SUP'COM
Higher School of Communication of Tunis

المدرسة العليا للمواصلات بتونس
ÉCOLE SUPÉRIEURE DES COMMUNICATIONS DE TUNIS

Graduation Project Report

2023/2024

Enhancing Visual Ego-Localisation through Cross-View Image Registration

Realized by:

LATRACH Aymen

Host Organizations:

Academic Supervisor:

Prof. Leila NAJJAR

Professional Supervisors:

Dr. Reza BAHMANYAR

Dr. Houda CHAABOUNI

Ecole Supérieure des Communications de Tunis
SUP'COM
2083 Cité Technologique des Communications - Elghazala - Ariana - Tunisie
Tél. +216 70 240 900 – site web : www.supcom.mincom.tn

Acknowledgements

First and foremost, I want to acknowledge the divine guidance of Allah. Without His blessings and the opportunities He has provided, I would not have reached this point. I am humbled and deeply grateful for the opportunities and growth this journey has brought into my life.

I would like to begin this report by expressing my sincere gratitude to the German Aerospace Center (DLR) and the Digital Research Center of Sfax (CRNS) for providing me with the invaluable opportunity to undertake this project through their collaborative partnership.

I would like to express special appreciation to my supervisors, Houda Chaabouni from CRNS and Reza Bahmanyar from DLR, whose guidance and support went far beyond the professional level. Their mentorship was crucial in shaping my experience, as they steered this work in the right direction with unwavering dedication. They generously invested their time, shared their expertise, and demonstrated immense patience, all of which were instrumental in ensuring the success of this project.

I would also like to extend my heartfelt thanks to my colleagues in the Photogrammetry and Image Analysis department. Our collaborative discussions were crucial in generating innovative ideas.

I am sincerely grateful to my academic supervisor, Leila Najjar, for her ongoing

mentorship and the invaluable insights that have played a crucial role in shaping my academic journey.

I would like to express my heartfelt gratitude to my family and friends for their unwavering support and encouragement throughout this academic journey. Your belief in me has been a constant source of motivation.

Table of Contents

List of Figures

List of Tables

Chapter 1

General Context and Problem Definition

1.1 Introduction

In this chapter, we begin by outlining the broader context of our project, covering the academic background and offering insights into the host organization. We then move into an in-depth exploration of the problem definition, highlighting the key challenges that have influenced our project's direction. Finally, we detail our specific contributions to the project, providing a thorough overview of our research efforts.

1.2 General Context

1.2.1 Academic Context

This work entitled "Enhancing Visual Ego-Localisation through Cross-View Image Registration" is carried out within the framework of the Graduation Internship presented in order to obtain the National Diploma in Telecommunications Engineer-

ing of the Higher School of Communication of Tunis for the academic year 2023/2024. The internship was hosted in collaboration between The German Aerospace Center (DLR)[1] and The Digital Research Center of Sfax (CRNS) [2].

1.2.2 Host Organisms

In the following sections, we will outline the organizational structure of DLR, starting with the institute level, moving through the department, and concluding with the specific team where we completed our internship. We will also introduce CRNS, the research center that played a key role in supporting and enabling the success of this project.

DLR[1]

The German Aerospace Center (DLR), or Deutsches Zentrum für Luft- und Raumfahrt, is a cornerstone of scientific and technological innovation in Germany. As a leading national research institution, DLR focuses on advancing aeronautics, space exploration, transportation, and energy. In aeronautics, DLR conducts pioneering research in areas such as aerodynamics, propulsion systems, and air traffic management, all aimed at enhancing the safety, efficiency, and sustainability of air travel. In space exploration, DLR plays a crucial role, contributing to satellite missions, Earth observation, and international collaborations through the European Space Agency (ESA). Additionally, DLR is dedicated to developing sustainable transportation and energy solutions, including high-speed trains, electric vehicles, and renewable energy sources.

DLR's expansive network of research facilities across Germany houses cutting-edge laboratories and simulation tools, enabling innovative research. By collaborating

with national and international partners, DLR promotes technology transfer and innovation, bridging the gap between theoretical research and real-world applications. The center's active involvement in various space missions highlights its commitment to expanding scientific knowledge and technological capabilities. Furthermore, DLR places a high priority on education and training, offering students, researchers, and professionals opportunities to contribute to and benefit from its extensive expertise. Through these efforts, DLR remains a driving force in shaping the future of aerospace, aviation, and energy technologies both in Germany and globally.

