Language-Driven Semantic Change Detection in Urban Maps via Multi-Modal Deep Learning

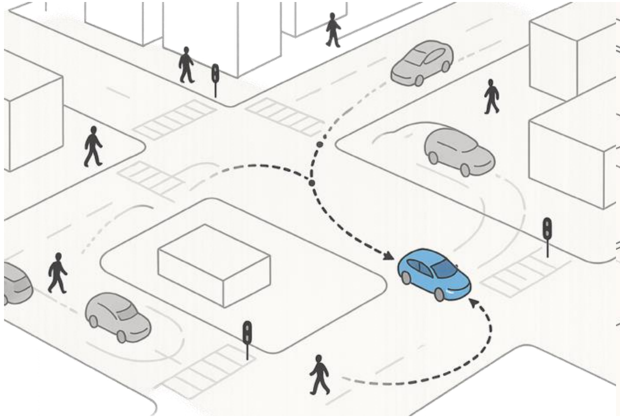
HARVEY MUDD COLLEGE

ION INSTITUTE OF NAVIGATION

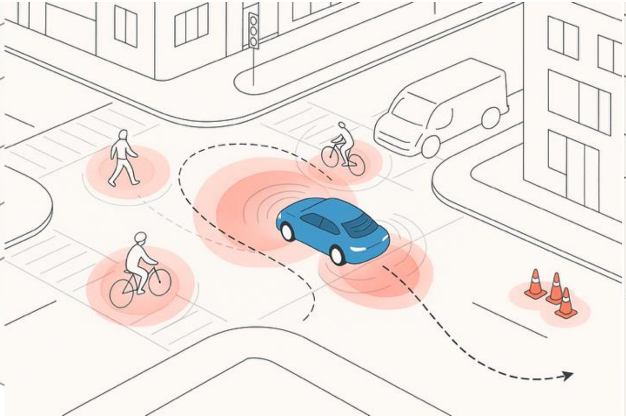
Huaze Liu, Zihao Gao, and Adyasha Mohanty, MADD Lab

Session C1, Sept 10, 2025

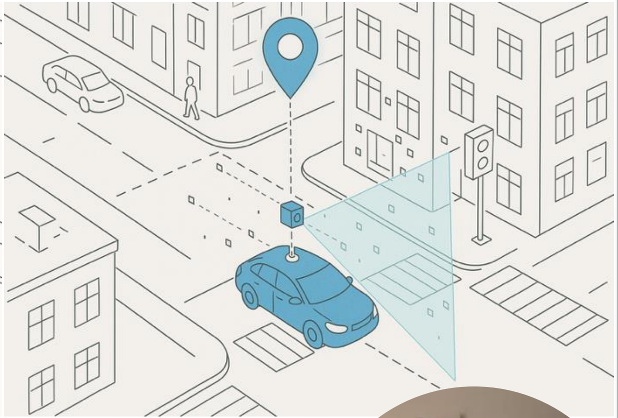
High-integrity maps are essential for autonomous navigation



Trajectory planning

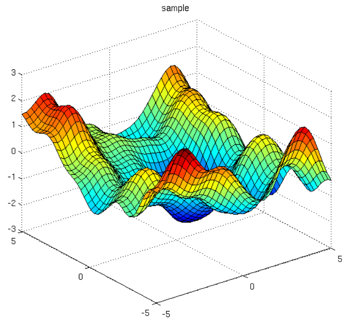


Obstacle avoidance

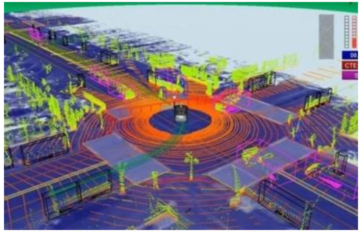


Accur
localiz

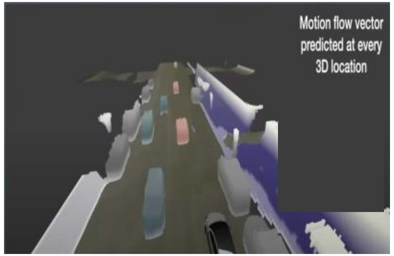
Existing Approaches



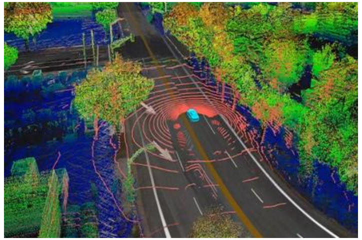
Heuristic or statistical methods[1]



