Multimodal Radio and Vision Fusion for Robust Localization in Urban V2I Communications

From Single to Multi-User Scenarios

Can Zheng, Jiguang He, Chung G. Kang, Guofa Cai, Henk Wymeersch

April 15, 2026

Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

Extended Works

- Multi-User Case
- Preliminary Simulation Results

Conclusion and Future Work

- Summary
- Conclusion
- Future Work

Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

Extended Works

- Multi-User Case
- Preliminary Simulation Results

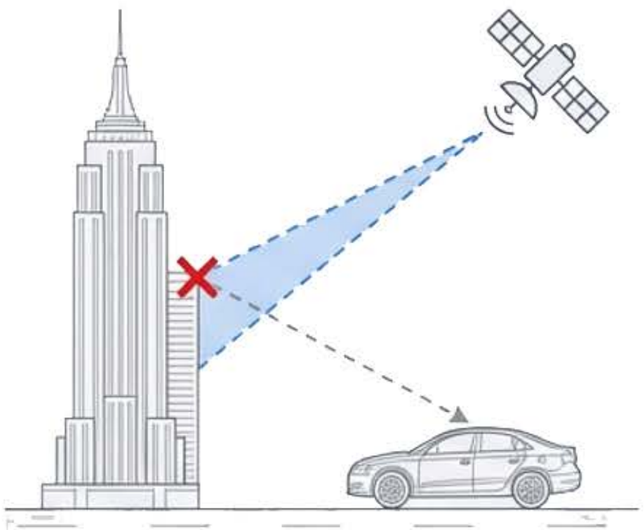
Conclusion and Future Work

- Summary
- Conclusion
- Future Work

Research Background

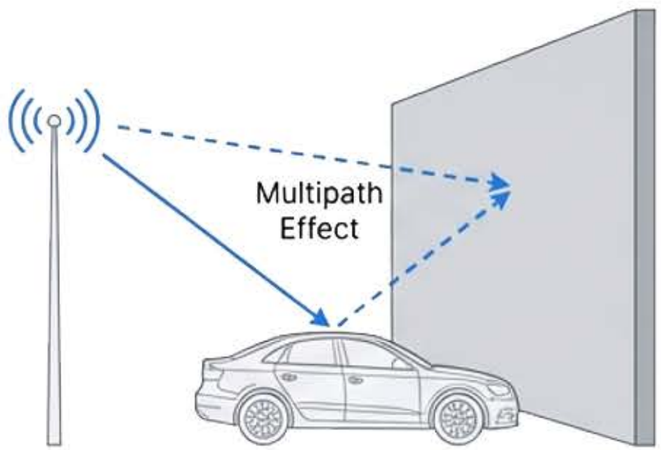
- The Core Need: High-accuracy localization is the foundation for autonomous driving and smart city infrastructure.
- The Challenge: In "urban canyons" (streets with tall buildings), GPS signals are easily blocked or reflected, causing large positioning errors.

GPS Failure



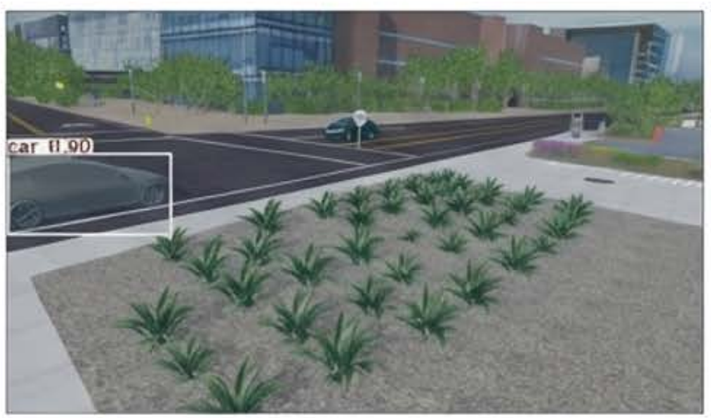
High-rise obstructions block LoS to satellites, rendering GPS unreliable or unavailable in dense urban area.

Radio Failure



Distorted by multipath interference and clutter.

