A-KIT VIO FOR AUV POSITIONING

A Project Report

Submitted to the APJ Abdul Kalam Technological University

in partial fulfillment of requirements for the award of degree

Bachelor of Technology

in

Computer Science And Engineering

by

Dona Sebastian(227)

Gauri P Nair(232)

Vinaya V(261)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

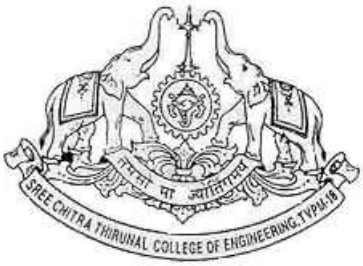
SREE CHITRA THIRUNAL COLLEGE OF ENGINEERING

KERALA

September 2025

DEPT. OF COMPUTER SCIENCE & ENGINEERING SREE CHITRA THIRUNAL COLLEGE OF ENGINEERING TRIVANDRUM

2022-26



CERTIFICATE

This is to certify that the report entitled **A-KIT VIO FOR AUV POSITIONING** submitted by **Dona Sebastian** (227), **Gauri P Nair** (232), **Vinaya V** (261) to the APJ Abdul Kalam Technological University in partial fulfillment of the B.Tech. degree in Computer Science And Engineering is a bonafide record of the project work carried out by him under our guidance and supervision. This report in any form has not been submitted to any other University or Institute for any purpose.

Prof. Kavitha K V
(Project Guide)
Associate Professor
Dept.of CSE
SCT College of Engineering
Trivandrum

Dr. Subu Surendran
(Project Coordinator)
Professor
Dept.of CSE
SCT College of Engineering
Trivandrum

Dr. Soniya B
Professor and Head
Dept.of CSE
SCT College of Engineering
Trivandrum

DECLARATION

We hereby declare that the project report **A-KIT VIO FOR AUV POSITIONING**, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala is a bonafide work done by us under supervision of Prof. Kavitha K V

This submission represents our ideas in our own words and where ideas or words of others have been included, we have adequately and accurately cited and referenced the original sources.

We also declare that I have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in my submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

Dona Sebastian

Gauri P Nair

Vinaya V

Trivandrum

20-10-2025

Abstract

Accurate localization in underwater environments remains one of the most significant challenges in autonomous navigation due to the absence of GPS and the presence of visual disturbances such as turbidity and poor lighting. This project, titled A-KIT VIO for AUV Positioning, aims to develop a robust positioning system for Autonomous Underwater Vehicles (AUVs) by integrating Visual-Inertial Odometry (VIO) with an Adaptive Kalman-Informed Transformer (A-KIT) framework. The proposed system fuses data from a camera and an Inertial Measurement Unit (IMU) through an Extended Kalman Filter (EKF), enhanced by a Set Transformer that dynamically learns and optimizes process noise parameters. This approach enables the model to adapt to varying underwater conditions and minimize localization errors. The project will be implemented and tested in a Gazebo simulation environment, which accurately models underwater physics such as drag, buoyancy, and current flow. The ultimate goal is to achieve a highly accurate and adaptive localization system that ensures reliable AUV navigation in GPS-denied environments, supporting applications in underwater exploration, surveillance, and marine research.

Acknowledgement

We take this opportunity to express my deepest sense of gratitude and sincere thanks to everyone who helped us to complete this work successfully. We express our sincere thanks to Dr. Soniya B, Head of Department, Computer Science And Engineering, Sree Chitra Thirunal college of Engineering for providing us with all the necessary facilities and support.

We would like to place on record my sincere gratitude to our project guide Prof. Kavitha K V, Associate Professor, Computer Science And Engineering, Sree Chitra Thirunal college of Engineering for the guidance and mentorship throughout this work.

Finally I thank my family, and friends who contributed to the succesful fulfilment of this seminar work.

Dona Sebastian

Gauri P Nair

Vinaya V

Contents

List of Figures

List of Abbreviations

Abbreviation	Full Form
A-KIT	Adaptive Kalman Informed Transformer
EKF	Extended Kalman Filter
IMU	Inertial Measurement Unit
VIO	Visual-Inertial Odometry
AUV	Autonomous Underwater Vehicle
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
ZUPT	Zero Velocity Update
SE(3)	Special Euclidean Group in 3D (for pose representation)
EM	Expectation–Maximization Algorithm
GT	Ground Truth
KF	Kalman Filter
AEKF	Adaptive Extended Kalman Filter
AKIT-VIO	Adaptive Kalman Informed Transformer for Visual-Inertial Odometry