The Remote Sensing Technology Institute: IMF[3]

The Remote Sensing Technology Institute (IMF) is a premier research institution focused on advancing remote sensing technology and its diverse applications. At the forefront of scientific and technological progress, IMF specializes in creating operational image processors that extract geometric and semantic information from digital images captured by satellites, aircraft, and UAVs. Driven by the growing number of missions, including active involvement in the Copernicus program and collaborations with emerging satellite constellations, IMF has achieved significant operational and scientific milestones. The institute is dedicated to delivering standardized, high-quality data to an expanding user community, with a strong focus on pre-processing raw satellite data.

Notable projects include developing ground processing software for hyperspectral missions like DESIS and EnMAP, covering calibration, systematic and radiometric processing, orthorectification, and atmospheric correction. IMF's research spans a variety of fields, including photogrammetry, hyperspectral data analysis, atmospheric correction, and real-time applications such as traffic monitoring using air-

borne sensors. Recognized internationally for its atmospheric correction processors, IMF's innovations include cutting-edge software like PACO and MAJA, which have advanced the field significantly. With a commitment to innovation, IMF continues to push the boundaries of remote sensing, contributing both to operational advancements and scientific discovery.

Departement of Photogrammetry and Image Analysis[4]

The Photogrammetry and Image Analysis department at IMF specializes in developing operational image processors to extract both geometric (photogrammetry) and semantic (image analysis) information from remote sensing images captured by satellites, aircraft, and UAVs. Driven by the growing number of missions, such as the Copernicus program, and advancements in machine learning and computational power, the department has achieved significant operational and scientific accomplishments. A core objective is to provide standardized, high-quality data to a growing user community, necessitating robust pre-processing of raw satellite data.

Currently, the department is engaged in developing ground processing software for hyperspectral missions, including DESIS and EnMAP, addressing calibration, processing, orthorectification, and atmospheric correction. Research areas within the department include photogrammetry, hyperspectral data analysis, atmospheric correction, and real-time methods for applications such as traffic monitoring using airborne sensors. The team is particularly skilled in direct georeferencing, object detection, and building extraction through computer vision techniques. Their work in hyperspectral data analysis also spans land and water spectrometry, atmospheric correction, and algorithm development, and they are internationally recognized for their contributions to atmospheric correction processors.

Team of Traffic Monitoring[5]

The traffic monitoring team specializes in scientific research on automated image analysis for a range of traffic applications, including real-time responses during disasters and large-scale events. Their work encompasses the development of machine learning techniques and benchmark datasets for tasks such as vehicle detection, tracking, and area-based segmentation of traffic zones. Taking a comprehensive approach, the team designs full systems and process chains that integrate both hardware and software, enabling seamless image processing from data acquisition to result analysis.

Data is primarily collected through flight campaigns using custom-built sensor systems deployed on aircraft, helicopters, or UAVs, utilizing high-resolution 3K and 4K optical cameras for rapid, large-scale data acquisition and processing. The team's efforts also support authorities and organizations in security and disaster response tasks. Additionally, the outputs of these automated processes feed into wider transportation research, contributing to improved traffic models, more accurate estimates of land transport emissions, validation of mobile GNSS receivers, and enhanced ego-localization for road users based on aerial imagery. Through this multifaceted research, the team significantly advances both practical applications and theoretical insights in traffic monitoring and transportation science.

CRNS[2]

The Digital Research Center of Sfax, founded in July 2012, is a public institution dedicated to driving research and development in the rapidly evolving field of information and communication technologies (ICT). Its mission is multifaceted, encompassing leading-edge research, fostering innovation, and promoting the transfer of technology into practical applications.

With a focus on positioning itself at the forefront of ICT advancements, the center emphasizes applied research and innovation. This focus allows it to develop sophisticated and relevant solutions aligned with the needs of a fast-paced technological environment. The center's research is strategically directed toward ICT applications in online services, particularly within the e-tourism, e-health, and e-government sectors, addressing growing demands in areas vital to societal and economic advancement.

To support its goals, the Digital Research Center of Sfax is staffed by a skilled and diverse team, including senior researchers, many of whom are university professors, alongside Ph.D. students specializing in computing and telecommunications. This combination of experience and fresh perspectives fosters a vibrant, forward-thinking research environment.

