**Multi-Modal Battlefield Anomaly Detection for AI-Enabled Situational Awareness**

Modern operational environments generate massive volumes of heterogeneous battlefield data, including GEOINT imagery, intelligence reports, operational logs, and sensor feeds. Although current systems provide access to these data sources, analysts are still required to manually correlate information across modalities to identify operationally significant anomalies and inconsistencies. This process is time-consuming, cognitively demanding, and increasingly impractical in modern data rich operational environments where rapid decision-making depends on timely and actionable intelligence. This work presents a focused multi-modal anomaly detection framework that combines GEOINT imagery and textual intelligence to support AI-enabled situational awareness and operational analysis.

The proposed framework leverages recent advances in computer vision, large language models (LLMs), and multi-modal representation learning to identify anomalous battlefield activities and inconsistencies between visual and textual intelligence sources. Rather than attempting to solve every aspect of battlefield understanding, the effort focuses specifically on cross-modal anomaly detection and explainable operational reasoning. The architecture follows a Modular Open Systems Approach (MOSA) to support interoperability, extensibility, and future integration with evolving defense systems and operational data pipelines.

The first stage of the framework focuses on ingesting and preprocessing heterogeneous data sources, including satellite imagery, drone imagery, and textual intelligence reports. Preprocessing pipelines perform data normalization, metadata extraction, and segmentation of textual and visual content to prepare the information for downstream machine learning tasks. The framework is designed to support both historical analysis and near-real-time operational workflows.

The second stage extracts semantic representations independently from each modality. For imagery, a Vision Transformer-based encoder is used to capture spatial patterns, terrain changes, infrastructure modifications, and unusual object activity from GEOINT data. For textual intelligence, an LLM-based natural language processing pipeline extracts operationally relevant entities, events, and contextual relationships from intelligence reports and operational communications. These representations provide a compact semantic understanding of each modality while reducing dependence on handcrafted features or rule-based systems.

To enable cross-modal reasoning, visual and textual representations are mapped into a shared embedding space using contrastive learning techniques inspired by CLIP-style

architectures. This shared semantic space enables similarity analysis and consistency verification between imagery and textual intelligence. For example, the framework can identify situations where textual reports indicate routine operational conditions while imagery reveals unusual troop movement, defensive preparation, or unexpected infrastructure changes. Such inconsistencies may represent operationally significant anomalies that require analyst attention.

Anomaly detection is performed using embedding-distance analysis and unsupervised clustering methods to identify deviations from expected operational patterns. Finally, an LLM-based reasoning module generates analyst-friendly summaries and natural language explanations describing the detected anomalies and their potential operational relevance. By combining focused multi-modal fusion with explainable AI techniques, the proposed framework aims to reduce analyst workload, improve situational awareness, and provide scalable decision-support capabilities for future operational and simulation environments.