Vision Failure

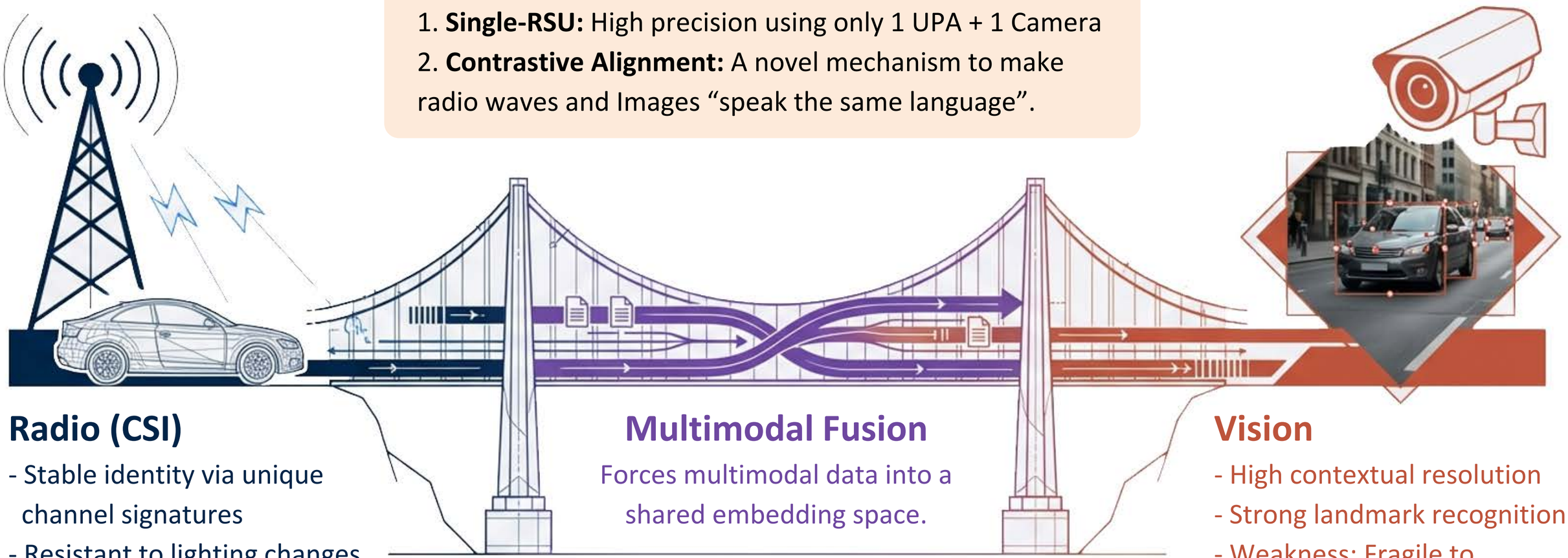


Blinded by obstruction, lighting, or distance.

The Solution: “Eyes”+ “Ears”

Core Contributions

1. **Single-RSU:** High precision using only 1 UPA + 1 Camera
2. **Contrastive Alignment:** A novel mechanism to make radio waves and Images “speak the same language”.



Radio (CSI)

- Stable identity via unique channel signatures
- Resistant to lighting changes
- Weakness: Noisy in multipath

Multimodal Fusion

Forces multimodal data into a shared embedding space.

Vision

- High contextual resolution
- Strong landmark recognition
- Weakness: Fragile to blockage/weather

Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

Extended Works

- Multi-User Case
- Preliminary Simulation Results

Conclusion and Future Work

- Summary
- Conclusion
- Future Work

System Model

Signal Model

A road side unit (RSU) is equipped with a single camera and a uniform planar array (UPA) antenna.

- The channel response associated with the *m*-th antenna and *k*-th subcarrier:

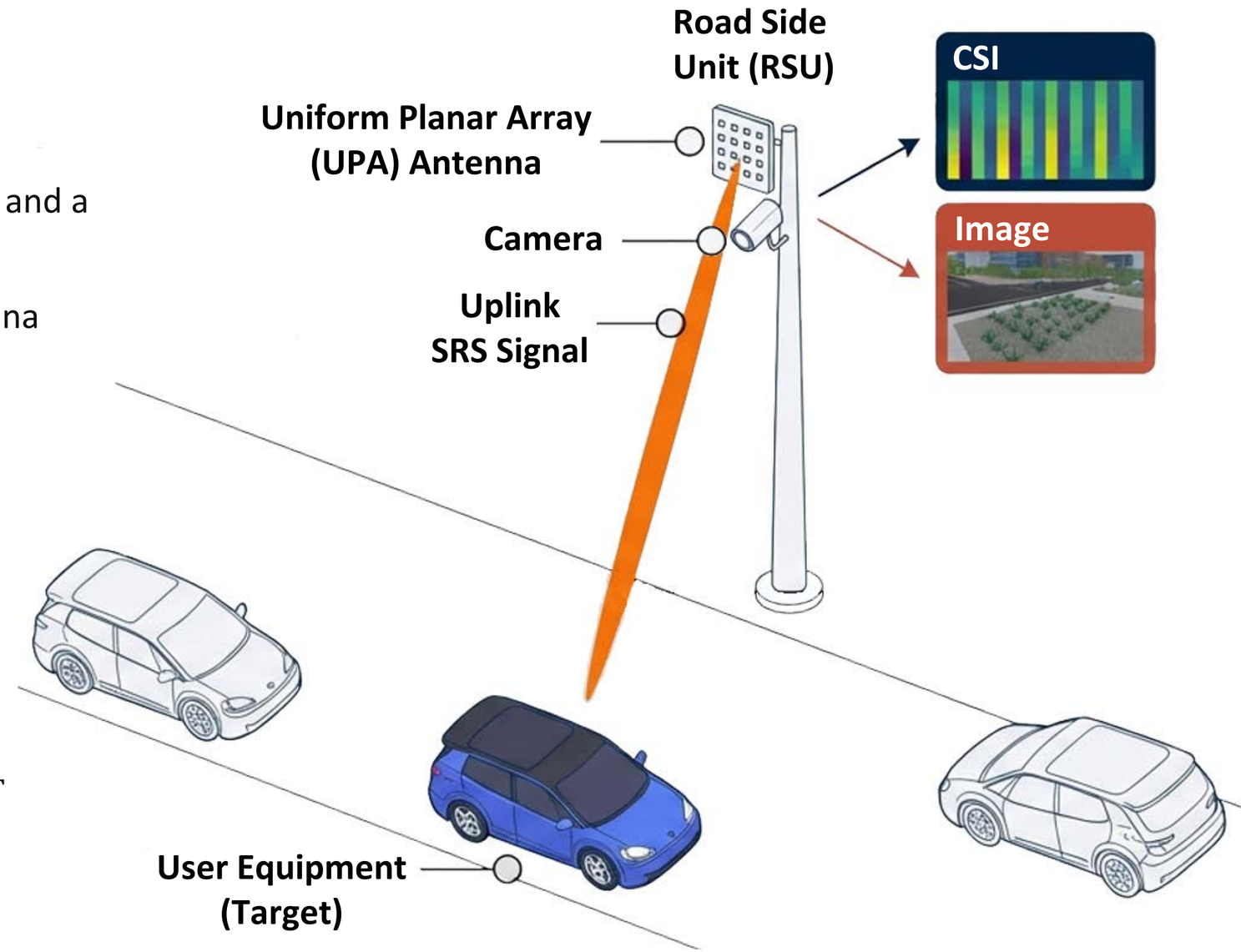
Complex Gain — Time of Arrival (ToA)

