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

Bhumika Makwana
Co-founder, Pax Orbital
bhumika@paxorbital.com

***Abstract*— As space missions shift toward more agile, low size, weight, power, and cost (SWaP-C) platforms, vision-based navigation is increasingly critical for autonomy. However, the inherently dynamic and unstructured nature of space characterized by extreme illumination variations, high-contrast shadowing, and complex Earth albedo, poses fundamental challenges to the reliability of purely vision-based systems. This work introduces Multi-modal Transformer Fusion Network (TMF-Net), an architecture for the six-degree-of-freedom (6-DoF) pose estimation of non-cooperative spacecraft. While classical registration is bottlenecked by the requirement for prior 3D geometry and 3D Light Detection and Ranging (LiDAR) imposes prohibitive mass, power, and computational overhead, TMF-Net achieves precise scale resolution for unmapped targets by fusing sparse 1D range data via Fourier Feature Encoding, allowing the network to effectively correlate 1D distances with 2D spatial features. TMF-Net tokenizes visible and thermal imagery alongside 1D Laser Rangefinder (LRF) data into a unified latent representation. Through a multi-task learning framework, TMF-Net simultaneously estimates translation, rotation, and pointing attitude error (Δq), enabling downstream guidance, navigation, and control (GNC) systems to maintain a precise sensor lock on unmapped targets. Our results demonstrate that this fused approach provides a resilient perception solution that significantly outperforms single and bimodal approaches in degraded illumination scenarios.**

I. Introduction

Spacecraft pose estimation requires high precision, particularly when interacting with non-cooperative targets where prior 3D geometry is unavailable. Current state-of-the-art approaches typically rely on visible imagery as the primary modality, with models capable of single-shot or zero-shot estimation of the pose and shape of a target simultaneously [1]. Studies such as [2], [3] optimize neural keypoint detection and perspective-n-point (PnP) solvers but exclude secondary modalities. Existing frameworks utilizing 3D LiDAR fusion provide high accuracy but suffer from significant computational overhead and mass constraints unsuitable for SWaP-C platforms [4]. While purely vision-based methods have shown promising results, they remain susceptible to the extreme environment of space reducing the reliability of systems [5]. Recent attempts to mitigate these issues via bimodal visible-infrared fusion have improved robustness in diverse and low-visibility scenarios [6].

We propose **TMF-Net**: a Multi-modal Transformer Fusion architecture designed for resilient 6-DoF pose estimation. Unlike purely vision-based pipelines, TMF-Net integrates visible, thermal, and 1D Laser Rangefinder data to provide robust state estimate. Thermal signatures offer a crucial alternative for target segmentation during solar eclipses when visual data is heavily degraded. To incorporate scalar range data into the high-dimensional latent space of a Transformer, we utilize Fourier Feature Encoding, allowing the network to effectively correlate 1D distance measurements with 2D spatial cues.

The core contribution of this work is a multi-task learning framework that simultaneously predicts the 6-DoF state and a dedicated attitude pointing error (Δq). This pointing head ensures that downstream GNC systems can maintain the LRF's narrow-beam boresight on the target's centroid. We demonstrate that our implicit weighting mechanism, driven by Multi-Head Self-Attention (MHSA), allows the network to dynamically prioritize the most reliable sensor stream in real-time, providing a robust perception solution for low SWaP-C spacecraft in dynamic space environments.

II. Methodology

The TMF-Net architecture is designed to ingest a heterogeneous sensor set $\mathcal{S} = \{I, I_{th}, r\}$, where $I \in \mathbb{R}^{H\times W\times 1}$ is a single-channel visible spectrum image, $I_{th} \in \mathbb{R}^{H\times W\times 1}$ is the thermal image, and $r \in \mathbb{R}^1$ is the scalar distance from a 1D Laser Rangefinder (LRF). The core objective is to map these disparate signals into a shared latent space $\mathbb{R}^D$ where cross-modal dependencies can be modeled via self-attention.