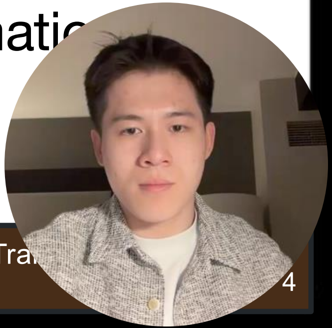
SLAM-based approaches[2]



Deep networks for occupancy maps[3]



Online HD Map Estimati

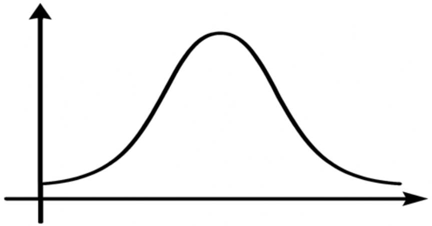


[1] Zou, L., & Sester, M. ArXiv preprint [2] Harithas, S. S., & Krishna, M. ArXiv preprint [3] Katyal, K., & Hager, G. D. IEEE Tra Robotics. [4] Gu, X., Ivanovic, B., & Pavone, M. International Conference on Intelligent Vehicles

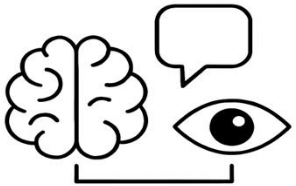
HARVEY MUDD COLLEGE

Gaps and Opportunities

Single modality approaches lack robustness → **Can we use learned methods and perform sensor fusion?**



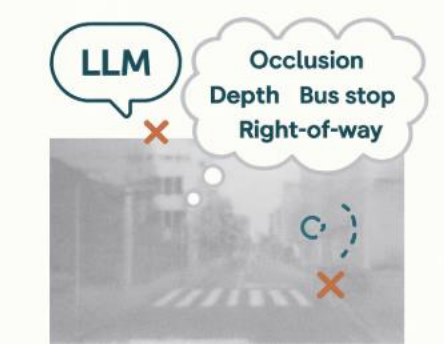
Strict noise assumptions are more suited for static environments → **Can we characterize uncertainty without strict assumptions on the noise distribution?**



Lack of semantic reasoning → **Can we use mu vision and lang models?**

Multi-modal Vision and Language Models

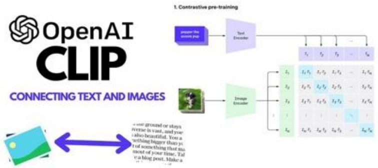
Vision models see objects but lack contextual meaning



Language models can't ground objects visually

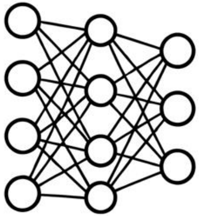
Large multi-modal models bridge this gap

Can enable "**Zero-Shot Learning**" for unseen scenarios!

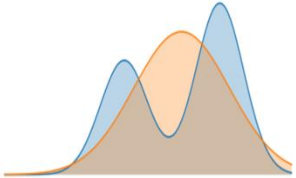


Our Contributions

We propose a LiDAR-Camera sensor fusion framework for quantifying dynamic map uncertainty as well as comprehensive scene change understanding



We use novel large vision-language models to perform zero-shot semantic segmentation for more robust change detection



We propose online KL divergence-based consist
tracking algorithm and evaluate its efficacy under
weather conditions