List of Symbols

Symbol	Units / Type	Description
$\mathbf{u}_k$	m/s², rad/s	Raw IMU measurements at timestep k (linear acceleration and angular velocity)
I_k	image	Camera image captured at timestep k
$\mathbf{x}_k$	vector	True system state vector at timestep k
$\hat{\mathbf{x}}_{k\|k-1}$	vector	Predicted EKF state vector before measurement update
$\hat{\mathbf{x}}_{k\|k}$	vector	Updated EKF state vector after incorporating visual measurement
$\mathbf{P}_{k\|k-1}$	m², m²/s², rad²	Predicted EKF covariance before measurement update
$\mathbf{P}_{k\|k}$	m², m²/s², rad²	Updated EKF covariance after measurement update
$\mathbf{Q}_k$	m², m²/s², rad²	Process noise covariance matrix predicted by the Set Transformer
$\mathbf{R}_k$	m², rad²	Measurement noise covariance matrix predicted by the Set Transformer
$\mathbf{y}_k$	m, rad	Innovation or residual: difference between predicted and measured visual features
$\mathbf{S}_k$	m², rad²	Innovation covariance: represents uncertainty of the innovation
$\mathbf{K}_k$	matrix	Kalman gain used for updating EKF state and covariance
$f(\cdot)$	function	Nonlinear process model describing motion prediction using IMU data
$h(\cdot)$	function	Nonlinear measurement model mapping state to predicted observations
F_k	matrix	Process Jacobian $\partial f/\partial x$
H_k	matrix	Measurement Jacobian $\partial h/\partial x$
$\mathbf{z}_k^{pred}$	vector	Predicted visual measurements derived from current state estimate
$L_{innovation}$	scalar	Loss term penalizing large innovations
L_{cov}	scalar	Loss enforcing covariance consistency
L_{gt}	scalar	Loss representing deviation from ground-truth trajectory
L_{total}	scalar	Weighted sum of all loss terms used for training
$\lambda_1, \lambda_2, \lambda_3$	scalar	Hyperparameters weighting different loss terms
$\mathbf{x}_k^{gt}$	vector	Ground-truth state at timestep k
θ	parameters	Learnable parameters of the Set Transformer

Chapter 1

Introduction

1.1 Background

Accurate localization plays a critical role in the operation of Autonomous Underwater Vehicles (AUVs), which are increasingly used for ocean exploration, environmental monitoring, underwater infrastructure inspection, and defense applications. However, achieving precise positioning underwater remains a major challenge due to the absence of GPS signals and poor visibility conditions caused by turbidity, light scattering, and water depth. Traditional navigation systems, such as acoustic positioning or dead reckoning, suffer from drift and cumulative errors over time, making them unreliable for long-duration missions.

Recent advancements in Visual-Inertial Odometry (VIO) have shown promise in solving these challenges by fusing data from visual sensors (cameras) and inertial sensors (IMUs) to estimate the AUV's pose (position and orientation). However, underwater environments introduce nonlinearities and uncertainties that degrade the accuracy of conventional VIO systems. This project introduces an Adaptive Kalman-Informed Transformer (A-KIT) framework, which enhances the Extended Kalman Filter (EKF) with learning-based adaptability to dynamically adjust noise parameters and improve localization precision.

1.2 Problem Statement

Accurate positioning of AUVs (Autonomous Underwater Vehicles) in GPS-denied underwater environments using Adaptive Kalman-Informed Transformer (A-KIT) for visual-inertial odometry (VIO) .

1.3 Objectives

The objective of this project is to design a reliable and adaptive localization framework for Autonomous Underwater Vehicles (AUVs) operating in GPS-denied environments. The system focuses on improving pose estimation accuracy by combining advanced learning mechanisms with sensor fusion techniques. The proposed approach aims to integrate visual and inertial data in a unified model that dynamically adapts to complex and changing underwater conditions. The specific objectives of the project are outlined as follows:

1. To enhance AUV positioning by integrating a Set Transformer network with an Adaptive Kalman Filter, enabling dynamic adjustment of system noise parameters to improve pose estimation accuracy. This integration allows the model to learn and adapt to varying environmental conditions and sensor uncertainties, ensuring more stable and accurate localization.

2. To employ temporal attention mechanisms to learn sensor behavior over time, allowing the system to adapt in real time to complex marine dynamics such as current variations, turbidity, and motion disturbances. This improves the robustness and reliability of AUV navigation over long durations.

3. To support the development and evaluation of Visual-Inertial Odometry (VIO) models under challenging underwater conditions, including turbidity and lighting variations, using simulation software such as Gazebo. This enables testing and validation of the proposed system in a controlled environment before real-world deployment.

1.4 Scope of the Project

The scope of this project focuses on the simulation-based development and testing of a visual-inertial localization system using the A-KIT framework. The project involves the following activities:

- Modeling the underwater environment using Gazebo simulation tools, incorporating effects such as drag, buoyancy, and lighting variations.
- Implementing sensor models for camera, inertial measurement unit (IMU), and depth sensors to generate realistic input data for analysis.
- Developing and integrating Visual-Inertial Odometry (VIO), Set Transformer, and Extended Kalman Filter (EKF) algorithms for adaptive pose estimation and localization.

The following components are excluded from the current phase of the project:

- Real-world hardware deployment and testing on physical AUVs.
- Integration with external acoustic positioning or communication systems.

1.5 Report Overview

This report is organized into several chapters, each addressing a key stage of the project. Chapter 1 provides an introduction that outlines the background of the study, defines the problem statement, lists the project objectives, and explains the overall scope of the work. Chapter 2 presents a detailed literature survey that reviews previous research related to underwater localization, visual-inertial odometry (VIO) systems, and adaptive Kalman-informed transformer (AKIT) models. Chapter 3 describes the methodology followed in the project, including the system design, architectural framework, tools, simulation environment, and algorithms implemented. The subsequent sections include the project schedule, which outlines the timeline and major milestones of the development process, followed by the conclusion that summarizes the outcomes and future directions. Finally, the report concludes with a list of references that provide the academic and technical sources supporting the work.