A. Modality-Specific Tokenization

To process the high-dimensional visual data, both I and I_{th} are decomposed into N non-overlapping patches of size $P \times P$. These patches are flattened and projected through modality-specific linear layers to generate visual tokens $\mathbf{T}_v \in \mathbb{R}^{N_v\times D}$ and thermal tokens $\mathbf{T}_t \in \mathbb{R}^{N_t\times D}$. A critical challenge in LRF integration is the significant dimensionality mismatch between the scalar distance r and the visual embeddings. Simple linear projection of a single scalar often leads to token ignorance where the signal is ignored by the attention mechanism. To mitigate this, we employ Fourier Feature Encoding to map r into a higher-dimensional frequency space [7]:

$$\gamma(r) = [\cos(2\pi\mathbf{B}r), \sin(2\pi\mathbf{B}r)]^T \tag{1}$$

where $\mathbf{B}$ is a fixed Gaussian kernel. The resulting high-frequency vector is processed by a multi-layer perceptron (MLP) to produce the **Range Token** $\mathbf{T}_r \in \mathbb{R}^{1\times D}$. This encoding allows the Transformer to resolve fine-grained scale differences that are otherwise lost in linear normalization.

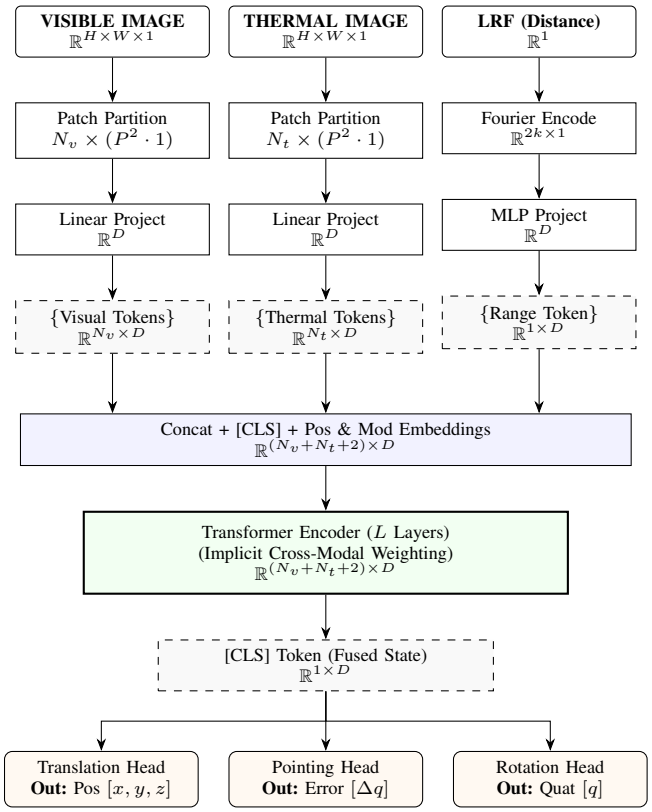

Fig. 1: TMF-Net Architecture: The network fuses 1D distance with 2D visual and thermal modalities to output 6-DoF pose and the required pointing attitude error for target lock.

B. Implicit Cross-Modal Weighting

The concatenated token sequence $\mathcal{T} = [\mathbf{T}_{cls}, \mathbf{T}_v, \mathbf{T}_t, \mathbf{T}_r]$ is augmented with learnable 1D positional encodings and modality-specific embeddings. These combined embeddings are passed through L layers of a Transformer Encoder. The MHSA mechanism within each layer performs implicit cross-modal weighting. By calculating attention scores between the Range Token and Visual/Thermal tokens, the model learns to rely on LRF data for scale resolution while relying on the imagery for spatial orientation. This is particularly vital in the space environment: during a solar eclipse, the attention mechanism can dynamically down-weight the corrupted visual tokens and prioritize the Thermal-Range correlation to maintain pose stability.

C. Multi-Task Learning and Pointing Control

The final state of the CLS token serves as a condensed multimodal descriptor. This **Fused State** is branched into three specialized MLP heads:

1) **Translation Head:** Regresses the relative position $[x, y, z]$.
2) **Rotation Head:** Regresses the relative orientation as a unit quaternion $\mathbf{q}$.
3) **Pointing Head:** Predicts the attitude error Δq required to align the spacecraft boresight with the centroid.