Outline

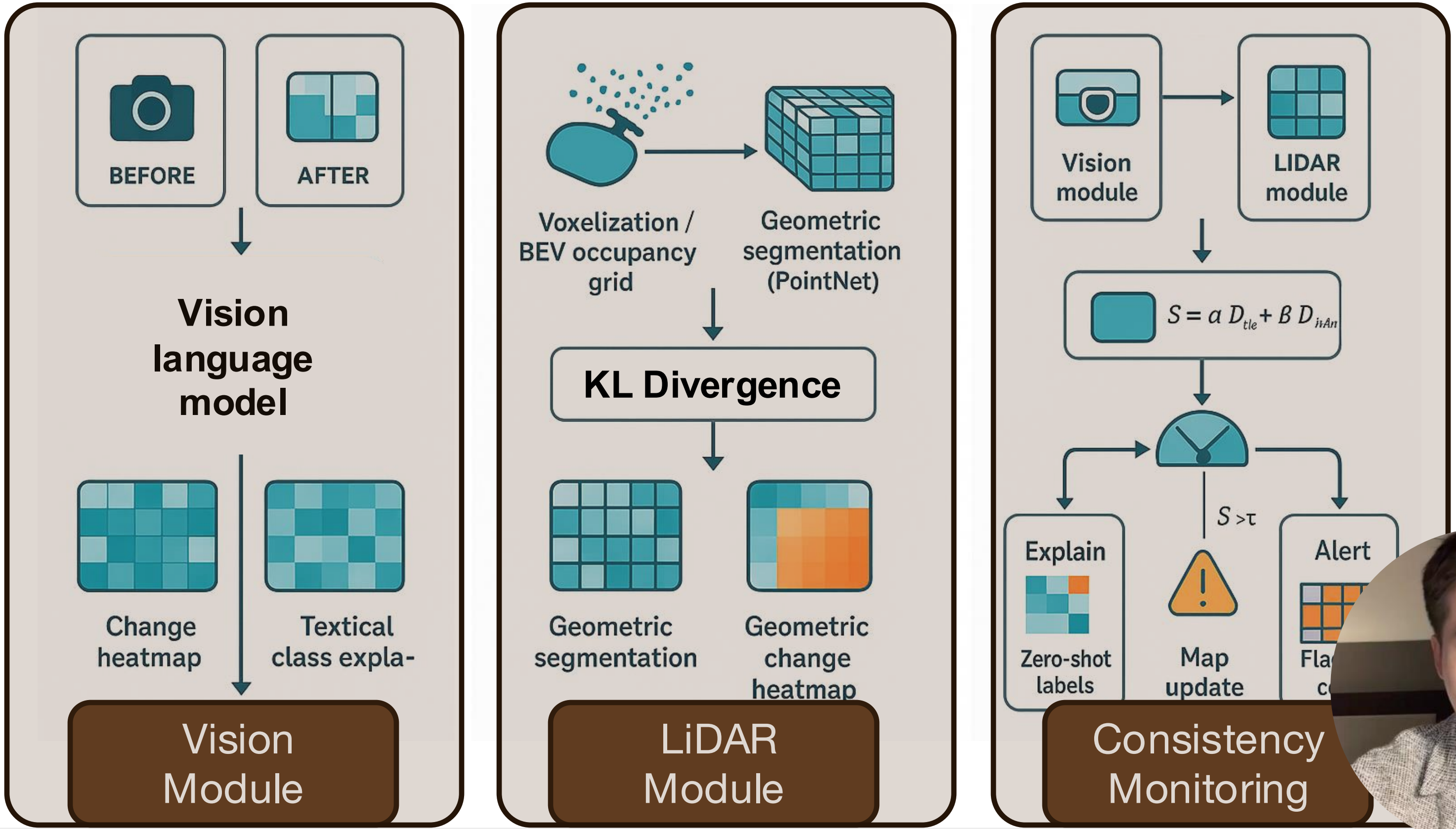
- **Proposed Framework**
  - **Vision Module**
  - **LiDAR Module**
  - **Consistency Monitoring and Sensor Fusion**

- Experiments
  - Virtual KITTI dataset setup
  - Key experimental parameters, metrics and baselines

- Key Results
  - Selected change detection accuracy results for individual se
    modalities
  - Selected sensor fusion results on adverse weather conditions

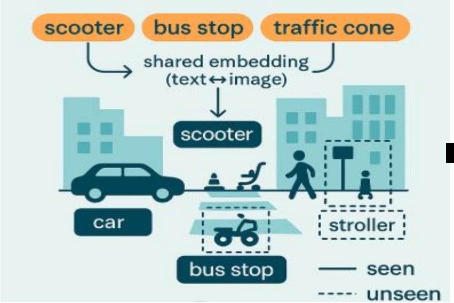
Harvey Mudd College

Proposed Framework



HARVEY MUDD COLLEGE

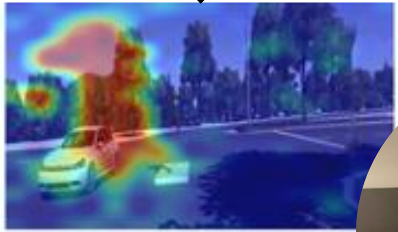
Vision Module



Zero-Shot
Object Detection[6]