Chapter 2

Literature Survey

2.1 Introduction

The literature survey reviews the existing research works related to underwater localization, adaptive filtering, and sensor fusion techniques. The goal is to understand current methodologies, evaluate their effectiveness, and identify the research gaps that motivate the proposed project, A-KIT VIO for AUV Positioning. Various methods such as Extended Kalman Filters (EKF), Adaptive EKF, and learning-based sensor fusion frameworks have been studied for their potential to enhance accuracy in GPS-denied environments.

2.2 RINS-W: Robust Inertial Navigation System on Wheels

This work by Brossard *et al.* [1] introduces a Robust Inertial Navigation System on Wheels (RINS-W) that enhances the performance of IMU-only navigation systems, particularly when GPS data is unavailable. The paper focuses on designing a robust system capable of accurately estimating the pose (position and orientation) of wheeled robots using only inertial measurements.

The proposed method integrates a Recurrent Neural Network (RNN) model with an Extended Kalman Filter (EKF). The RNN is trained to recognize specific motion states—such as stationary, straight motion, or turning—directly from raw IMU sensor

data (accelerometer and gyroscope readings). The network outputs zero-velocity updates (ZUPTs) and motion constraints, which are then incorporated into the EKF pipeline to refine pose estimation in real time.

Key Achievements:

- The first framework to integrate RNN-based motion classification and zero-velocity updates within an EKF pipeline.
- Demonstrated superior accuracy compared to previous inertial-only navigation methods on real driving datasets.
- Efficiently operates with low-cost IMUs and can run on embedded hardware in real-time conditions.

Limitations:

- The system relies solely on IMU data, which leads to drift accumulation over long durations or dynamic movements.
- Lacks external corrections such as GPS, vision, or depth sensors, making recovery from accumulated errors impossible in extended missions.
- The model is domain-specific—optimized for wheeled platforms—and does not generalize well to non-wheeled systems like AUVs or aerial drones.

Comparison and Identified Gap: RINS-W effectively combines learned motion constraints with probabilistic filtering, demonstrating that deep learning can enhance inertial navigation. However, it remains limited to terrestrial wheel-based motion and fixed motion classes. The proposed A-KIT VIO for AUV Positioning extends these ideas to the underwater domain by integrating visual information with inertial data and introducing a Set Transformer-based adaptive Kalman filter. This approach generalizes beyond predefined motion categories, dynamically adapting to complex marine dynamics and achieving robust localization under GPS-denied and visually degraded conditions.

2.3 VINet: Visual-Inertial Odometry as a Sequence-to-Sequence Learning Problem

Clark *et al.* [2] propose VINet, the first deep learning–based end-to-end Visual-Inertial Odometry (VIO) framework that estimates motion trajectories directly from raw visual and inertial data. Unlike traditional VIO pipelines that require manual synchronization, calibration, and hand-crafted feature engineering, VINet leverages a sequence-to-sequence learning approach to automate these steps and jointly learn motion estimation.

The VINet architecture combines a Convolutional Neural Network (CNN) for visual feature extraction with a multi-rate Long Short-Term Memory (LSTM) network that fuses inertial and visual features, handling different sampling frequencies between sensors. A differentiable SE(3) composition layer ensures that the predicted motion outputs remain physically consistent in three-dimensional space. This design enables VINet to perform sensor fusion in a unified, data-driven manner.

Key Achievements:

- First deep learning–based end-to-end framework for visual–inertial odometry.
- Reduced dependence on manual calibration and synchronization between camera and IMU sensors.
- Demonstrated superior accuracy to classical methods such as MSCKF and OKVIS on benchmark datasets like KITTI and EuRoC.

Limitations:

- Requires ground-truth pose data for supervised training, which is often impractical for underwater applications.
- The LSTM-based fusion lacks dynamic adaptability to varying sensor quality or environmental changes.
- Does not incorporate uncertainty modeling or probabilistic reasoning as in Kalman-based filters, leading to limited interpretability.

Comparison and Identified Gap: VINet successfully demonstrates that deep neural networks can replace hand-engineered visual-inertial fusion processes with learned representations. However, the lack of uncertainty estimation and online adaptability limits its deployment in complex and dynamic underwater environments. The proposed A-KIT framework combines the interpretability of Kalman filtering with the adaptability of attention-based transformers, allowing dynamic adjustment of noise parameters and improved reliability in GPS-denied and visually degraded conditions.

2.4 A New Adaptive Extended Kalman Filter for Cooperative Localization

This study by Huang *et al.* [3] presents an improved approach for localization using an Adaptive Extended Kalman Filter (AEKF) developed specifically for cooperative localization tasks, including applications for Autonomous Underwater Vehicles (AUVs). The primary objective of this research is to enhance the accuracy of position estimation in environments where GPS signals are unavailable or unreliable.

The proposed AEKF model focuses on the adaptive estimation of process and measurement noise covariance matrices, which are often unknown and vary with time. To achieve this, the authors utilize an online Expectation–Maximization (EM) algorithm that updates noise parameters continuously during system operation. This enables the filter to maintain improved convergence and accuracy without requiring pre-collected calibration data.

Key Achievements:

- The AEKF significantly improves localization accuracy compared to conventional EKF implementations.
- It estimates both process and measurement noise parameters online, offering dynamic adaptability.
- The approach eliminates the need for offline data, enabling real-time operation.