$$\hat{h}_{m,k} = \sum_{l} \alpha_l \exp(-j2\pi(f_c + k\Delta f)\tau_l) a_m(\theta_l, \phi_l)$$

Multipath — Steering Vector

The Objective

- **Goal:** Estimate UE 3D position $\mathbf{p}_{\mathrm{UE}} = (x_{\mathrm{UE}}, y_{\mathrm{UE}}, z_{\mathrm{UE}})^{\mathrm{T}}$
- **Input:** CSI Matrix, Image



Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

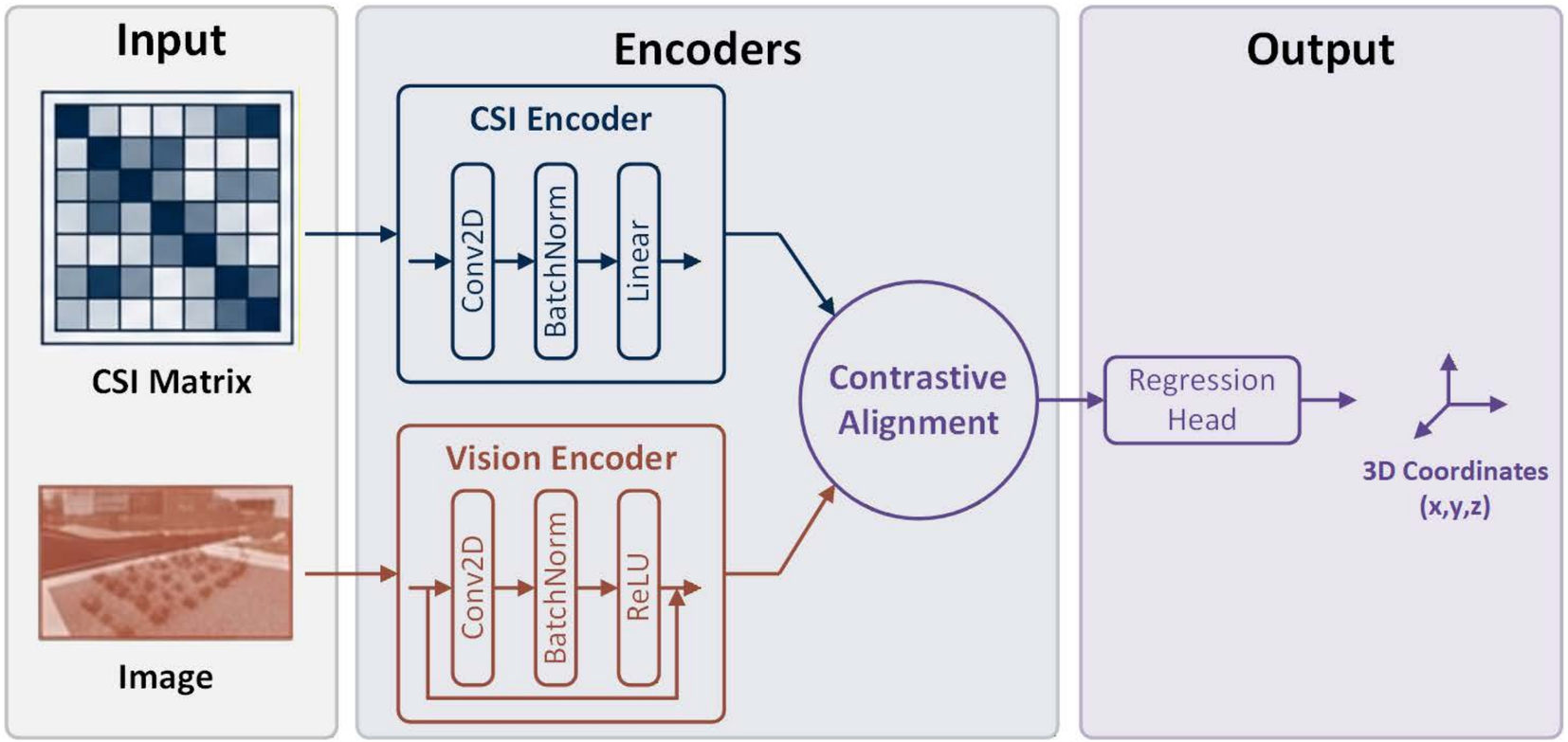
Extended Works

- Multi-User Case