Beyond theoretical research, the center focuses on developing practical techniques and tools in fields such as smart agriculture and smart environments. These efforts contribute not only to theoretical understanding but also to the creation of real-world applications, underscoring the center's commitment to impactful, solution-oriented research.

1.3 Problem Definition

Consumer-grade GPS sensors, commonly used in vehicles, are essential for various applications, including Advanced Driver Assistance Systems (ADAS), robot navigation, and geospatial data analysis. However, these sensors have inherent limitations that impact localization accuracy—an especially critical factor for autonomous vehicles. [6] [7]

- Satellite Geometry Arrangement and Accuracy: The accuracy of GPS data

is highly dependent on the geometric arrangement of visible satellites. To achieve reliable positioning, a sufficient number of satellites must be in view. When fewer satellites are visible, the system's ability to accurately determine location diminishes, leading to larger errors. A stronger geometric spread of satellites reduces positional uncertainty, while poor satellite coverage or signal obstruction can significantly degrade accuracy. The more satellites available, the better the precision of the positioning data.[8]

- Ionospheric and Tropospheric Effects: The Earth's atmosphere affects GPS signals, introducing delays due to ionospheric and tropospheric conditions. Variations in atmospheric conditions can alter the speed of GPS signals, leading to errors in distance calculations. Although selective availability (SA) has been mostly eliminated, atmospheric effects still contribute to inaccuracies. [9]
- Receiver Noise and Sensitivity: Consumer-grade GPS receivers often have limited sensitivity to weak signals, particularly in challenging environments. The presence of signal noise, along with the receiver's noise-filtering capabilities, impacts the accuracy of position estimations. [10]
- Multipath interference arises when GPS signals reflect off surfaces before reaching the receiver, creating multiple signal paths. Without proper mitigation, this effect can introduce significant errors in position estimates, especially in urban canyons or areas with reflective surfaces. [11]

Precise localization is critical in autonomous systems, especially in applications like autonomous driving, where highly accurate position data is required for safe and effective operation. Autonomous vehicles must make split-second decisions based on real-time positioning, often needing localization accuracy within a centimeter-level

range. However, consumer-grade GPS sensors typically provide accuracy within 3-10 meters, which is insufficient for the precision required by these systems.

With technological advancements, high-precision GNSS (Global Navigation Satellite System) receivers, which incorporate signals from multiple constellations (e.g., GPS, Galileo, GLONASS, BeiDou), as well as real-time kinematic (RTK) techniques, are becoming increasingly viable for meeting the localization requirements of autonomous driving, however their implementation is constrained by the requirement for signal reference stations, making them a costly solution.

1.4 Previous Attempts for Ego-Localization

In this section, we aim to present the previous attempts for ego-localization. We will begin by discussing multi-sensor navigation systems and their limitations, which highlight the growing need for more robust approaches, such as visual ego-localization. After addressing the challenges of traditional methods, we will then explore the state-of-the-art in visual ego-localization, specifically focusing on advancements relevant to our task.

1.4.1 Multi-Sensor Navigation Systems

Multi-sensor navigation systems have successfully mitigated the inaccuracies of GPS by integrating high-precision sensors. These advanced sensors work together to provide more accurate and reliable positioning, often achieving the centimeter-level precision required for autonomous systems. While IMUs (Inertial Measurement Units) provide accurate short-term trajectory estimates [12, 13], they are prone to drift over time, requiring regular calibration and synchronization. Adding radar

sensor to autonomous vehicle navigation systems can improve the precision and dependability of position and orientation estimates by enhancing GNSS and IMU fusion, while radar's long-range detection and weather-resilience qualities enhance overall robustness. Alternatively, various sensors including 360 cameras, depth cameras, and LiDAR—have been widely utilized for Ego-Localization and pose estimation [14, 15, 16]. Meanwhile, other self-localization techniques based on pre-constructed 3D High Definition (HD) maps [17, 18] face limitations due to the significant time and resources required for map creation and ongoing maintenance.

The integration of high-end sensors significantly increases the overall cost of the system, making it less accessible for widespread use. Additionally, these systems still face limitations, such as sensor fusion complexity, environmental challenges, and the need for frequent calibration and maintenance. These challenges highlight the importance of visual ego-localization systems, as visual data is more affordable, easier to integrate, and simpler to synchronize, offering a more accessible alternative.

