CHAPTER 2: STATE-OF-THE-ART

2.1 Computer Vision

Computer Vision (CV) has transitioned from classical, heuristic-based image processing to advanced deep learning architectures capable of understanding complex spatial relationships in real-time. In the rigorous context of aerospace flight testing, CV systems must overcome extreme environmental challenges, including high-velocity motion blur, variable atmospheric lighting, perspective changes, and partial occlusions. The primary requirement is high-precision localization of aircraft structural components to replace traditional, time-consuming manual workflows.

2.2 Real-time Object Detection

Modern object detection has shifted towards "one-stage" and "transformer-based" detectors to satisfy the strict latency requirements of Flight Test Instrumentation (FTI). Unlike older two-stage detectors, these architectures prioritize inference speed without significantly compromising mean Average Precision (mAP).

- **YOLO (You Only Look Once):** This CNN-based/Hybrid approach treats detection as a single regression problem, moving from image pixels directly to bounding box coordinates and class probabilities. Known for extreme efficiency, the latest iterations like **YOLO26x** achieve a latency of approximately 2 ms with a mAP (50:95) of 56.3%.
- **D-FINE (Transformer-based):** This model utilizes Vision Transformers to understand the global context of the whole image. It is particularly effective at precise edge localization and cutting exact object boundaries. The **D-FINE-X** version reaches a 56.5% mAP (50:95) with 61.7M parameters.
- **RF-DETR (Re-frame DETR):** Representing the state-of-the-art in 2026, this model specializes in small object detection through advanced transformer re-framing. It is critical for identifying distant aircraft or small components like antennas, achieving a peak accuracy of 59.0% mAP (50:95).

2.3 Universal Visual Segmentation

Segmentation provides a pixel-level mask, offering a deeper geometric understanding of the aircraft than simple bounding boxes. This precision is vital for identifying exact aircraft silhouettes against complex backgrounds.

- **DeepLab v3 (CNN-Based):** A solid baseline that uses Atrous Spatial Pyramid Pooling (ASPP) to capture multi-scale context and semantic information. While robust, its inference latency (50–300 ms) and reliance on traditional encoders often make it too slow for high-speed aircraft tracking.
- **YOLO11x-seg:** Optimized for speed, this model uses a specialized head for instance segmentation. It achieves a 57.5% precision with a remarkable latency of only 11.8 ms, making it the most viable candidate for real-time aircraft identification and geometry learning.

- **SAM 3 (Segment Anything Model):** A foundational transformer-based model (ViT) with zero-shot capabilities, meaning it can segment any object without specific training. However, its massive parameter count (310 M) and high latency (2500 ms) limit it to offline post-processing or high-end server environments rather than edge-based FTI hardware.

2.4 Keypoint Detection and Descriptors

To calculate the "Closure Rate" (Cv) during air-to-air refueling maneuvers, the system must track specific 2D points—such as the nose and antennas—with high stability and temporal consistency.

- **SIFT (Scale-invariant Feature Transform):** The classic gold standard for robustness, using local gradient histograms to ensure points are uniquely identifiable. However, its computational cost (~100-200+ ms) makes it unsuitable for real-time FTI applications.
- **FAST & ORB: FAST** provides rapid detection through pixel neighborhood comparisons, while **ORB** (Oriented FAST and Rotated BRIEF) combines this with a modified descriptor to maintain perspective and rotation invariance. ORB offers a balanced real-time performance (~17 ms) suitable for SLAM and mobile devices.
- **Superpoint (Fully CNN):** A state-of-the-art deep convolutional network that simultaneously detects keypoints and generates descriptors. By learning high-level features, it remains stable under complex structures, processing at ~13 ms on GPU and proving highly resistant to noise.

2.5 Actual Tracking System Trends and Investigation Lines

Current trends indicate a decisive move toward NMS-free (Non-Maximum Suppression) architectures and end-to-end transformers that eliminate manual post-processing bottlenecks. A key investigation line for this project is the integration of these disparate models into a unified, automated pipeline. This includes the generation of synthetic data from 3D engines (such as Unreal Engine) to train deep learning models where real flight data for specific aircraft parts is scarce or difficult to label.