Segment Anything Model
(SAM) provides initial
masks [7]

Prompts
refine
masks



Semantic change detecti
KL Divergence[8]

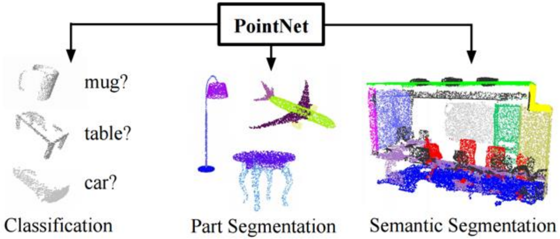
[6] Kirillov, A. et. al. ArXiv Preprint [7] Ding, X. et. al. ECCV [8] Kullback, S., & Leibler, R. A. The Annals of Mathematical Statistic

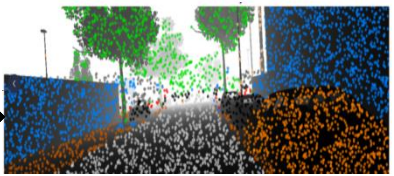
LiDAR Module

- PointNet[9]: main architectural backbone
- Chosen because lightweight and efficient

Depth images and semantic labels of old and new maps

Preprocessing



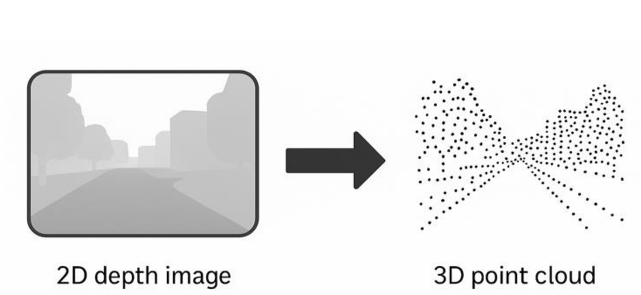


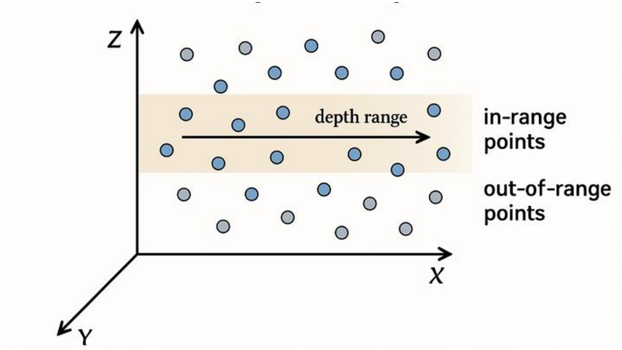
Segmented point clouds

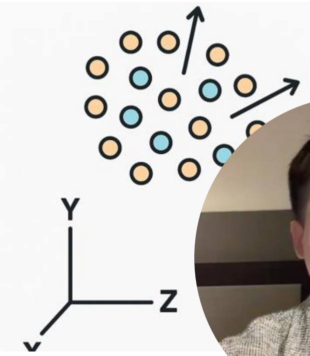
Sem
dete
D

LiDAR Module: Preprocessing and Key Modifications

- Convert depth images to point clouds using camera intrinsics and range filtering to avoid simulator boundary artifacts
- Compute local surface normal via KD-tree search[15] to capture geometric structure for improved classification
- Assign point-wise semantic categories from ground-truth annotations with unified class labels for consistent analysis

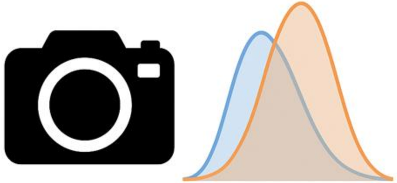






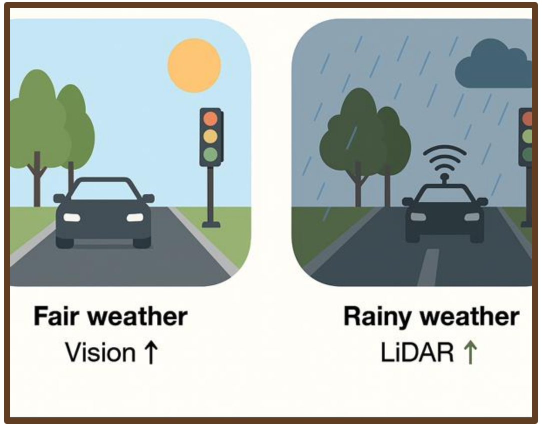
[15] Bentley, J. L. (1975). Communications of the ACM