Limitations:

- The algorithm assumes Gaussian noise, which may not hold in real underwater environments.
- It guarantees only local convergence, making performance sensitive to initial conditions.
- The model relies on simplified dynamic assumptions and lacks the flexibility of deep learning or attention-based frameworks.

In summary, the work by Huang *et al.* introduces a reliable adaptive filtering mechanism that improves noise estimation and localization accuracy. However, it still depends on rigid probabilistic assumptions and lacks the capability to learn complex nonlinear sensor relationships. The proposed A-KIT framework in this project builds on these insights by integrating adaptive Kalman filtering with transformer-based attention mechanisms to achieve better robustness and adaptability in dynamic underwater conditions.

2.5 Attention Is All You Need

Vaswani et al. [4] propose the **Transformer**, a deep learning architecture that processes input in parallel using self-attention, replacing traditional sequential models like RNNs and LSTMs. The Transformer enables efficient learning of dependencies across sequences by allowing each element to attend to all others simultaneously.

Methodology: The Transformer framework relies on the following core concepts:

1. **Self-Attention Mechanism:** Enables each element of the input to dynamically weigh the importance of other elements, allowing the model to focus on the most relevant parts for prediction.
2. **Parallel Processing:** Unlike RNNs, the model can process all input tokens simultaneously, improving training efficiency.
3. **Positional Encoding:** Since self-attention is permutation-invariant, positional encodings are added to input embeddings to provide sequence order information.

Key Achievements:

- Enabled highly parallelizable training, reducing sequential computation bottlenecks.
- Captured long-range dependencies more effectively than RNN-based models.
- Achieved state-of-the-art performance on machine translation and various NLP tasks at the time of publication.

Limitations:

- Requires positional encoding to maintain sequence order information.
- Computational complexity is quadratic with respect to input sequence length: $O(n^2)$.
- Limited by a fixed-length context window, making very long sequences challenging to handle efficiently.

Comparison and Identified Gap: The Transformer revolutionized sequence modeling by eliminating recurrence and leveraging self-attention. However, its high computational cost and limited context length remain challenges, motivating research into efficient attention variants, sparse attention mechanisms, and memory-augmented transformers for longer sequences.

2.6 Set transformer: A framework for attention-based permutation-invariant neural networks

Lee et al. [5] propose the Set Transformer, a neural network architecture designed to process sets of data in a way that is permutation-invariant, models interactions among set elements, and is scalable to large input sets. This framework is particularly useful for tasks where the order of elements does not matter but their relationships are important.

Methodology: The Set Transformer framework consists of three main components that enable efficient and expressive set processing:

1. **ISAB (Induced Set Attention Block):** Summarizes the entire set into a small set of learned global features, allowing the model to capture the overall structure of the input efficiently.

2. **SAB (Self-Attention Block):** Models detailed interactions between all elements in the set, enabling each element to attend to others based on learned importance.

3. **PMA (Pooling by Multihead Attention):** Aggregates the set into a fixed-size vector by learning how to combine elements into a global representation suitable for downstream tasks.

Key Achievements:

- Developed a permutation-invariant architecture capable of modeling complex interactions among set elements.
- Introduced scalable attention mechanisms that allow efficient processing of large sets.
- Designed learnable pooling (PMA) using seed vectors to optimally aggregate set information for final tasks.

Limitations:

- Computationally expensive for very large sets.
- Requires a fixed number of seed vectors in PMA, limiting flexibility.
- Does not provide explicit modeling of uncertainty in predictions.

Comparison and Identified Gap: The Set Transformer effectively captures set-wise relationships while maintaining permutation invariance, addressing challenges in tasks where order should not affect the output. However, it is computationally intensive and lacks uncertainty estimation, which may limit its application in safety-critical or resource-constrained environments. Integrating uncertainty modeling and adaptive pooling strategies could further enhance its practicality for real-world problems.

2.7 Adaptive Kalman-Informed Transformer (A-KIT)

Cohen and Klein [6] propose the Adaptive Kalman-Informed Transformer (A-KIT), a hybrid model that integrates the principles of the {Extended Kalman Filter (EKF) with the learning capability of Set Transformers. The goal of A-KIT is to enable adaptive estimation of process and measurement noise covariance, thereby improving positioning accuracy under dynamic and uncertain conditions such as fast motion and sensor miscalibration.

Methodology: The A-KIT framework operates through a hybrid sequence of steps combining model-based prediction and transformer-based learning:

1. Predict the next state using the process model of the EKF.
2. Collect a new set of observed features from available sensors.
3. Use a Set Transformer to predict an adaptive measurement noise covariance matrix (R_t) dynamically based on the observed feature set.
4. Perform the projection step and compute the residual error.
5. Calculate the Kalman Gain to weigh prediction and measurement updates.
6. Correct both the state and covariance matrices using the adaptive parameters learned by the transformer.

Key Achievements:

- Introduced a novel fusion of transformer-based learning with probabilistic filtering for dynamic noise estimation.
- Achieved adaptive noise prediction without manual calibration or fixed parameters.
- Demonstrated a 35%–49% improvement in position accuracy compared to traditional and adaptive EKF methods on real AUV datasets.
- Maintained strong consistency by grounding learning outputs in physically interpretable Kalman update logic.