- Preliminary Simulation Results

Conclusion and Future Work

- Summary
- Conclusion
- Future Work

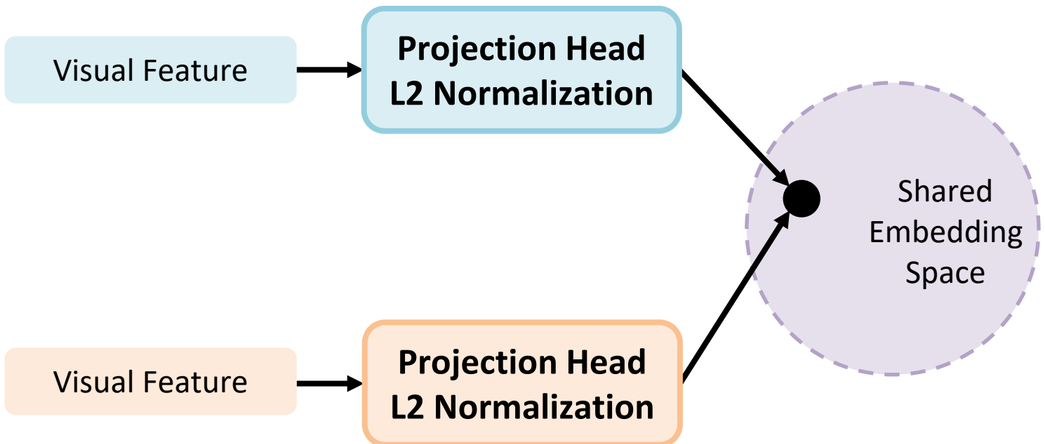
Multimodal Fusion Architecture (1)



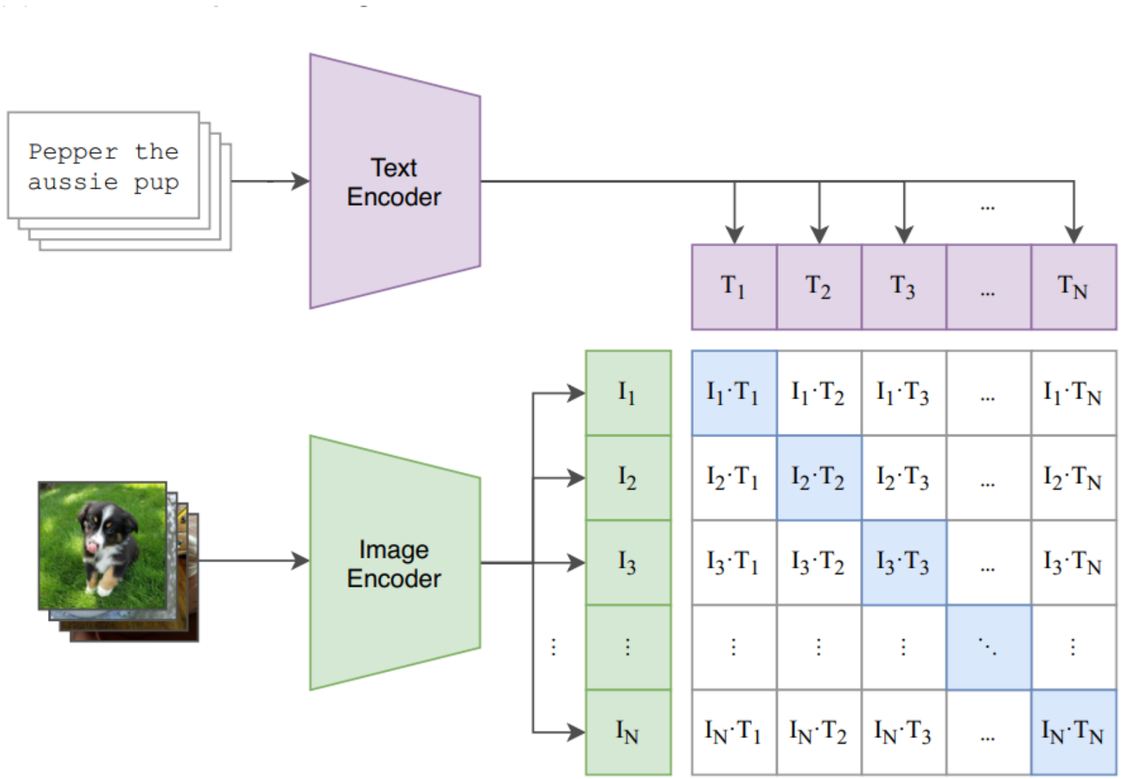
Three main stages

- Dual-Stream Extraction: Extract features from both image and radio data.
- Cross-Modal Alignment: Map them to the same space.
- Fusion & Regression: Join them together to predict 3D coordinates.

