"The image shows a busy city scene with road filled with cars, motorcycles, and bicycles. Sidewalk **has** pedestrians and riders. Buildings, fence, and pole are **in the background**. Vegetation is **near the edge**, and sky **above**. People are scattered across the scene, some **near the** road, others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

$$\mathcal{L}_p^{(i)}(f_{\text{pool}}^{(i)}, v_{\text{CLIP}}^{(i)}) = 1 - \frac{f_{\text{pool}}^{(i)} \cdot v_{\text{CLIP}}^{(i)}}{\|f_{\text{pool}}^{(i)}\| \, \|v_{\text{CLIP}}^{(i)}\|}. \tag{6}$$

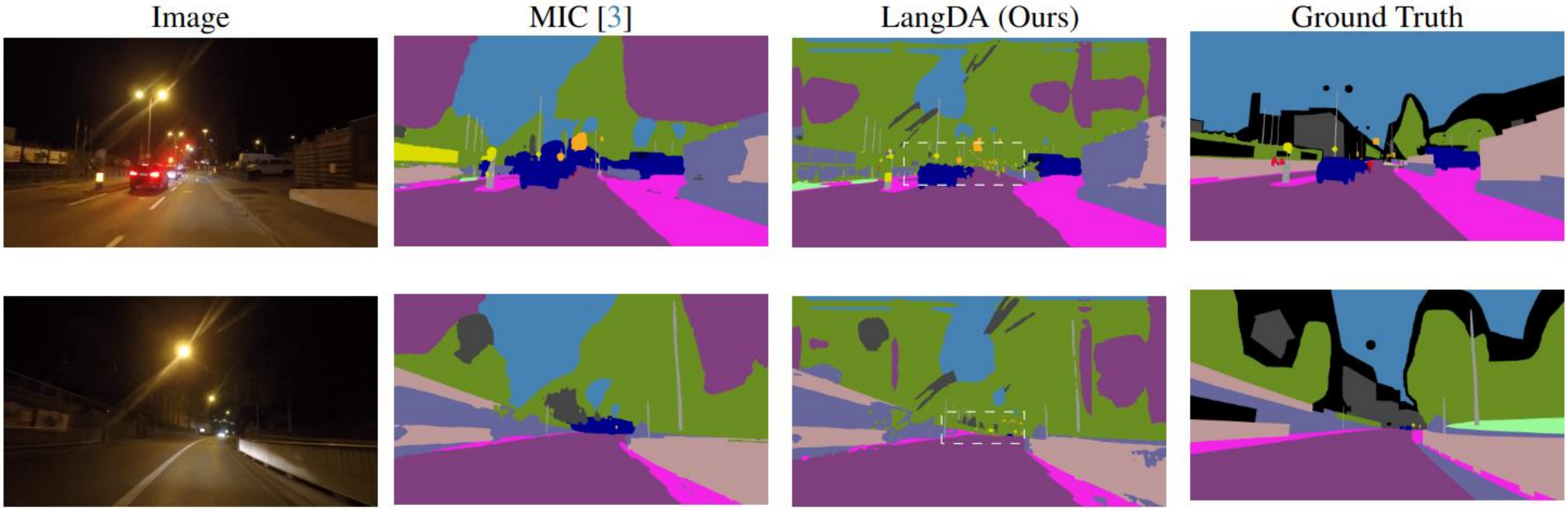
Qualitative Results: Synthetic-to-Real Adaptation



Qualitative Results: Normal-to-Adverse-weather Adaptation

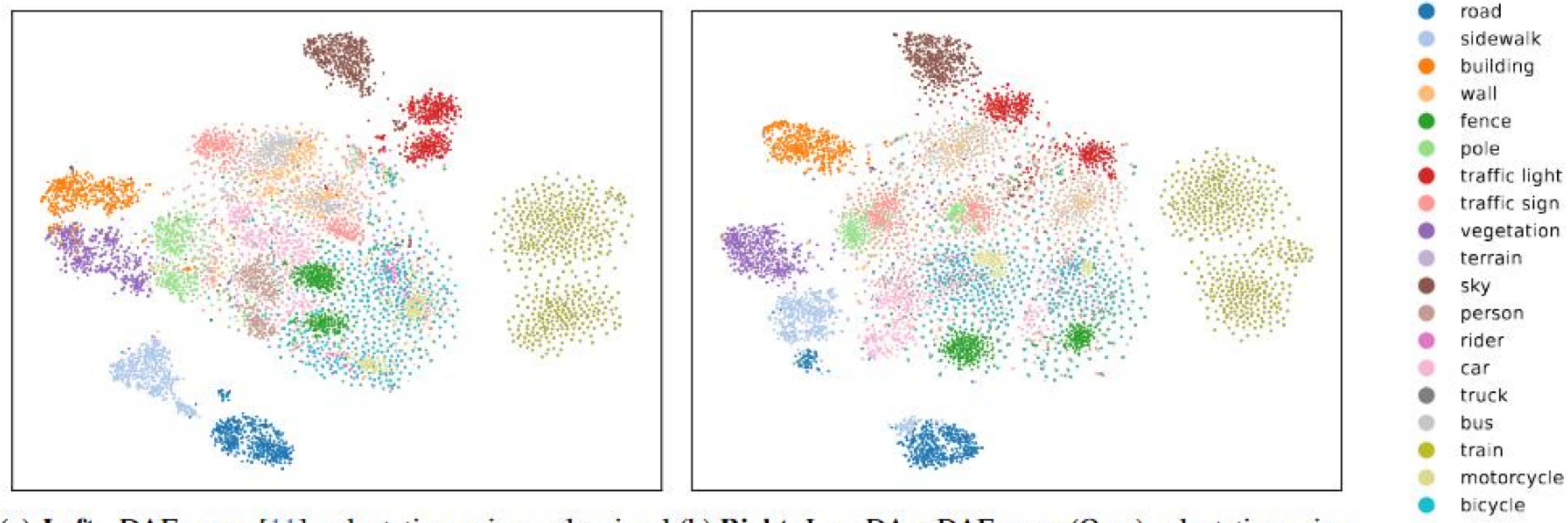


Qualitative Results: Day-to-night Adaptation



UNIVERSITY OF WATERLOO | FACULTY OF ENGINEERING

t-SNE



(a) **Left**: DAFormer [11], adaptation using only visual images.

(b) **Right**: LangDA + DAFormer (Ours), adaptation using both visual images and contextual language descriptions.

Figure 9. **t-SNE of DAFormer and LangDA (Ours)** After aligning language and visual features, we observe more well-defined boundaries and improved class clustering.

Quantitative Result

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) + DAFormer	SegFormer	✓	✓	**62.0**
HRDA [12]	SegFormer	✓		65.8
LangDA (Ours) + HRDA	SegFormer	✓	✓	**66.3**
MIC [13]	SegFormer	✓		67.3
CoPT [24]	SegFormer	✓	✓	67.4
LangDA (Ours) + MIC	SegFormer	✓	✓	**70.0**

Ablations

Table 6. Ablation on different prompting and aligning techniques on Synthetic-to-Real adaptation benchmark: Synthia $\rightarrow$ Cityscapes.

	Context-aware Caption Generation	**Image-level Alignment**	**% mIoU↑**
1	✓	✓	**70.0**
2	–	✓	68.7
3	✓	–	65.7

Ablations

Table 7. Ablation on applying contextual scene description on source only, target only, and source + target.

Image Captions	% mIoU↑
Source only	**70.0**
Target only	69.1
Source + Target	68.0

UNIVERSITY OF WATERLOO

FACULTY OF ENGINEERING

Questions?