Limitations:

- The model is computationally intensive, limiting real-time applicability on low-power systems.
- Requires labeled pose data during training, making unsupervised adaptation challenging.
- May overlook inter-sensor correlations, as it focuses primarily on adaptive noise estimation per sensor set.

Comparison and Identified Gap: A-KIT successfully bridges deep learning and classical filtering by introducing transformer-driven adaptivity into the EKF pipeline. However, while it improves flexibility and accuracy, it remains computationally demanding and lacks domain-specific validation for complex underwater conditions. The present project, A-KIT VIO for AUV Positioning, extends this framework by integrating visual-inertial data fusion and validating the approach in simulated underwater environments using Gazebo. This ensures improved robustness to turbidity, lighting variation, and dynamic underwater forces while retaining the adaptive, transformer-based advantages of A-KIT.

2.8 Conclusion

The reviewed works highlight the evolution of transformer-based models for different data types and tasks. The Transformer [4] enabled parallel sequence processing and long-range dependency modeling but is limited by computational cost and fixed context length. The Set Transformer [5] extended attention to permutation-invariant set data, effectively modeling element interactions while remaining scalable, though uncertainty modeling remains unaddressed. The A-KIT [6] combines classical filtering with transformers to achieve adaptive noise estimation and improved positioning, but requires labeled data and is computationally intensive.

In summary, transformer-based architectures show strong potential for complex and dynamic data tasks, with ongoing challenges in efficiency, uncertainty modeling, and real-time deployment.

Chapter 3

System Design

3.1 Architecture of AKIT (Adaptive Kalman Informed Transformer) for Visual-Inertial Odometry

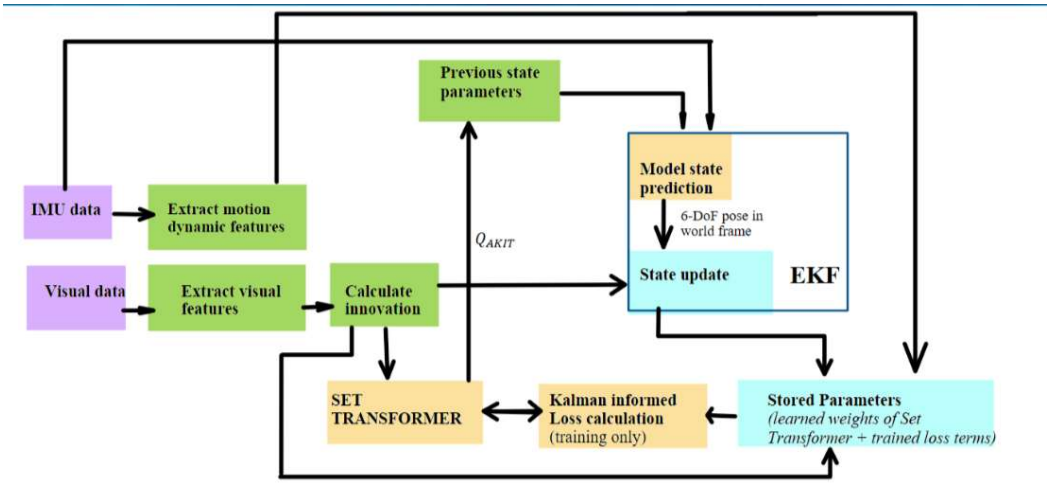


Figure 3.1: Architecture of the AKIT framework for Visual-Inertial Odometry

This figure 3.1 illustrates the complete pipeline of AKIT, highlighting how IMU and camera data are fused to estimate the 6D pose of a robot or device. The system first predicts motion using IMU measurements, computes innovation from visual landmarks, and updates the state via an EKF. The Set Transformer dynamically predicts sensor noise covariances and is trained using Kalman-informed losses, forming a feedback loop to improve estimation accuracy. Inputs and Feature Extraction in AKIT for VIO. This figure 3.1 shows the raw sensor inputs and the initial feature

extraction stage of the pipeline:

- **IMU Data:** Raw accelerometer and gyroscope measurements providing linear acceleration and angular velocity for motion prediction.
- **Visual Data:** Camera images capturing 2D features such as corners, textures, and CNN descriptors.
- **Feature Extraction:**
  - Detects keypoints and descriptors from images.
  - Tracks these features across frames.
  - Triangulates 2D correspondences into 3D landmarks for later use in EKF innovation.
- **EKF State Prediction (Motion Model):** Uses IMU measurements to forecast the next state and uncertainty:

$$\hat{\mathbf{x}}_{k|k-1} = f(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_k) + \mathbf{w}_k \quad \text{(predicted state vector)} \tag{3.1}$$

$$\mathbf{P}_{k|k-1} = F_k \mathbf{P}_{k-1|k-1} F_k^\top + \mathbf{Q}_k \quad \text{(predicted covariance)} \tag{3.2}$$

  Here, $\hat{\mathbf{x}}_{k|k-1}$ contains position, velocity, orientation, and sensor biases; F_k is the process Jacobian; $\mathbf{Q}_k$ is process noise covariance representing IMU uncertainty.