1.4.2 State-of-the-art for Visual ego-localization

In crowded urban areas, cross-view localization using satellite imagery has shown substantial potential for correcting noisy GPS signals [19, 20] and enhancing navigation accuracy [21]. Additionally, visual data is relatively affordable and easy to integrate into various vehicles such as bicycles, motorcycles, and e-scooters, enabling data collection without requiring advanced systems.

Several frameworks have explored the effectiveness of visual ego-location and pose estimation using visual data from both ground and aerial perspectives. For instance, in [22], a framework called CCVPE introduces an end-to-end method for cross-view pose estimation using a ground-level image with a varying field of view and an aerial

image. This method estimates the 3 Degrees-of-Freedom camera pose of a query image by matching its image descriptor to descriptors of local regions in the aerial image. CCVPE employs two separate image encoders (without weight sharing), fuses the encoders' descriptors at a central layer, and then uses a decoder to produce the output. It achieves median localization errors of 1.42 meters and 3.47 meters on the VIGOR and KITTI datasets, respectively. CCVPE also supports ground images with different horizontal fields of view and incorporates an orientation prior to improve localization without retraining, though it faces challenges in generalizing to new datasets.

Similarly, [23] introduces a method called HC-Net for fine-grained cross-view ego-location, aligning a warped ground image with a GPS-tagged satellite image covering the same area through homography estimation in an end-to-end network. HC-Net processes spherical-transformed ground images and GPS-tagged satellite images as inputs, generating a homography matrix between them. This method achieves median localization errors of 1.59 meters and 4.57 meters on the VIGOR and KITTI datasets, respectively.

In [24], a model named PureACL presents a multi-camera fusion approach that significantly enhances localization accuracy. This approach leverages a spatial embedding that utilizes both intrinsic and extrinsic camera information, improving the extraction of spatially aware visual features and achieving median spatial accuracy errors below 0.5 meters in the lateral and longitudinal directions.

1.4.3 Open Perspectives for Visual Ego-location

Visual ego-location continues to open exciting avenues for research and practical applications, particularly in the context of dense urban environments. Cross-view

localization using satellite imagery has emerged as a promising approach to enhance navigation accuracy and correct noisy GPS signals, which are common challenges in these areas. With the advantage of affordable and widely available visual data, this technology has the potential to be integrated into a diverse range of vehicles.

One key area is improving generalization across datasets, as current methods struggle to adapt to unseen data. Developing more robust feature representations and leveraging geometric transformation could address this challenge. Additionally, scalable and lightweight solutions are essential for real-world deployment, particularly for integration into low-cost vehicles.The flexibility of data acquisition in such setups opens the possibility for large-scale data collection, which can subsequently be used to enhance the generalization of deep learning approaches. This potential for mass data gathering serves as a key motivation for our project.

1.5 Our Contributions

Our contributions focus first on the exploration of our localization dataset, which is specifically characterized to address the challenges of the task at hand. This includes thorough exploration, preprocessing, and labeling of the data. Next, we present our visual localization methodology, which, as we will demonstrate, enables the correction of noisy GPS signals using only visual data. Our approach employs a fine-to-grain localization process based on an image registration framework, leveraging deep learning techniques to address the visual localization problem effectively.

1.6 Conclusion

In conclusion, this chapter has established a solid foundation for understanding our project. We began by presenting the broader context, including the academic background and an overview of the hosting organization. This was followed by a detailed analysis of the problem definition, outlining the primary challenges that have shaped the direction of our work. As we progress through this academic journey, the upcoming sections will reveal our specific contributions, deepening our narrative and highlighting the impact of our research efforts .We will begin by addressing a primary aspect of any computer vision problem: the dataset. This includes presenting the benchmarks for the ego-localization task. Subsequently, we will introduce our dataset, detailing the process from acquisition to preparation. With this foundation established, we will transition to our methodology for the visual ego-localization task, exploring each component in depth and explaining its purpose step by step. Finally, we will present our results and engage in a discussion of their implications, including any limitations.

Chapter 2

Kokovi Dataset

2.1 Introduction

In the upcoming chapter, we explore the various datasets essential to our experimental work. We begin with an overview of publicly available datasets that are widely recognized and used in current research. Following this, we provide an in-depth examination of our own dataset, including detailed insights and explanations. This chapter serves as an essential introduction to the diverse data sources that form our experimental analyses and contribute to a deeper understanding of the research context.