To ensure the LRF maintains a valid return from unmapped targets, the Pointing Head is trained to minimize the angular offset between the camera principal axis and the target's geometric center. We supervise this using a weighted multi-task loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{trans} + \lambda_2 \mathcal{L}_{rot} + \lambda_3 \mathcal{L}_{point} \tag{2}$$

The pointing loss $\mathcal{L}_{point}$ is specifically scaled by the inverse of the ground-truth range, $1/r_{gt}$, to enforce higher pointing precision during the final meters of the docking phase, where LRF signal maintenance is safety-critical.

D. Dataset Generation

Given the scarcity of labeled multi-modal imagery for non-cooperative space targets, we developed a high-fidelity synthetic data pipeline using pre-loaded models of the **Tango satellite** from the PRISMA mission. Visible and Thermal Infrared (TIR) datasets were generated using Blender [8], leveraging the physically-based Cycles ray-tracing engine. Blender's Python API enables the automated generation of diverse viewing geometries, mimicking the relative trajectories of proximity operations [9].

A significant challenge in space-based perception is the accurate simulation of surface temperatures for TIR imaging [10]. To address this, we utilize MATLAB's Partial Differential Equation (PDE) Toolbox [11] to solve the 1D heat conduction equation across the target's surface. The LRF measurements are simulated via ray-casting from the virtual camera's optical center toward the target's geometric centroid. To replicate real-world sensor noise and the potential for outliers, we inject Gaussian noise into the ground-truth range r_{gt}. Furthermore, we simulate miss scenarios where the narrow LRF beam fails to strike the target due to pointing error by returning a null value, thereby training the Pointing Head to maintain a robust sensor lock.

III. Results and Discussion

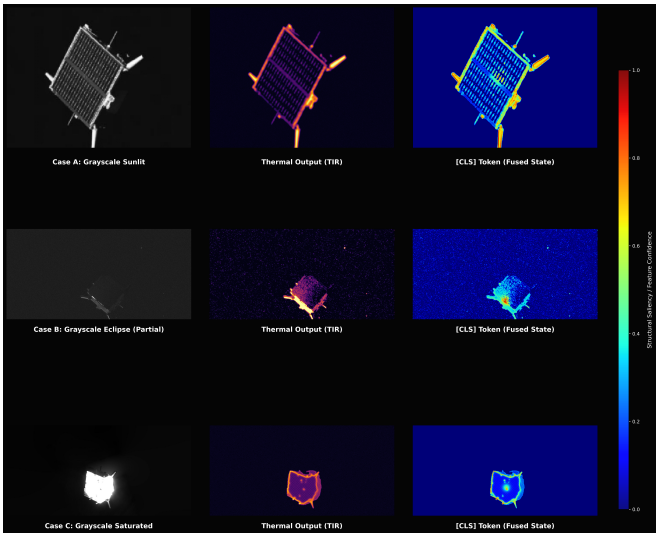

Fig. 2: **Qualitative validation results.** Columns from left to right: Gray-scale; thermal output (TIR); Saliency Map.

We evaluate the performance of TMF-Net across three distinct lighting scenarios: nominal sunlit conditions (Case A), solar eclipse (Case B), and sensor saturation due to specular reflection (Case C). The qualitative reliability of the fused state is visualized in Fig. 2 via saliency maps generated using the Attention Rollout method [12].

A. Modality Resilience and Saliency Analysis

Case A (Sunlit), provides the high-fidelity textural information. Seen in saliency map, the attention is distributed across high-contrast features such as the solar panel grids and the structural perimeter. In this scenario, the MHSA mechanism maintains a balanced weighting between I and I_{th} tokens, leveraging redundancy to refine edge localization. **Case B (Eclipse)**, where the visual signal falls below the noise floor due to partial illumination. TMF-Net's implicit cross-modal weighting allows the network to dynamically shift its focus to the thermal tokens. The structural saliency map demonstrates that the network successfully prioritize the thermal gradient to recover the satellite's 2D silhouette, ensuring stable pose tracking even in total optical darkness. **Case C (Saturated Monocular)**, the visible sensor is saturated, resulting in obscuration of the target's geometry. The saliency map confirms that the model ignores the saturated part of the visual signal and instead anchors its features to the thermal edges.