- **Innovation and EKF Update (Measurement Correction):** Incorporates visual measurements to correct the predicted state:

$$\mathbf{y}_k = \mathbf{z}_k^{pred} - h(\hat{\mathbf{x}}_{k|k-1}) \quad \text{(innovation/residual)} \tag{3.3}$$

$$\mathbf{S}_k = H_k \mathbf{P}_{k|k-1} H_k^\top + \mathbf{R}_k \quad \text{(innovation covariance)} \tag{3.4}$$

$$\mathbf{K}_k = \mathbf{P}_{k|k-1} H_k^\top \mathbf{S}_k^{-1} \quad \text{(Kalman gain)} \tag{3.5}$$

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \mathbf{y}_k \quad \text{(state update)} \tag{3.6}$$

$$\mathbf{P}_{k|k} = (I - \mathbf{K}_k H_k) \mathbf{P}_{k|k-1} (I - \mathbf{K}_k H_k)^\top + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^\top \quad \text{(covariance update)} \tag{3.7}$$

  H_k is the measurement Jacobian, $\mathbf{R}_k$ is measurement noise covariance predicted by the Set Transformer, and the retract operation is applied for quaternion

updates.

- **Kalman-Informed Losses:** Guide the Set Transformer to produce consistent covariances:

$$L_{innovation} = \frac{1}{N} \sum_{k=1}^{N} \mathbf{y}_k^\top \mathbf{S}_k^{-1} \mathbf{y}_k \tag{3.8}$$

$$L_{cov} = \frac{1}{N} \sum_{k=1}^{N} \|\mathbf{P}_k^{pred} - \mathbf{P}_{k|k}\|_F^2 \tag{3.9}$$

$$L_{gt} = \frac{1}{N} \sum_{k=1}^{N} \|\hat{\mathbf{x}}_{k|k} - \mathbf{x}_k^{gt}\|_2^2 \tag{3.10}$$

$$L_{total} = \lambda_1 L_{innovation} + \lambda_2 L_{cov} + \lambda_3 L_{gt} \tag{3.11}$$

  These losses ensure that the predicted covariances and state estimates are consistent with EKF corrections and, if available, ground truth trajectory.

- **Set Transformer (Adaptive Noise Prediction):** Dynamically predicts the process and measurement noise covariances based on sensor data and past innovations:

  - **Inputs:** IMU measurements, visual features, past innovation history.
  - **Outputs:** Adaptive $\mathbf{Q}_k$ (process noise) and $\mathbf{R}_k$ (measurement noise).
  - **Feedback Loop:** Kalman-informed losses (innovation, covariance, ground truth) are *pack propagated* to train the Set Transformer, improving future noise predictions.

3.2 Training Algorithm for AKIT

The AKIT training involves learning the adaptive noise prediction using the Set Transformer with EKF-in-the-loop. The algorithm uses IMU and visual data as inputs and backpropagates Kalman-informed losses to train the transformer.

Algorithm 1 AKIT Training Procedure

Require: IMU data $\{\mathbf{u}_k\}_{k=1}^N$, visual images $\{I_k\}_{k=1}^N$, optional ground truth $\{\mathbf{x}_k^{gt}\}_{k=1}^N$
Ensure: Trained Set Transformer predicting adaptive $\mathbf{Q}_k$ and $\mathbf{R}_k$

Initialize Set Transformer parameters θ and EKF state $\hat{\mathbf{x}}_{0|0}, \mathbf{P}_{0|0}$
for each sequence of length N **do**
  for $k = 1$ to N **do**
    Feature Extraction: extract keypoints/descriptors from I_k and triangulate 3D landmarks
    Set Transformer Forward: input IMU $\mathbf{u}_k$, visual features, past innovations; output $\mathbf{Q}_k, \mathbf{R}_k$
    EKF Prediction:

$$\hat{\mathbf{x}}_{k|k-1} = f(\hat{\mathbf{x}}_{k-1|k-1}, \mathbf{u}_k), \quad \mathbf{P}_{k|k-1} = F_k \mathbf{P}_{k-1|k-1} F_k^\top + \mathbf{Q}_k$$

    EKF Update:

$$\mathbf{y}_k = \mathbf{z}_k^{pred} - h(\hat{\mathbf{x}}_{k|k-1}),$$
$$\mathbf{S}_k = H_k \mathbf{P}_{k|k-1} H_k^\top + \mathbf{R}_k$$
$$\mathbf{K}_k = \mathbf{P}_{k|k-1} H_k^\top \mathbf{S}_k^{-1},$$
$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \mathbf{K}_k \mathbf{y}_k$$
$$\mathbf{P}_{k|k} = (I - \mathbf{K}_k H_k)\mathbf{P}_{k|k-1}(I - \mathbf{K}_k H_k)^\top + \mathbf{K}_k \mathbf{R}_k \mathbf{K}_k^\top$$

    Compute Kalman-Informed Losses:

$$L_{innovation} = \mathbf{y}_k^\top \mathbf{S}_k^{-1} \mathbf{y}_k,$$
$$L_{cov} = \|\mathbf{P}_k^{pred} - \mathbf{P}_{k|k}\|_F^2$$
$$L_{gt} = \|\hat{\mathbf{x}}_{k|k} - \mathbf{x}_k^{gt}\|_2^2$$
$$L_{total} = \lambda_1 L_{innovation} + \lambda_2 L_{cov} + \lambda_3 L_{gt}$$

    Backpropagate: update Set Transformer parameters θ using $\nabla_\theta L_{total}$
  end for
end for
return Trained Set Transformer