2.2 Benchmark Datasets

A review of the literature reveals a variety of datasets developed to support tasks in visual ego-location and pose estimation. Each dataset, however, has distinct characteristics and inherent limitations, as they are tailored to specific tasks within the field. Below are some examples of these Benchmark datasets:

2.2.1 KITTI Dataset

The KITTI dataset [25] has been recorded from a moving platform while driving in and around Karlsruhe, Germany. It includes camera images, laser scans, GPS measurements and IMU accelerations from a combined GPS/IMU system. The main purpose of this dataset is to push forward the development of computer vision and robotic algorithms targeted to autonomous driving.

KITTI is one of the most prestigious dataset for self driving, provide a number of excellent benchmarks for the evaluation of stereo vision, optical flow, scene flow, visual odometry, SLAM, object detection and tracking, road/lane detection, semantic segmentation. The stereo camera faces the driving direction and has a horizontal FoV of $90\circ$, KITTI is calibrated, synchronized and rectified autonomous driving dataset capturing a wide range of interesting scenarios with raw data set is divided into the categories 'Road', 'City', 'Residential', 'Campus' and 'Person'.

The original KITTI dataset do not contain any aerial images, therefore most visual cross view ego-location frameworks use of the collected aerial images from [26] with ground resolution 0.20 m/pixel Each aerial patch covers a 100 m $\times$ 100 m ground area, As assumed in [26], ground images are located within a 40 m $\times$ 40 m area in the corresponding aerial patches.

For many years, KITTI has been a significant benchmark in the field. However, it has certain limitations, starting with the aerial images added later on from Google Maps, which may introduce changes in the scenes due to time gaps between captures. The satellite images in KITTI have a resolution of 512 $\times$ 512 pixels, covering an area of approximately 100 $\times$ 100 meters, resulting in a Ground Sample Distance (GSD) that may not capture finer road features crucial for cross-view ego-location tasks.

While these satellite patches provide a broad structural overview of the scene from above, a single ground-level query image often contains finer details that describe specific scene features. This is why frameworks employing four-camera views or 360-degree views can often work more effectively with larger GSDs; in contrast, KITTI's satellite GSD can pose challenges for fine-grained localization. Additionally, because KITTI uses GPS data as part of its ground truth positioning, there is an inherent positional error, particularly notable in scenes captured in dense urban areas.

2.2.2 VIGOR Dataset

The VIGOR dataset [27] includes geo-tagged ground-level panoramic images and satellite images collected across four major U.S. cities. Unlike earlier cross-view image retrieval datasets, VIGOR's satellite patches seamlessly cover target areas, with each patch corresponding to a 72.96m $\times$ 72.96m ground area and a Ground Sample Distance (GSD) of 0.114 m/pixel. The orientation of each satellite patch is aligned with the ground panorama such that the vertical centerline of the panorama points north in the satellite image.

Each patch typically overlaps with neighboring patches by approximately 50% to the north, south, east, and west, meaning that each ground image is covered by four satellite patches. A total of 90,618 aerial images were collected to cover central urban areas in New York City (Manhattan), San Francisco, Chicago, and Seattle, using the Google Maps Static API. Additionally, 238,696 street-view panorama images were captured through the Google Street View Static API, mostly at a zoom level of 2 along most streets. Each panorama image GPS location in the dataset is unique, with an average interval of about 30 meters between samples.

The raw image resolutions for aerial and ground views are 640 $\times$ 640 and 2048 $\times$

1024 pixels, respectively. Notably, VIGOR's data was collected exclusively in urban areas, where GPS signals are often noisier than in suburban settings. This dataset also provides raw GPS data for meter-level accuracy evaluation, aligning with the precision goals of localization applications.

2.3 Kokovi Data acquisition campaigns

2.3.1 Overview

Various benchmark datasets for autonomous driving have been published in recent years, each with a slightly different focus. Datasets designed for autonomous driving are generally useful for visual ego-location tasks, as most include color and gray-scale images—typically in mono, stereo, or panoramic formats—and also provide localization information from GPS/GNSS.

Despite their strong performance, deep convolutional neural networks (DCNNs) still face challenges when encountering occlusions. This common issue arises when distinctive features visible in the ground-level view, such as lane markings, unique road shapes, or sidewalk structures, are not discernible in satellite imagery. Satellite images are often sourced from Google Street View, where data quality is limited to more permanent features like buildings and general road topology, with a Ground Sample Distance (GSD) that may lack the necessary detail. Additionally, time gaps between ground and satellite captures can result in inconsistencies.