Consistency Monitoring



Vision KL Divergence
Semantic Richness

Weighted Sum



LiDAR KL Divergence
Geometric Reliability

Outline

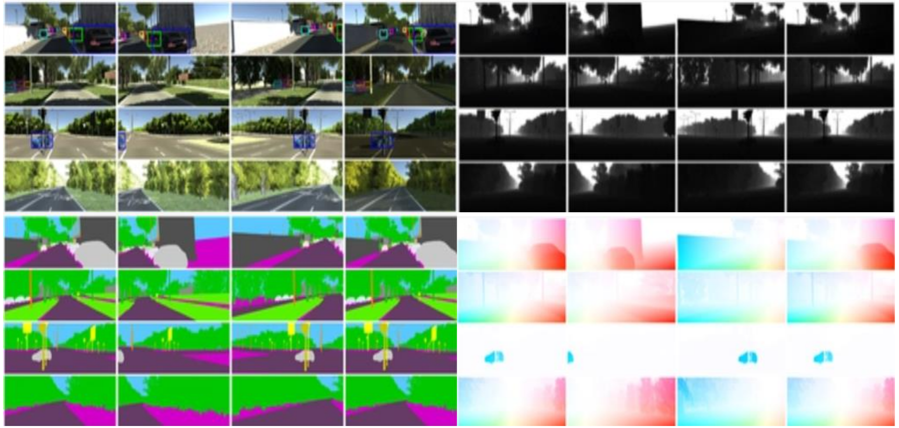
- Proposed Framework
  - Vision Module
  - LiDAR Module
  - Consistency Monitoring and Sensor Fusion
- **Experiments**
  - **Virtual KITTI dataset setup**
  - **Key experimental parameters, metrics and baselines**
- Key Results
  - Selected change detection accuracy results for individual se
    modalities
  - Selected sensor fusion results on adverse weather conditions

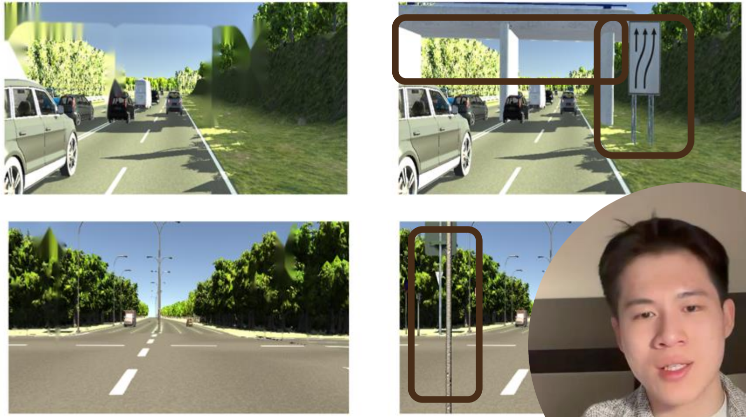
Virtual KITTI Dataset[10]

- Pixel-level ground truth
- Stress-test in controlled conditions
- Multiple object categories
- Several sequences for fair evaluation

Modification

Objects removed programmatically to simulate map-change



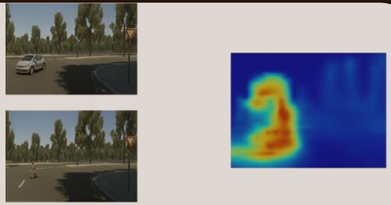


Baselines and Metrics

Baselines

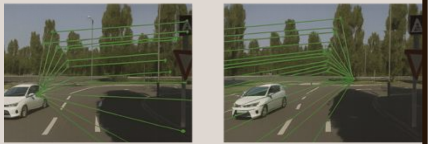
Contrastive Language-Image Pretraining

CLIP[11] : Patch-difference change maps using ViT-B/32 embeddings.



Local Feature Transformer

LoFTR[12] : Dense local feature matching with transformer



Jaccard Distance[14]: Voxel overlap metric for LiDAR maps

Fusion: Weighted Sum of Vision and LiDAR Scores

Metrics

KL divergence[8] (↓)

v.s. ground-truth change map

Pearson correlation[13] (

spatial agr

[8] Kullback, S., & Leibler, R. A. The Annals of Mathematical Statistics [11] Radford, A., et al. ICML [12] Sun, J., et al. *CVPR* [13] Pearson, K., 1896 [

Evaluation Questions

How well do the predicted anomaly distributions align with ground-truth changes induced by simulated infrastructure removal?