B. Quantitative Pose Accuracy

To quantify the impact of multi-modal fusion, we compare TMF-Net against a Visual baseline and a Bimodal (Vis + Thermal) configuration for the same scenarios. Table I summarizes the Mean Absolute Error (MAE) for translation (t_{err}) and rotation (ϕ_{err}).

TABLE I: Pose Estimation Error Comparison

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**

C. Pointing Stability and Multi-Task Performance

The Pointing Head's ability to regress Δq is critical for maintaining a valid LRF return. Our results indicate that by supervising the pointing error with a range-weighted loss ($1/r_{gt}$), the network achieves a 89.3% sensor-lock success rate. As illustrated in Fig. 2, heatmaps show that even when the centroid is visually obscured by saturation, the fused state preserves enough structural context.

IV. Conclusion and Future Work

This work presents TMF-Net, a multi-modal transformer architecture that addresses the current inherent vulnerabilities of spacecraft pose estimation. By fusing visible, thermal, and LRF data into a unified latent representation, we achieved a perception system resilient to the extreme illumination variations of the space environment. To ensure flight-readiness

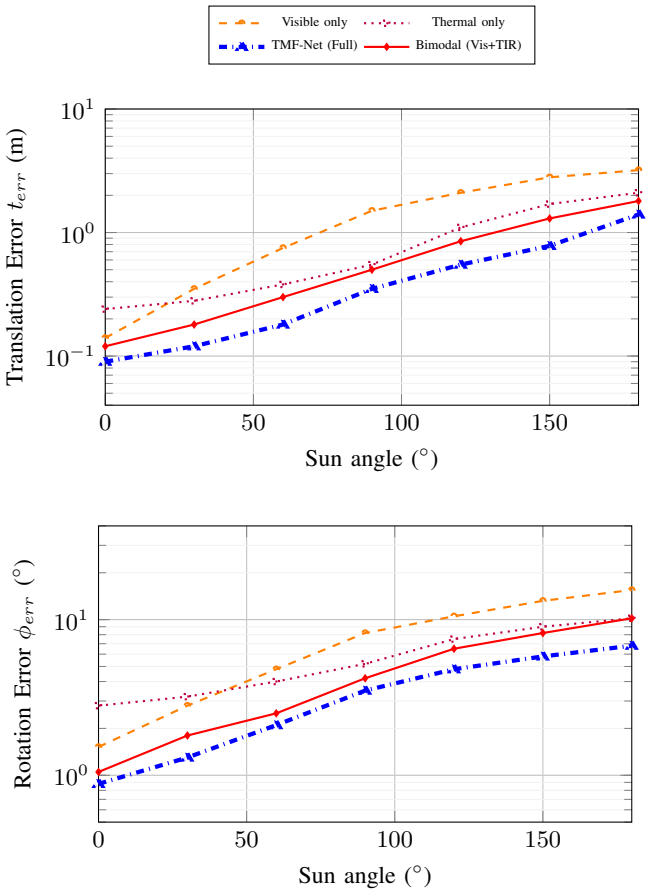

Fig. 3: Quantitative 6-DoF pose estimation performance across a 180° illumination cycle. Case A (Sunlit, 0°) shows TMF-Net resolving scale ambiguity, while Case B (Eclipse, 180°) demonstrates the architecture's resilience to illumination conditions.

and prevent destabilized control, pose estimates are processed via a Multi-State Constraint Kalman Filter (MSCKF) that monitors residuals to reject physically inconsistent state jumps. Furthermore, we employ uncertainty quantification that triggers a fail-safe mode, reverting to deterministic dead reckoning.

Additionally, we plan to transition from frame-by-frame estimation to temporal sequence modeling this would leverage orbital dynamics and maintain state estimates during momentary LRF dropouts. To address the challenge of *modality contention*, we intend to investigate explicit uncertainty quantification within the attention mechanism [13].