Simultaneous capture of ground and aerial data is necessary to enable reliable matching based on static, permanent features and to ensure consistency between different perspectives. The potential use of advanced sensors capable of achieving a finer GSD from an airborne view would also improve the accuracy of cross-view matching tasks.

To address this data requirement, the Traffic Monitoring team [5] in the Photogrammetry and Image Analysis department [4] at IMF [3], DLR [1], initiated simultaneous ground and aerial data collection, resulting as a part of the Kokovi Dataset. This campaigns was designed to support ego-location tasks through two data acquisition campaigns. The first campaign, "München-Pasing," was conducted in Pasing, approximately 7 km from Munich city center, Germany, on October 23, 2023. The second campaign took place about 1 km from Munich's city center on October 28, 2024.

Our work was conducted between these two campaigns, beginning with an analysis of the data available at the start of the internship and participating in annotation and preprocessing. We also participated in the second data acquisition campaign. However, by the time this work was finalized, we were unable to test and explore our methodology on the second part of the dataset due to time constraints. Nonetheless, this represents a crucial step for thoroughly evaluating and improving the generalization of our approach. In the following sections, we will delve into the dataset specifications to provide a comprehensive understanding of its structure and features.

2.3.2 Acquisition Campaign : München-Pasing

The campaign setup (see Figure 2.1) was as follows: a reference vehicle—an e-bike—was used to collect ground perspective data. This included capturing the front-view camera perspective and GPS data via a smartphone, which served as a consumer-grade GPS device for tracking the ground position over time. Simultaneously, a helicopter hovered above the reference vehicle to capture an aerial view, providing a comprehensive overview of the scene and allowing us to establish a ground-truth reference position for later analysis.

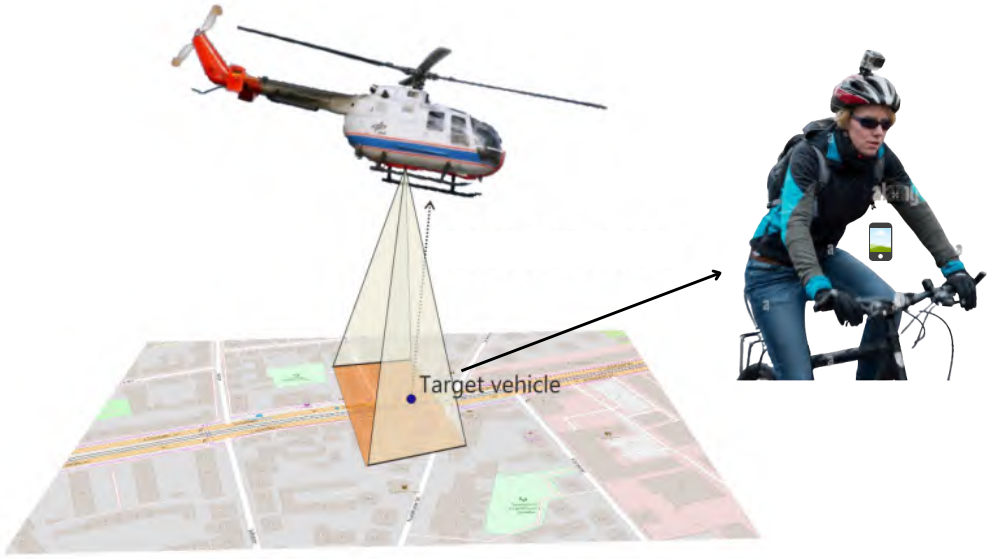

Figure 2.1: Campaign Setup

The region was chosen based on the characteristics of the urban areas available at the time, enabling us to test GPS accuracy and validate the positional error of consumer-grade devices in these environments. Below, we explore each sensor's specifications in detail:

Airborne sensors

When planning for the campaign, we first defined the desired trajectory, followed by the helicopter's hovering position relative to the camera's field of view. This ensured that the reference vehicle remained visible in all aerial images. Given the challenge for helicopter pilots to maintain a precise vertical hover above a moving vehicle, we set an altitude that provided a broad field of view, making it easier to continuously capture the reference vehicle from above.