3.3 Tools and Technologies Used

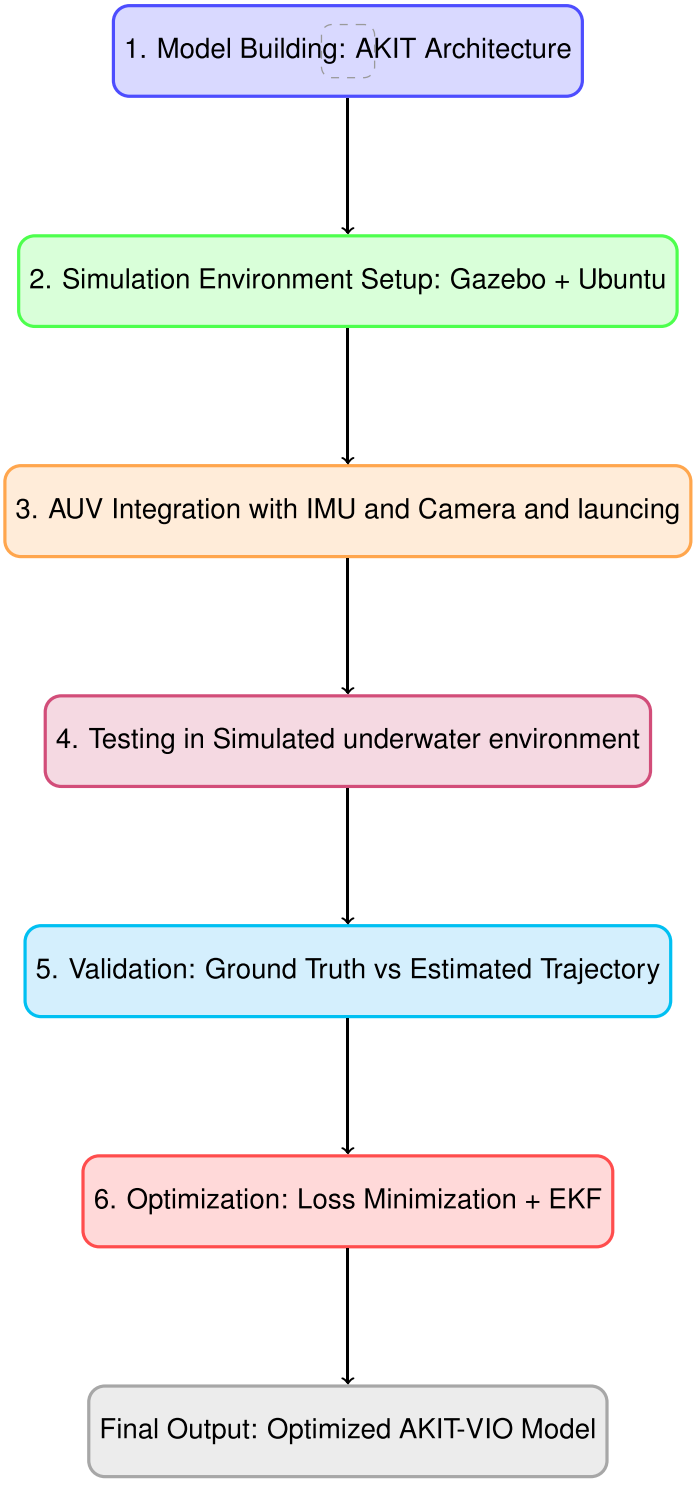
Table 3.1: Technical Stack Used for AKIT-based Visual-Inertial Odometry and simulation

Category	**Technology / Tool**	**Description and Role**
Operating System	Ubuntu 20.04	Stable and widely supported Linux distribution for robotics and ROS integration. Provides real-time kernel support, package management, and hardware compatibility for sensors and simulators.
Simulation Environment	Gazebo 11	Used to simulate the robot environment and sensor dynamics. Provides realistic camera and IMU models to generate synthetic datasets for VIO training and validation. Integrated with ROS topics for sensor data streaming.
Middleware / Communication Framework	ROS Noetic	Handles sensor data acquisition, message passing, and synchronization between IMU, camera, and control nodes. Topics: `/camera/image_raw`, `/imu/data`. TF tree used for spatial transformations.
Programming Language	Python 3.9 / C++17	Python used for ML model training and evaluation; C++ used for ROS nodes and real-time state estimation components.
Visual Processing Libraries	OpenCV, cv_bridge, NumPy	Used for visual feature extraction, tracking, image preprocessing, and conversion between ROS image messages and OpenCV matrices.
Inertial Data Processing	SciPy, NumPy, ROS `sensor_msgs`	Handles raw IMU data (linear acceleration, angular velocity), performs pre-integration, bias correction, and synchronization with visual timestamps.

Category	Technology / Tool	Description and Role
Machine Learning Framework	TensorFlow	Implements the Adaptive Kalman Informed Transformer (AKIT) model. Transformer layers process visual embeddings and IMU temporal features; fusion occurs at attention blocks.
Deep Learning Components	Transformer Encoder, Multi-Head Attention, Feedforward Layers, Positional Encoding	The Transformer backbone encodes temporal dynamics of IMU features and spatial cues from visual embeddings, adaptively fusing them using attention-weighted Kalman correction.
State Estimation Core	Extended Kalman Filter (EKF)	Provides probabilistic prediction–correction framework. IMU prediction used for process model update; Transformer-informed visual measurements refine pose and covariance.
Optimization and Loss Functions	Innovation Loss, Covariance Consistency Loss, Ground Truth Pose Loss	Ensures stability and consistency during training: • Innovation loss minimizes residual error between prediction and visual observation. • Covariance consistency loss constrains predicted uncertainty. • Ground truth loss aligns predicted pose with labeled data.
Visualization and Evaluation Tools	RViz, Matplotlib, TensorBoard	Used for visualizing trajectories, camera–IMU alignment, attention maps, and training metrics.