How accurately can each individual modality detect semantic changes in the map under normal and degraded conditions?

Can fusing information from Vision and LiDAR improve map-ch
detection in diverse conditions?

Outline

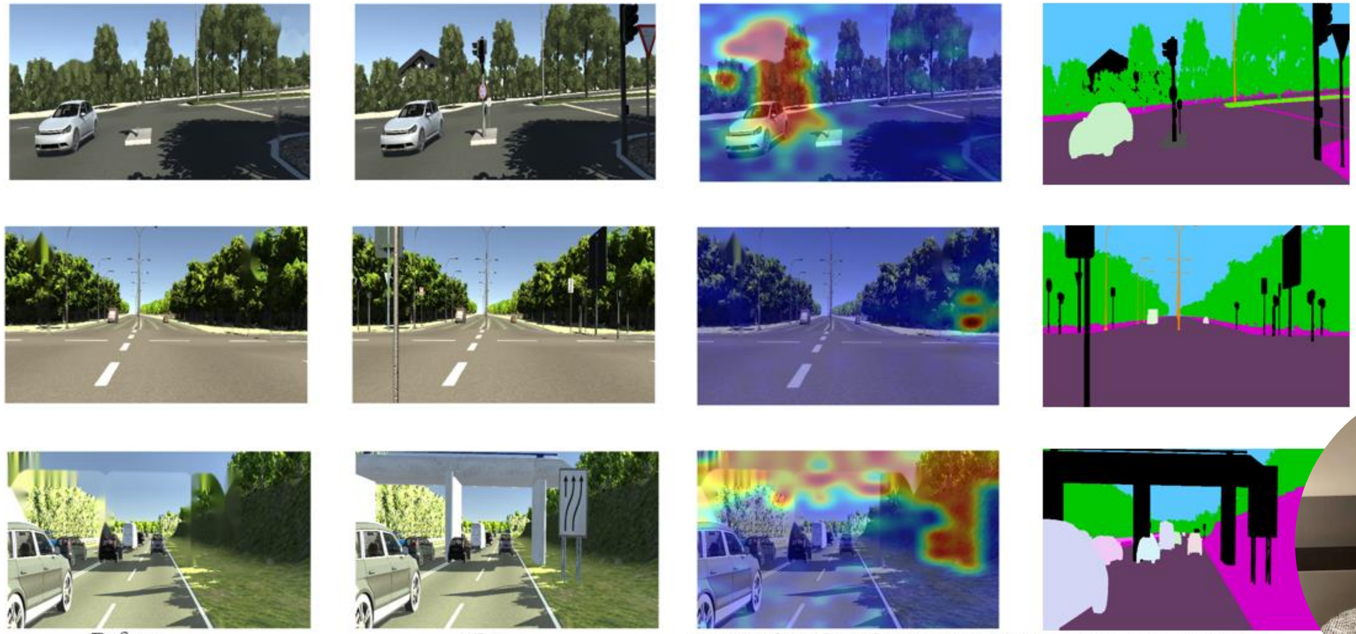
- Proposed Framework
  - Vision Module
  - LiDAR Module
  - Consistency Monitoring and Sensor Fusion

- Experiments
  - Virtual KITTI dataset setup
  - Key experimental parameters, metrics and baselines

- **Key Results**
  - **Selected change detection accuracy results for individua modalities**
  - **Selected sensor fusion results on adverse weather condit**

Vision-Only Alignment with Ground-Truth Changes

DINOv2 + segmentation captures semantic differences from missing or changed infrastructure.