Multimodal Fusion Architecture (2)



The Semantic Gap: The pixels and wireless signals “speak” different languages.
Solution: Features from the exact same physical location are forced to align mathematically.



Figs: Summary of CLIP model’s approach.

Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *International conference on machine learning*. PmLR, 2021.

Joint Optimization

- **Localization loss:** $\mathcal{L}_L = \frac{1}{B}\sum_{i=1}^{B} \|\mathbf{p}_{\mathrm{UE}}^{(i)} - \hat{\mathbf{p}}_{\mathrm{UE}}^{(i)}\|_2^2$
- **Contrastive loss:** $\mathcal{L}_C = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})}\right).$
- **Total loss:** $\mathcal{L}_{\mathrm{Total}} = \frac{1}{2\exp(s_C)}\mathcal{L}_C + \frac{1}{2\exp(s_L)}\mathcal{L}_L + \frac{1}{2}(s_C + s_L).$

Kendall, Alex, Yarin Gal, and Roberto Cipolla. "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics." *Proceedings of the IEEE conference on computer vision and pattern recognition (ICLR)*. 2018.

Adaptive Weighting

- Dynamically balances the two losses during training based on task uncertainty.
- Prevents one objective from dominating the network.



Localization Loss (MSE)
Optimize the 3D coordination prediction accuracy.

Contrastive Loss
Optimize the 3D coordination prediction accuracy.

Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

Extended Works

- Multi-User Case
- Preliminary Simulation Results

Conclusion and Future Work

- Summary
- Conclusion
- Future Work

Simulation Settings

Environment:

Realistic outdoor urban city street with moderate vehicular traffic.

Hardware Parameters:

- Carrier frequency: 60 GHz
- Antennas: 16 elements (4×4 UPA)
- Bandwidth: 50 MHz
- Subcarriers: 8 of 512

Baselines:

- Traditional: MUSIC and ToA
- CSI-only
- Vision-only (Genie-aided)





Figs: DeepVerse 6G DT 31 Scenario.

https://deepverse6g.net/

Simulation Results (1)

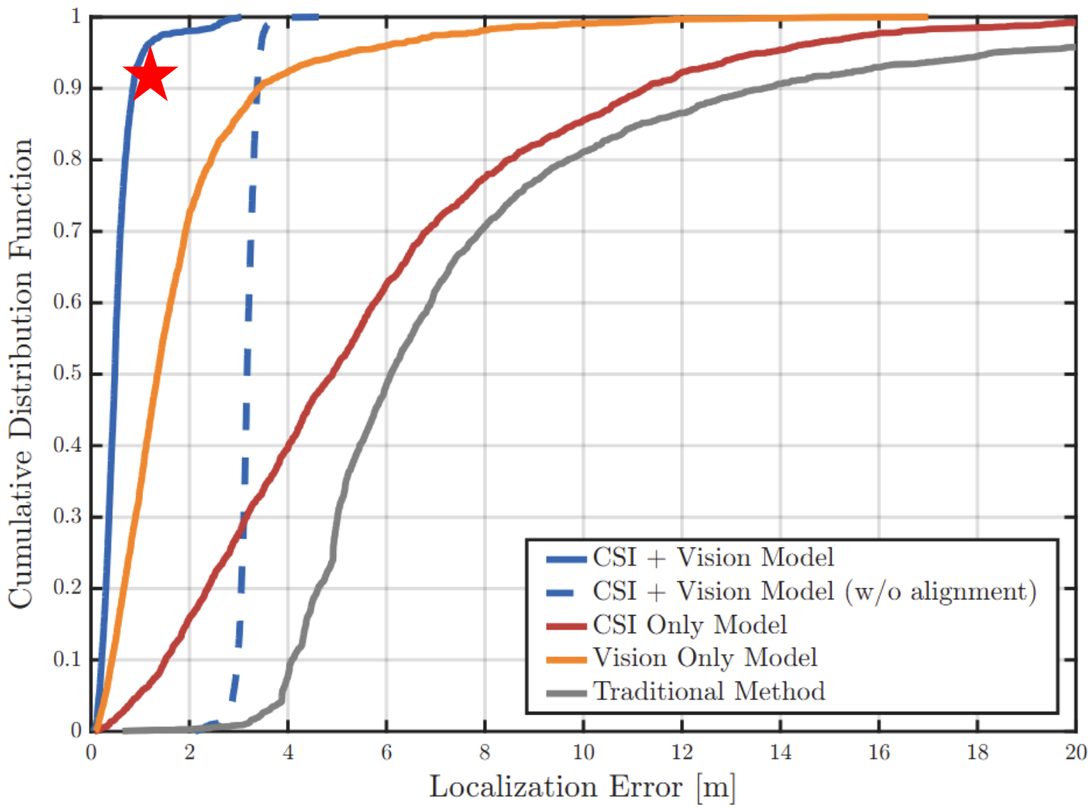


Fig: CDF of the localization error for different methods.