The 4K sensor system mounted on the helicopter was responsible for capturing the aerial perspective as the reference vehicle followed the predefined route. The sensor's images cover the area surrounding the reference vehicle with a ground sample distance (GSD) of approximately 3.5 cm per pixel. With a 50 mm focal length, the camera covers an area of about 500 m $\times$ 350 m at an altitude of approximately 700 m above ground level. The image capture rate was set at 1 Hz throughout the campaign, and each image has a resolution of 14204 $\times$ 10652 pixels.

Vehicle sensors

The reference vehicle used in the campaign was an e-bike, with an OSMO Action camera mounted on the cyclist's helmet to capture ground-level footage. The camera was set to video recording mode at 30 fps with a resolution of 1080p, positioned approximately 2 meters above ground. This mounting position provides a wide range of motion, as the camera moves with the cyclist's head. However, this introduces challenges in relying on extrinsic camera parameters, which were not recorded and can sometimes be difficult to estimate.

Despite this, the setup mimics real-world scenarios, where ground-view data acquisition often involves similar characteristics. For instance, platforms like Mapillary [28] enable users to contribute street-level images and mapping data, which can vary widely in camera perspectives. Many research projects have leveraged such data, showcasing its utility even with variable extrinsic parameters.

Data Summary

In summary, Table 2.1 provides the specifications of both airborne and vehicle-based sensors, while Figure 2.2 presents different synchronized views captured at the same

time : (a) Aerial view captured from a pre-chosen hovering position of the helicopter; (b) Zoomed-in section of the aerial image highlighting the cyclist (circled in red); (c) Ground-level perspective from the helmet-mounted camera of the cyclist.

Table 2.1: Specifications of Airborne and Vehicle-Based Sensors

Sensor Type	**Platform**	**Specifications**	**Goals**
4K Sensor System	Helicopter	Resolution: 14204 x 10652 pixels Ground Sample Distance (GSD): 3.5 cm/pixel Focal Length: 50 mm Coverage Area: 500 m × 350 m at 700 m altitude Capture Rate: 1 Hz	Captures aerial view of reference vehicle and surroundings
OSMO Action Camera	E-bike (cyclist helmet)	Resolution: 1080p Frame Rate: 30 fps Mount Height: 2 m above ground	Captures ground view perspective
GPS (Smartphone)	E-bike	Consumer-grade GPS	Provides ground-level GPS data for reference vehicle

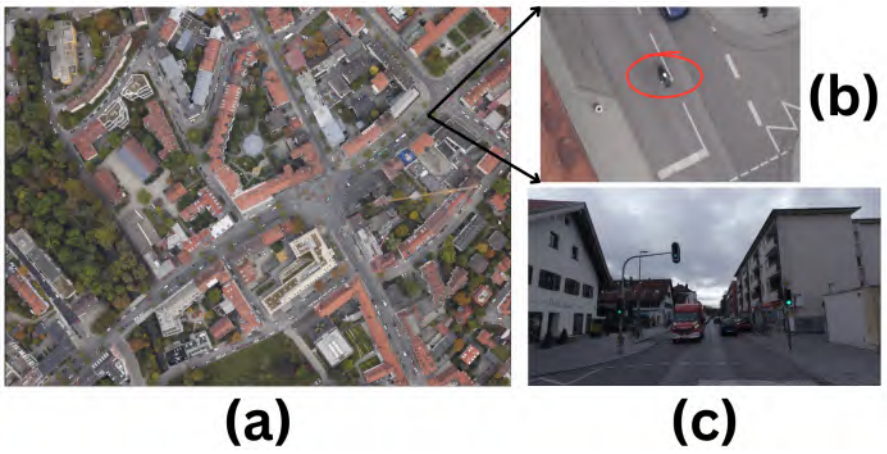

Figure 2.2: Example of Our Simultaneous Ground and Aerial Visual Data Collection

2.4 Annotation and Preprocess

Highly accurate reference vehicle trajectories are essential for our task, as they serve as ground truth data for validating our framework. In this study [29], they present a new method for generating reference vehicle trajectories in real-world conditions using simultaneously acquired aerial imagery from a helicopter. The reference trajectory is derived by performing a forward intersection of the vehicle's position in each image with the Digital Terrain Model (DTM). this method demonstrates an accuracy level of approximately 10 cm.

To determine the vehicle's position in each image, annotations were applied to the aerial images. After annotation, we were able to define the reference trajectory of the vehicle. The next step was to preprocess the data to synchronize the GPS position with the reference position and the ground view with the aerial imagery. It is worth noting that, during data acquisition, the helicopter could only hover in a fixed position for a few minutes at a time before it had to move, which sometimes caused the vehicle to move out of the aerial view. This required verifying the exact times when the reference vehicle and its surroundings were captured in the aerial view.