Before After DINOv2-based patch-pattern contrast Semantic differences

Per-Modality Accuracy in Detecting Semantic Changes

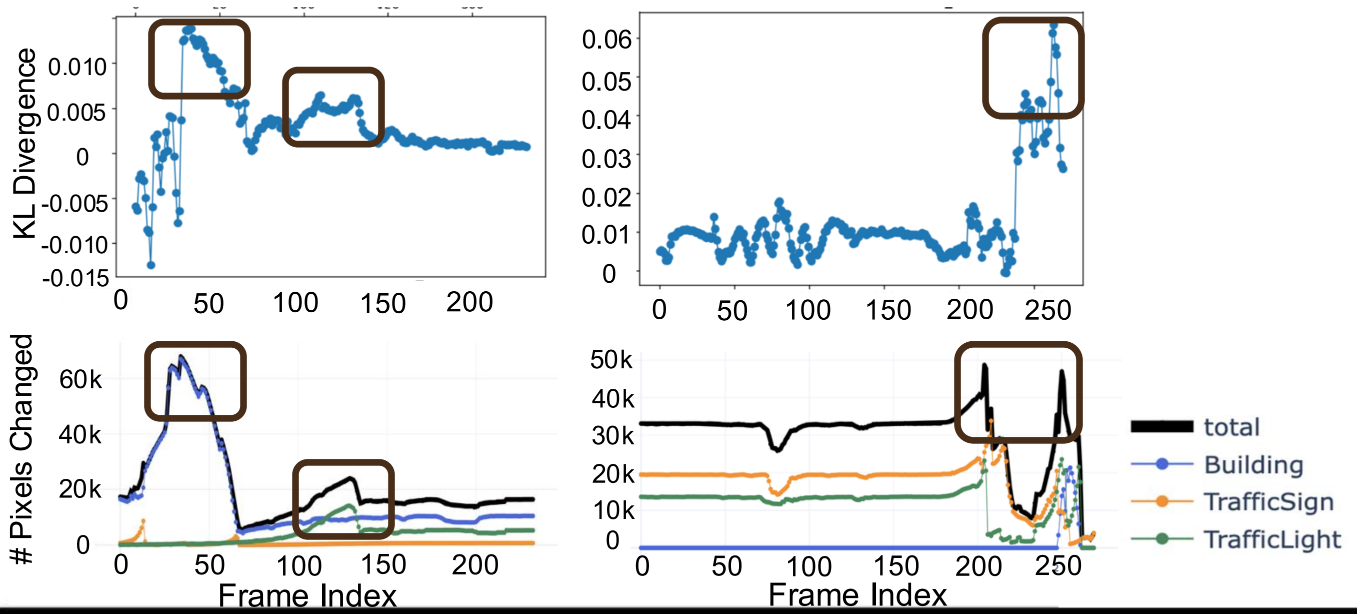
Our Vision Module method achieves 95% overall True Positive Rate vs. ~60 – 75% for baselines.

Category	Ours	CLIP[11]	LoFTR[12]
Building	84.8	60.3	55.4
Traffic Light	83.9	60.4	50.8
Traffic Sign	81.6	60.4	48.1
Overall	**95.0**	75.0	63.8

[11] Radford, A., et al. *International Conference on Machine Learning.* [12] Sun, J., et al. *IEEE/CVF Conference on Computer Vision and Pattern Rec*

Per-Modality Accuracy in Detecting Semantic Changes

Our LiDAR Module method shows KL divergence peaks fairly aligning with true map changes.



Point Cloud KLD

Tr
C

Fusion Preserves Robustness in Adverse Conditions

Our fusion method maintains strong alignment with ground truth under rain and fog, while baselines degrade sharply.

Normal Condition	Ours	CLIP[11] + Jaccard [14]	LoFTR[12] + Jaccard [14]
KL Divergenc e[8] (↓)	**0.11**	0.63	0.52
Pearson Corr.[13] (↑)	**0.72**	0.38	0.15

Rainy Condition	Ours	CLIP[11] + Jaccard [14]	LoFTR[12] + Jaccard [14]
KL Divergenc e[8] (↓)	**0.13**	0.89	0.73
Pearson Corr.[13] (↑)	**0.68**	0.37	[illegible]

[8] Kullback, S., & Leibler, R. A. The Annals of Mathematical Statistics [11] Radford, A., et al. ICML [12] Sun, J., et al. *CVPR* [13] Pearson, K., 1896

Conclusion

- Our sensor fusion framework with KL divergence-based scoring achieves high performance under normal conditions and maintains it in adverse weather.

- Real-time anomaly detection with spatial heatmaps can provide autonomous systems with change alerts and accurate localization, addressing the critical gap between static maps and dynamic urban environments for safer navigation.

- The integration of large vision-language models can enable t detection of novel infrastructure changes without requiring ret

Thank you!
Acknowledgements: MADD Lab

https://sites.google.com/g.hmc.edu/madd-lab/home

About Me

Full Paper