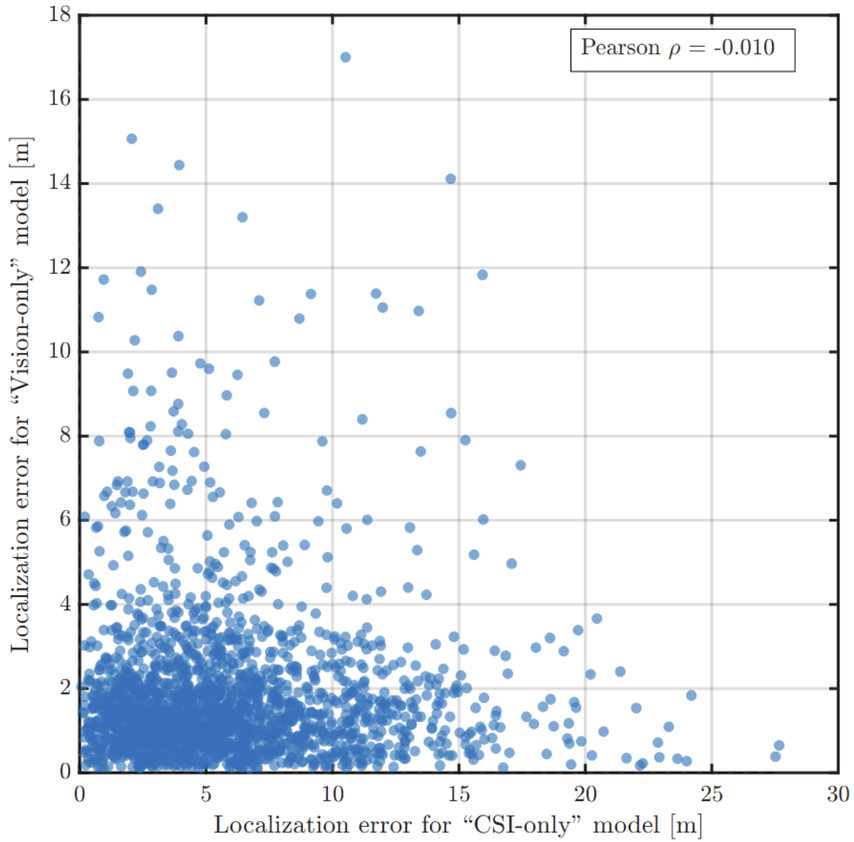


Fig: Localization error scatter point for single modal methods.

Superior Accuracy: 0.55 meters average error

High Reliability: 95% percentage - 1.06 meters

Complementary Nature: Error analysis shows CSI and vision errors are uncorrelated, perfectly covering each other's blind spots.

Simulation Results (2)

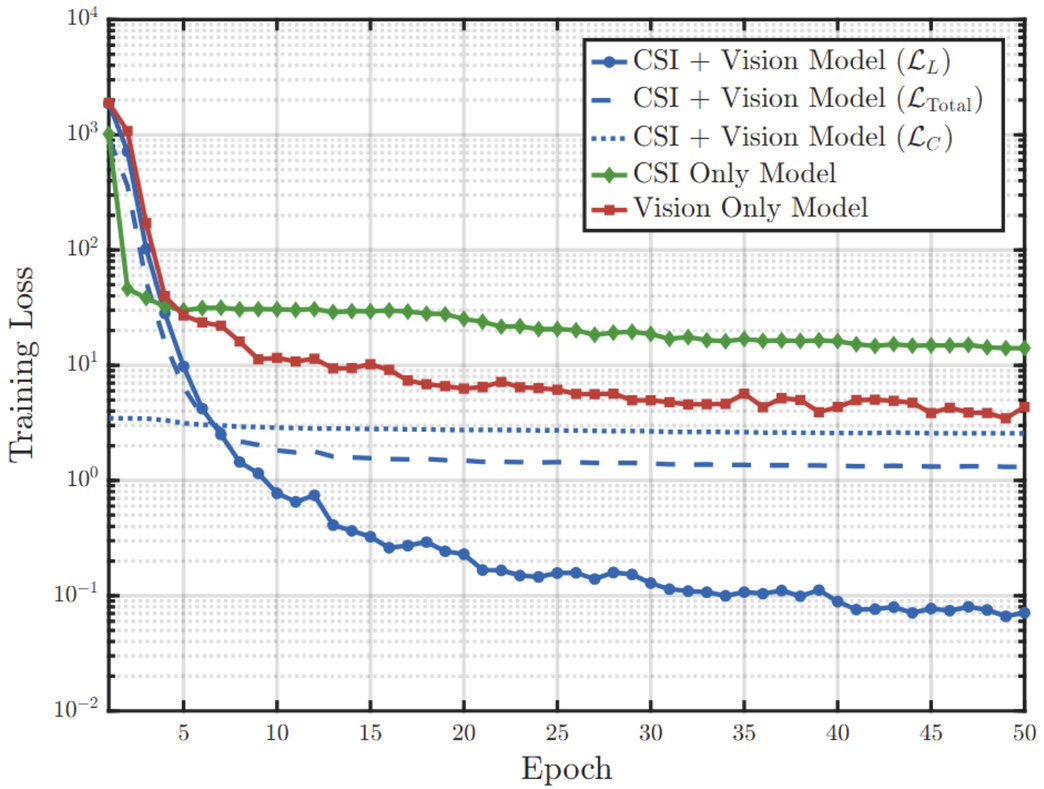


Fig: The curves of the training loss for different models.