2.6 Real-time Object Tracking Applications

The primary application of this research is the total automation of the flight test workflow. By replacing manual operator-dependent selection of critical points with an automated AI pipeline, localization capabilities can be scaled significantly. This ensures higher safety and precision standards during high-risk maneuvers, such as calculating the closure rate for a receiver aircraft approaching a tanker during air-to-air refueling.

CHAPTER 2: STATE OF THE ART

2.1. Introduction to Computer Vision in Critical Contexts

Computer Vision (CV) has undergone a revolutionary evolution, transitioning from classical image processing techniques based on heuristics and manually designed algorithms to advanced deep learning architectures capable of understanding complex spatial relationships in real-time. This transition has been driven by the availability of large datasets, increased computational capacity (especially with GPUs), and theoretical advances in neural networks.

In the rigorous and demanding context of aerospace flight testing, CV systems must overcome extreme environmental challenges that go beyond typical laboratory or controlled scenarios. These challenges include:

High-Velocity Motion Blur: Aircraft in test maneuvers can reach speeds where the relative motion between the camera (on another aircraft or on the ground) and the target generates significant blur, degrading image detail.

Variable Atmospheric Lighting: Lighting conditions change drastically during a flight (sunrise, sunset, clouds, night flight, reflections on cloud decks), affecting color and contrast consistency.

Perspective and Scale Changes: The relative position of the observed aircraft constantly varies, resulting in changes in perspective (front, side, top views) and scale (the aircraft can occupy from a few pixels to a large portion of the sensor).

Partial Occlusions: Elements such as wings, stabilizers, or the fuel itself during in-flight refueling can temporarily occlude critical parts of the aircraft that need to be monitored.

Complex and Dynamic Backgrounds: The sky, terrain, or sea as a backdrop present textures and patterns that can confuse algorithms, especially when contrast with the target is low.

The primary requirement within this domain is to achieve high-precision localization of the aircraft's structural components (such as the refueling probe/drogue, antennas, flaps, or nose). The ultimate goal is to replace traditional manual workflows, which are time-intensive and prone to human error, with an automated, objective, and scalable process. This chapter explores in detail the current technologies and methodologies that form the foundation for achieving this goal, focusing on real-time detection, segmentation, feature extraction, and tracking.

2.2. Real-Time Object Detection: The Speed vs. Accuracy Trade-off

The core of any visual automation system is the ability to detect and locate objects of interest within a frame. In the context of FTI (Flight Test Instrumentation), latency requirements are strict, often needing processing rates above 30 FPS (frames per second) to enable smooth tracking and immediate feedback. This has led to a shift away from classic two-stage detectors (like R-CNN and its variants), which prioritized accuracy at the expense of speed, towards "one-stage" architectures and, more recently, transformer-based models.

2.2.1. YOLO (You Only Look Once): The Efficiency Paradigm

YOLO represents a hybrid approach that treats detection as a single regression problem, directly mapping image pixels to bounding box coordinates and class probabilities. Its philosophy of "looking only once" at the entire image, rather than sequentially proposing regions of interest, is key to its speed.

Architecture: It combines a CNN backbone (like DarkNet or CSPNet) for feature extraction, a neck with techniques like PANet or FPN for multi-scale fusion, and detection heads that predict on a grid of cells.

Evolution and State of the Art: Since its inception in 2016, YOLO has evolved rapidly (v1-v11, and projected future versions like YOLO26x). The latest iterations incorporate attention mechanisms, optimized neck designs, and training with synthetic data. Models like the projected YOLO26x are reported to achieve a latency of approximately 2 ms with a mean Average Precision (mAP 50:95) of 56.3%, setting a benchmark for ultra-low latency applications.

Suitability for FTI: Its speed makes it an ideal candidate for the initial detection of the aircraft in the frame. However, its accuracy in delineating fine edges can be surpassed by other methods, and its performance on very small objects (at great distance) can be a weakness.

2.2.2. D-FINE: The Transformer Approach for Global Context

Models based on Vision Transformers (ViT) have demonstrated a superior ability to capture the global context of an image, allowing them to understand long-range relationships between pixels, which is beneficial for complex objects like aircraft.