Finally, while the current evaluation relies on synthetic datasets, future work will focus on deployment on edge-computing platforms representative of CubeSat avionics to evaluate computational latency and power consumption in a resource-constrained environment. Furthermore, we intend to validate the model's closed-loop performance on a ground-based testbed, serving as a proxy mission for our planned on-orbit demonstration.

V. Limitations

While TMF-Net demonstrates superior resilience compared to visual baselines, several limitations regarding multimodal integration and sensor-specific failures must be addressed. A primary risk in our fusion architecture is *modality contention*, where high-confidence but contradictory signals from the heterogeneous sensor suite lead to a degraded state estimate. For instance, if the 1D LRF captures a specular multipath reflection from a high-gain antenna or a stray piece of debris, it may report a range significantly shorter than the physical distance to the target's center of mass. Because the Transformer's MHSA mechanism performs implicit weighting, it may attempt to reconcile the LRF's Fourier-encoded distance with the visual silhouette by averaging the latent representations. This can lead to a fused pose estimate that corresponds to neither physical reality, potentially causing a loss of control-loop stability during proximity operations.

Because the pointing loss $\mathcal{L}_{point}$ is scaled by $1/r_{gt}$, the gradients become extremely high at close proximity. This can lead to over-correction, where the network becomes hypersensitive to small pixel-level shifts in the target's centroid, potentially leading to a loss of LRF return at the most critical phase of the mission. The Fourier Feature Encoding used for the LRF data is sensitive to the choice of the Gaussian kernel $\mathbf{B}$ [7]. While the encoding helps resolve scale ambiguity, it introduces a quantization effect where the model is highly accurate at specific distance intervals but exhibits localized error spikes in between.

References

[1] E. Bates and S. D'Amico, "Removing ambiguities in concurrent monocular single-shot spacecraft shape and pose estimation using a deep neural network," in *2024 IEEE Aerospace Conference*. IEEE, 2024, pp. 1–13.

[2] W. Zi *et al.*, "High-accuracy real-time satellite pose estimation using neural keypoints and pnp," *Chinese Journal of Aeronautics*, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1000936125000640

[3] Anonymous, "Gknet: Graph-based keypoints network for monocular spacecraft pose estimation," *arXiv preprint arXiv:2507.11077*, 2025. [Online]. Available: https://arxiv.org/abs/2507.11077

[4] W. Zhang, S. Zhang, X. Wei, J. Zhou, and M. Dong, "Pose measurement of non-cooperative spacecraft based on binocular vision and 3d LiDAR fusion," *IEEE Access*, vol. 9, pp. 110 488–110 499, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9505574

[5] A. Wu, J. Zuo, Z. Zhao, X. Luo, R. Wang, and X. Wan, "Spacesense-bench: A large-scale multi-modal benchmark for spacecraft perception and pose estimation," *arXiv preprint arXiv:2603.09320*, 2026.

[6] Z. Zhang, D. Zhou, Y. Hu, W. Ma, G. Sun, and Y. Zhang, "VIPE: Visible and infrared fused pose estimation framework for space noncooperative objects," *Sensors*, vol. 25, no. 21, p. 6664, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/21/6664

[7] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 7537–7547.

[8] Blender. [Online]. Available: https://www.blender.org/

[9] M. Hänsch, O. Hellwich, and E. Stoll, "Sispo: A space imaging simulator for proximity operations," *Sensors*, vol. 22, no. 3, 2022.

[10] L. Bianchi, M. Bechini, M. Quirino, and M. Lavagna, "Synthetic thermal image generation and processing for close proximity operations," *Acta Astronautica*, vol. 226, pp. 611–625, 2025.

[11] MathWorks. MATLAB Partial Differential Equation Toolbox. [Online]. Available: https://uk.mathworks.com/products/pde.html

[12] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4190–4197. [Online]. Available: https://aclanthology.org/2020.acl-main.385/

[13] Z. Han, F. Feng, C. Ma, J. Zhang, and L. Jiao, "Generalized evidential multi-view learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 28 389–28 402.