TMF-Net: Multi-modal Transformer Fusion for Relative Pose Estimation of Non-Cooperative Targets

Bhumika Makwana | Co-Founder, Pax Orbital

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION 2026 VIENNA

Objective

Enable reliable vision-based navigation for autonomous operation on increasingly agile, low size, weight, power, and cost (SWaP-C) space platforms.

Challenges

- **Extreme Space Environments:** Illumination variations, high-contrast shadowing, and complex Earth Albedo.
- **Prohibitive mass, power, and computational overheads** unsuitable for agile, low-resource satellite platforms.
- **Scarcity of real flight data.**

Key Idea

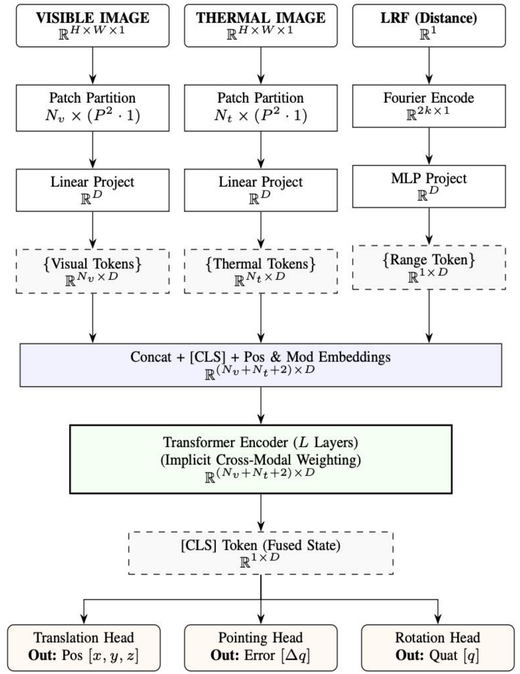
Fuse three complementary sensors in one latent space:

- Visible - Spatial orientation and texture.
- Thermal (TIR) - Target silhouette when optical data fails.
- 1D Laser Rangefinder - Absolute scale, lifted via Fourier Feature Encoding enabling correlation of 1D range with 2D cues.

Real-Time Stream Prioritization: The network adaptively prioritizes the most physically reliable sensor modality based on environmental conditions.

Methodology

Heterogeneous input $S = \{I, Ith, r\}$: visible image, thermal image, and scalar LRF range, mapped to a shared latent space RD where cross-modal dependencies are modeled by self-attention.



Range is encoded into frequency space so the Transformer can resolve fine scale:

$$\gamma(r) = \cos(2\pi Br), \sin(2\pi Br)^\top$$

A weighted multi-task loss supervises all heads, with the pointing term scaled by $1/rgt$ for tighter lock in the final approach metres:

$$Ltotal = \lambda 1 Ltrans + \lambda 2 Lrot + \lambda 3 Lpoint$$

Dataset Generation

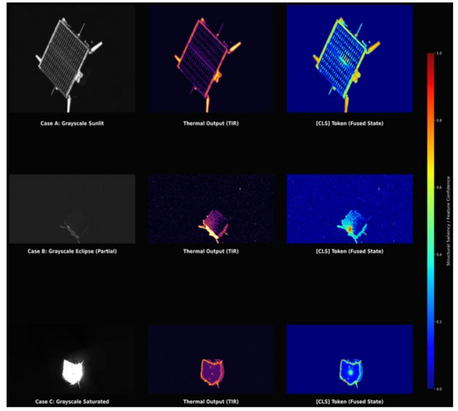
High-fidelity multi-modal data for the **Tango** satellite (PRISMA mission):

- Visible & TIR rendered in **Blender** over diverse proximity-operation trajectories.
- Surface temperatures from **MATLAB PDE Toolbox** (1D heat conduction) for realistic thermal output.
- LRF ray-cast to the centroid, with injected Gaussian noise and *null/miss returns* to train the pointing head.

(Case A) Sunlit - Attention spread across high-contrast structure.

(Case B) Eclipse - Visual falls below noise floor network shifts to thermal to recover the silhouette.

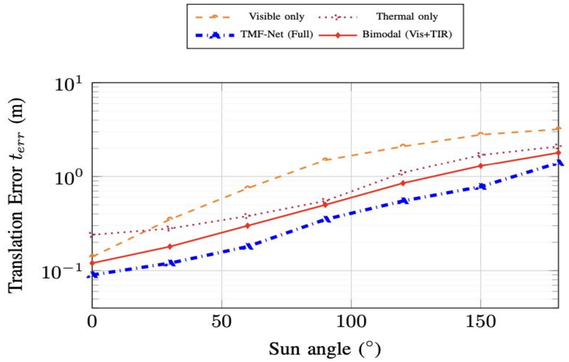
(Case C) Saturated - Ignores the blown-out visible signal, anchors to thermal edges.

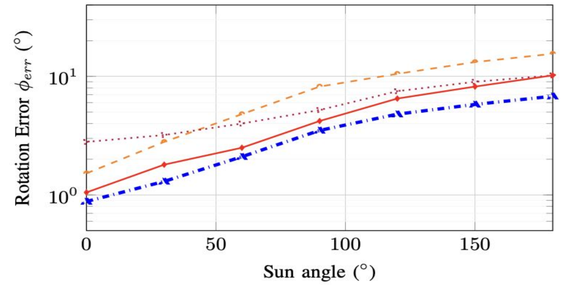


Saliency maps generated using the Attention Rollout method.

Experimental Results

Scenario	Visual Baseline		Bimodal (Vis+TIR)		TMF-Net (Full)	
	t_{err} (m)	ϕ_{err} (°)	t_{err} (m)	ϕ_{err} (°)	t_{err} (m)	ϕ_{err} (°)
Case A (Sunlit)	0.14	1.52	0.12	1.05	**0.09**	**0.88**
Case B (Eclipse)	1.21	12.4	0.92	9.80	**0.24**	**5.35**
Case C (Saturated)	2.85	4.10	1.97	2.15	**0.87**	**1.72**





89.3%
LRF sensor-lock success from the range-weighted pointing loss