Architecture: D-FINE uses a ViT encoder that divides the image into patches and processes them sequentially. A transformer decoder then generates the detection predictions.

Strengths: It is particularly effective in precise edge localization and cutting exact object boundaries, thanks to its holistic understanding of the scene. The D-FINE-X version, with 61.7M parameters, achieves a 56.5% mAP.

Considerations: Although faster than original DETR models, its computational cost is still higher than YOLO's, and its sensitivity to input resolution may be a factor to optimize for FTI.

2.2.3. RF-DETR (Re-frame DETR): The State of the Art in Small Object Detection

Representing the projected state of the art for 2026, RF-DETR addresses a critical limitation for FTI: robust small object detection.

Innovation: It introduces a "re-framing" mechanism within the transformer architecture, which allows for the adaptive re-evaluation and refinement of regions of interest, focusing on areas where small objects are likely.

Performance: Specialized in this task, it reaches a peak accuracy of 59.0% mAP (50:95), surpassing its contemporaries in scenarios where the target occupies few pixels, such as identifying distant antennas or sensors on an aircraft.

Relevance to the Project: This capability is critical for approach phases in air-to-air refueling, where the receiver may be initially far from the tanker, and its reference points are tiny in the image.

2.3. Universal Visual Segmentation: From Bounding Box to Pixel

While detection provides a bounding box, segmentation offers a pixel-level mask. This deeper geometric understanding is vital for identifying the exact silhouette of the aircraft against complex backgrounds (e.g., sky with diffuse clouds) and for isolating specific components (e.g., the refueling drogue).

2.3.1. DeepLab v3 (CNN-Based): A Robust Baseline

DeepLab v3 is a classic yet highly effective architecture for semantic segmentation.

Key Mechanism: It uses Atrous (Dilated) Convolutions and ASPP (Atrous Spatial Pyramid Pooling) to capture multi-scale context without drastically reducing spatial resolution. This allows it to recognize both large and small objects.

Limitations for FTI: Its typical inference latency (50–300 ms) and dependence on heavy encoders (backbones) like ResNet-101 make it too slow for real-time, high-speed aircraft tracking. Its use is limited to post-flight analysis or high-performance hardware.

2.3.2. YOLO11x-seg: Real-Time Instance Segmentation

Extending the YOLO philosophy, this variant adds an instance segmentation head optimized for speed.

Approach: It uses the same backbone and neck for feature extraction but adds a head that produces prototype masks and mask coefficients for each detection, fusing them to generate the final mask.

Performance: It achieves 57.5% precision with a remarkable latency of only 11.8 ms, making it the most viable candidate for real-time aircraft identification and geometry learning within FTI constraints.

2.3.3. SAM 3 (Segment Anything Model): The Foundational Model

SAM introduced the concept of a foundational model for segmentation, with impressive "zero-shot" capabilities.

Capabilities: It can segment any object indicated by a prompt (points, boxes, text), without having been specifically trained on that object category.

Architecture: Based on a massive ViT (310M parameters), a prompt encoder, and a lightweight mask decoder.

Applicability in FTI: Its enormous computational cost (~2500 ms latency) excludes it from execution on typical FTI edge hardware. However, it is an invaluable tool for the automated generation of labels in training data and for the offline post-processing of critical sequences where speed is not a priority.

2.4. Keypoint Detection and Descriptors: The Heart of "Closure Rate" (Cv) Calculation

To calculate kinematic metrics such as the closure rate (Cv) during air-to-air refueling maneuvers, the system must track specific 2D points on the aircraft (nose, tip of the drogue, wingtips) with high stability and temporal consistency. This requires not only detecting points but also describing them in a unique way to match them between frames.

2.4.1. SIFT (Scale-Invariant Feature Transform): The Robustness Gold Standard
SIFT is the classic algorithm, renowned for its invariance to scale, rotation, and partially to illumination.

Operation: It detects keypoints by identifying extrema in the difference of Gaussians at different scales. It then generates descriptors based on local gradient histograms.

Critical Limitation: Its computational cost (~100-200+ ms) makes it unsuitable for real-time FTI applications, relegating it to offline reference methods or calibration.