Category	Technology / Tool	Description and Role
Version Control and Build Tools	Git, CMake, catkin_make	Used for source management, building ROS packages, and maintaining modular node structure.

3.4 Project Workflow



Chapter 4

Project Schedule

Project Schedule				
Task	**Start**	**Finish**	**Status**	**Duration**
Project Idea Finalization	27 Jun 2025	14 Jul 2025	✅ Completed	2w
Project Approval	Nil	17 Jul 2025	✅ Completed	~1w
Papers Collection and Research	20 Jul 2025	20 Aug 2025	✅ Completed	~0.5
Preliminary Analysis	04 Sep 2025	08 Sep 2025	✅ Completed	~0.5
Feasibility Study	04 Sep 2025	08 Sep 2025	✅ Completed	2w
Literature Review	Nil	Sep 11, 2025	✅ Completed	~1w
Second Evaluation PPT Drafting	15 Sep 2025	25 Sep 2025	✅ Completed	1w
Architecture Design	15 Sep 2025	25 Sep 2025	✅ Completed	1w
Data Collection and Primary Environment Setup	26 Sep 2025	10 Oct 2025	✅ Completed	1w
Simulation Environment Creation (env1, env2, env3)	15 Oct 2025	03 Nov 2025	✏️ In progress	3w
AUV Sensor & Camera Integration	25 Oct 2025	07 Nov 2025	⏳ Not started	2w
Dataset Construction	10 Nov 2025	23 Nov 2025	⏳ Not started	2w
Data Preprocessing	24 Nov 2025	30 Nov 2025	⏳ Not started	1w
Model Building (EKF & Set Transformer)	01 Dec 2025	14 Dec 2025	⏳ Not started	2w
Model Training	15 Dec 2025	21 Dec 2025	⏳ Not started	1w
Model Testing and Validation	22 Dec 2025	28 Dec 2025	⏳ Not started	1w
Initial Model Optimisation	29 Dec 2025	03 Jan 2026	⏳ Not started	~0.5w
Launch AUV to Environment	04 Jan 2026	07 Jan 2026	⏳ Not started	~0.5w
Model Inference Testing (in env)	07 Jan 2026	10 Jan 2026	⏳ Not started	~0.5w
Final Model Optimisation	10 Jan 2026	15 Jan 2026	⏳ Not started	~0.5w

Figure 4.1: Project completion timeline

The project schedule outlines the timeline and progress of the A-KIT-based Visual–Inertial Odometry (VIO) system. As shown in Figure 4.1, the initial phases — including idea finalization, approval, literature review, feasibility study, and architecture design — are completed. The simulation environment setup is currently in progress, followed by upcoming stages such as sensor integration, dataset preparation, model training, and final testing for underwater localization validation.

Chapter 5

Conclusion

The Phase-1 work of this project focused on establishing a solid foundation for developing an adaptive and intelligent visual–inertial odometry (VIO) system for autonomous underwater vehicles (AUVs). Through an extensive literature review, various traditional and learning-based localization techniques—such as Adaptive Extended Kalman Filters, RINS-W, VINet, and the AKIT framework—were analyzed. This helped in identifying key research gaps related to noise adaptability, robustness to environmental disturbances, and multi-sensor data fusion under underwater constraints.

The proposed system, based on the Adaptive Kalman-Informed Transformer (AKIT), aims to integrate the reliability of Kalman filtering with the flexibility of transformer-based attention mechanisms. The architecture and training pipeline designed in this phase demonstrate how IMU and camera data can be fused efficiently using dynamic noise prediction and Kalman-informed feedback. Simulation-based development using tools such as Gazebo has been planned to evaluate system performance in controlled underwater environments before moving to hardware implementation.

Overall, this phase successfully defined the problem scope, objectives, methodology, and architectural framework for the project. The upcoming Phase-2 will focus on implementing the simulation model, fine-tuning the Set Transformer network, and performing quantitative evaluation of localization accuracy. This will move the project closer to achieving robust, adaptive, and real-time underwater positioning for AUVs.

References

[1] M. Brossard, A. Barrau, and S. Bonnabel, "Rins-w: Robust inertial navigation system on wheels," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019.

[2] R. Clark, S. Wang, H. Wen, A. Markham, and N. Trigoni, "Vinet: Visual-inertial odometry as a sequence-to-sequence learning problem," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1. AAAI, 2017.

[3] Y. Huang, Y. Zhang, J. Li, and W. Yu, "A new adaptive extended kalman filter for cooperative localization," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, no. 1, 2017.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, vol. 30, 2017.

[5] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *International conference on machine learning*. PMLR, 2019.

[6] N. Cohen and I. Klein, "Adaptive kalman-informed transformer," *Engineering Applications of Artificial Intelligence*, vol. 146, 2025.