Method	# Parameters	GFLOPS
Traditional	0	~0.001
CSI-only	152 K	~0.0081
Vision-only	11.2 M	1.82
Proposed	11.3 M	1.82

Table: Comparison of the network parameters and the inference cost per test sample.

Vision processing dominates the computational cost.

Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

Extended Works

- Multi-User Case
- Preliminary Simulation Results

Conclusion and Future Work

- Summary
- Conclusion
- Future Work

Multi-User Case

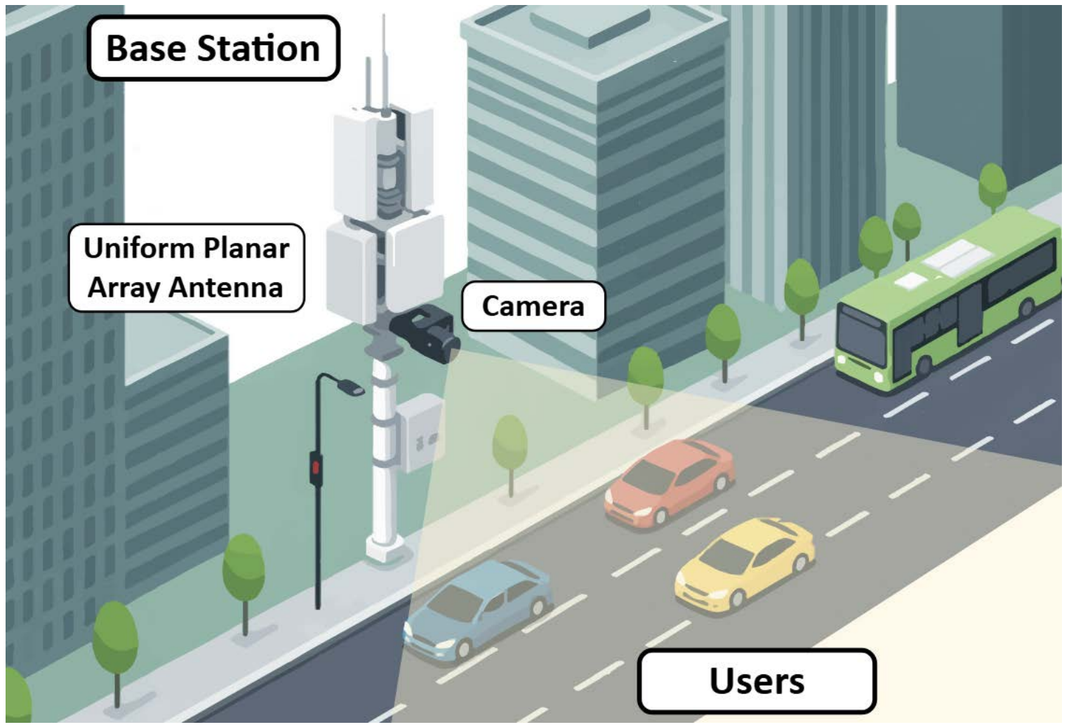


Fig: Illustration of the multi-user V2I multimodal localization scenario.

The Next Frontier:

Transitioning to dynamic, multiuser environments.

The Association Problem

- No natural one-to-one mapping between visual targets and communication users.
- There may be multiple vehicles in a single image. How to determine which CSI corresponds to which vehicle?

Our Solution:

Feature-level association, rather than matching.

Preliminary Simulation Results (1)

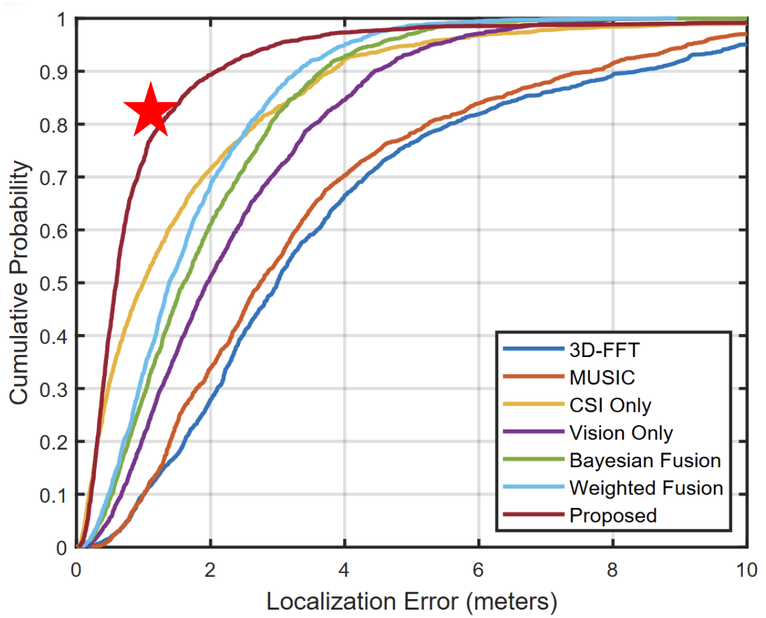


Fig: CDFs of the localization errors for all compared methods.

Maintains highly precise localization.