2.4.2. FAST & ORB: The Real-Time Compromise
FAST (Features from Accelerated Segment Test) is an extremely fast detector that evaluates a circle of pixels around a candidate point.

ORB (Oriented FAST and Rotated BRIEF): Combines the FAST detector with the BRIEF descriptor, adding orientation (for rotation invariance) and improving robustness. It offers balanced performance (~17 ms) suitable for SLAM and mobile devices.

Suitability: It is a viable option for FTI if using limited computing hardware, although its descriptors may be less distinctive than learning-based ones under large viewpoint or lighting changes.

2.4.3. SuperPoint (Fully CNN): The Learning-Based State of the Art
SuperPoint is a convolutional neural network trained in a self-supervised and supervised manner that revolutionized keypoint detection and description.

Architecture: It shares a single encoder for feature extraction and has two decoder heads: one for keypoint detection (as a probability map) and another for dense descriptor generation.

Advantages: By learning high-level features, it is exceptionally stable under perspective changes, illumination, and partial occlusion. It processes in ~13 ms on a GPU, is highly resistant to noise, and generates very powerful descriptors for matching.

Relevance for Cv: Its temporal stability and sub-pixel accuracy (if refinement is applied) make it the preferred choice for robustly tracking the critical points needed to reliably calculate the closure rate.

2.5. Current Trends and Research Lines
The field of real-time computer vision is constantly evolving. The most relevant current trends for this project are:

NMS-free (Non-Maximum Suppression) Architectures: NMS is a heuristic and costly post-processing step to remove redundant detections. New end-to-end models, especially transformers like DETR and its derivatives, eliminate the need for NMS, integrating duplicate suppression into the learning process, which simplifies the pipeline and reduces latency.

End-to-End Transformers: The unification of feature extraction, detection, segmentation, and even tracking into a single transformer architecture is an active research line. These architectures promise better global coherence and a simpler pipeline.

Integration of Disparate Models: A key challenge is how to optimally combine the best aspects of specialized models (e.g., YOLO for speed, RF-DETR for small objects, SuperPoint for keypoints) into a unified and automated pipeline. This involves research into feature fusion, computational attention management, and multimodal system design.

Synthetic Data for Training: The scarcity of labeled real flight data, especially for specific components of new aircraft or in critical maneuvers, is a bottleneck. A fundamental research line is the generation and use of photo-realistic synthetic data using 3D engines (Unreal Engine, Unity). This allows for the creation of infinite, perfectly labeled datasets to train deep learning models, followed by Domain Adaptation (DA) techniques to bridge the gap between the synthetic and real domains.

2.6. Applications in Real-Time Object Tracking for Flight Testing

The final application of all this technology is the total automation of the flight test workflow. The proposed system would integrate the analyzed components:

Initial Detection: A real-time detector (e.g., YOLO11x) locates the target aircraft in the video feed.

Segmentation and Refinement: A fast segmenter (e.g., YOLO11x-seg) or a prompt-based model (e.g., an optimized version of SAM) extracts the precise mask of the aircraft or a specific component.

Keypoint Extraction and Tracking: Over the region of interest, a detector like SuperPoint identifies and describes stable points. A tracking algorithm (Tracker) such as SORT, DeepSORT, or a tracking transformer maintains the identity and position of these points over time.

Metric Calculation: The stabilized 2D positions of key points (e.g., nose and drogue) are used, along with camera calibration data and possible auxiliary information (IMU), to calculate the closure rate (Cv) and other metrics of interest.

By replacing the manual, operator-dependent selection of critical points with this automated AI pipeline, the following is achieved:

Scalability: Multiple video streams and multiple points per aircraft can be processed simultaneously.

Objectivity: Variability between operators is eliminated.

Improved Safety and Precision: Consistent, real-time measurements are enabled during high-risk maneuvers like in-flight refueling, providing early alerts for out-of-limit parameters.

Cost and Time Reduction: The post-flight analysis cycle is drastically accelerated, and the human resources needed for repetitive tasks are reduced.

This chapter has established the technological foundation upon which the solution proposed in this Bachelor's Thesis will be built, justifying the selection of methodologies and pointing out the lines of innovation to be explored in the following chapters.