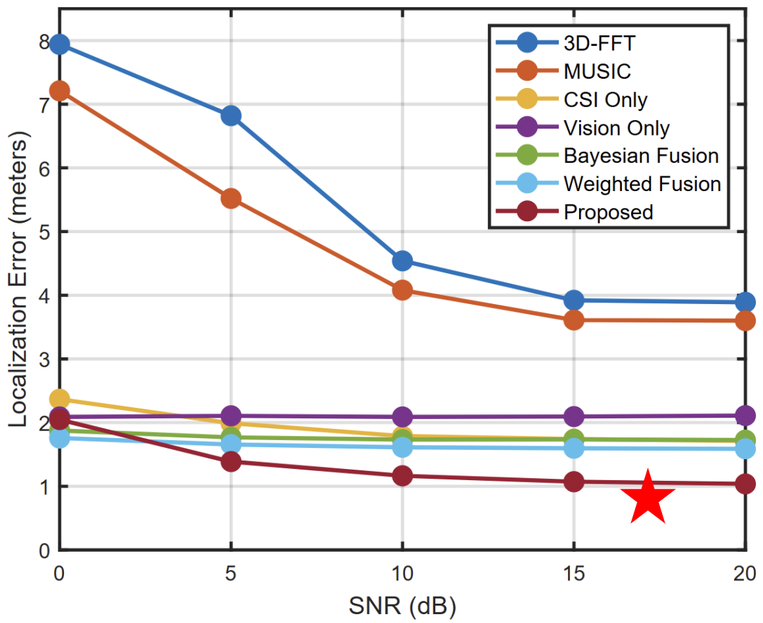


Fig: Mean localization error versus channel SNR for different methods.

Maintains highly precise localization.

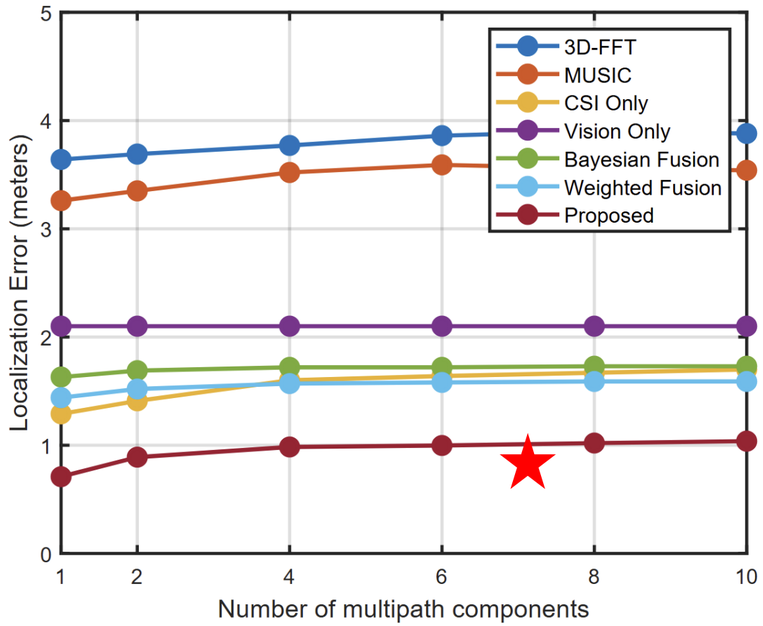


Fig: Mean localization error as a function of the number of multipath components for different methods.

Preliminary Simulation Results (2)



(a) Single user: Input image.



(b) Single user: Query 1.



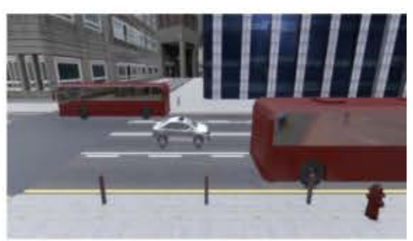
(c) Two users: Input image.



(d) Two users: Query 1.



(e) Two users: Query 2.



(f) Four users: Input image.



(g) Four users: Query 1.



(h) Four users: Query 2.



(i) Four users: Query 3.



(j) Four users: Query 4.

Fig: Cross-attention heatmaps under different numbers of active users.

Presentation Agenda

Background and Motivation

- Research Background
- The Solution

System Model

Proposed Method

- Multimodal Fusion Architecture
- Joint Optimization

Simulation Results

- Simulation Settings
- Simulation Results

Extended Works

- Multi-User Case
- Preliminary Simulation Results

Conclusion and Future Work

- Summary
- Conclusion
- Future Work

Conclusion and Future Work

Summary

Comparison Dimension	Work 1	Work 2
Task	Single-user localization	Multi-user localization
Core Problem	How to fuse CSI and vision	How to establish an instance-level correspondence between CSI and visual targets
Main Conclusion	Multimodal fusion is useful	Feature-level association should be performed when there are multiple users, not just late fusion.

Conclusion

- Combining vision and CSI from a single RSU effectively overcomes the limits of GPS and traditional wireless localization.
- Cross-modal contrastive alignment is the key to improving localization performance.
- We have successfully extended this work to complex multi-user environments.

Future Work

- **More Sensors:** We plan to expand the framework for full 4D+ environmental sensing. This will include incorporating radar, LiDAR, IMU, and inaccurate GPS data.
- **Trajectory Tracking:** We aim to extend our focus from instant localization to continuous and predictive trajectory mapping.