In the end, the preprocessing steps allowed us to create five sets, each containing a sequence of recorded positions over time. These sets represent the vehicle's trajectory in terms of both GPS recordings and reference positions, with corresponding ground and aerial views. Figure 2.3 illustrates these trajectories plotted on Aerial Imagery.

Figure 2.3: GPS Trajectory (Blue) and Reference Trajectory (Red) Overlaid on Aerial Imagery

2.5 Our Dataset in Numbers

As we can observe in Table 2.2, the frequency of the GPS device and the aerial image captures are not the same. This discrepancy arises because most consumer-grade GPS devices typically operate at a frequency of 1 Hz or close to it. On the other hand, the 4K system allows for a much higher frequency, resulting in a greater number of aerial images being captured. This increased frequency leads to more reference positions after annotation. To address this mismatch, we interpolated the GPS data sequence while respecting the timestamp of each aerial image capture. This interpolation ensures synchronization between the GPS data and the reference positions. Additionally, we can notice that in Sets 1 and 4, the number of reference positions is two points fewer than the number of aerial images. This is due to the reference vehicle being obscured by trees or other obstacles during the annotation process, preventing accurate positioning in those instances.

Table 2.2: Dataset Statistics

Set Number	**Number of GPS points**	**GPS Frequency (Hz)**	**Number of Reference Position Points**	**Reference Position Frequency (Hz)**	**Number of Aerial Images**
1	32	1	**52**	1.56	54
2	31	0.85	60	1.63	60
3	14	1	24	1.63	24
4	28	0.96	48	1.56	50
5	115	0.91	221	1.63	221
Total	**220**		**405**		**409**

As shown in Table 2.3, the average GPS error ranges between 2.8 and 6.8 meters, which aligns with the typical performance of consumer-grade GPS devices. Generally,

such devices exhibit an average error between 2 and 10 meters, depending on the characteristics of the area. Urban environments, in particular, often have more obstacles between the GPS device and satellites, leading to multipath effects that increase GPS errors. For our dataset, this was intentional, as the areas were chosen to reflect these challenging conditions.

Notably, Set 2 has the smallest average error of 2.8 meters. This can primarily be attributed to the fact that the sequence was captured in an open-sky environment with minimal surrounding buildings, thereby reducing potential obstructions. Additionally, the cyclist's velocity during this sequence was the lowest, which also contributed to the reduced error. Presenting these results highlights the challenges associated with relying on GPS measurements in such conditions and underscores the need for robust correction methods.

Table 2.3: Average GPS Error for Each Set

Set Number	**Average Error [m]**
1	4.53
2	2.8
3	4.94
4	6.8
5	4.7

In Figure 2.3, we provide an overview of the GPS trajectory alongside with the reference wchich surve as our ground truth. However, interpreting the errors can be challenging due to their dependence on time synchronization. This means that, at times, the blue trajectory may appear to align with or pass through the same position as the red reference trajectory in the aerial map, but there could still be a time lag that is not apparent when plotting all reference points alongside the GPS data.

To better illustrate the errors, we plotted the point-by-point positional error in meters for each set in Figure 2.4. One notable observation from this error plot is the occasional sharp increase in GPS error. For example, the error can range between 2 to 4 meters but then suddenly spike to 7 or 9 meters. This phenomenon is a direct result of multipath effects, which occur when transitioning from urban environments, such as roads flanked by tall buildings, to open-sky areas. This behavior is particularly evident in Set 1, where the sequence begins on a road surrounded by tall buildings, with the error averaging around 6.5 meters. However, as the trajectory transitions to an open-sky area around point 30, the error drops to approximately 2 meters, also in Set 3 the exact inverse way was presented and the transition went form an error around 4 meters before point 15 to an error around 7 meters and up to 9 meters.

The task of visual ego-localization remains an active research challenge, with current efforts striving to achieve an accuracy of 1 meter or even sub-meter precision. In comparison, GPS errors in urban areas, are significantly higher than this target. Therefore, our goal is to propose a correction method that consistently minimizes positional errors and enhances the performance of localisation through consumer-grade GPS devices. By exploring the GPS errors, we aim to gain a deeper understanding of our dataset, ultimately contributing to improved ego-localization accuracy.

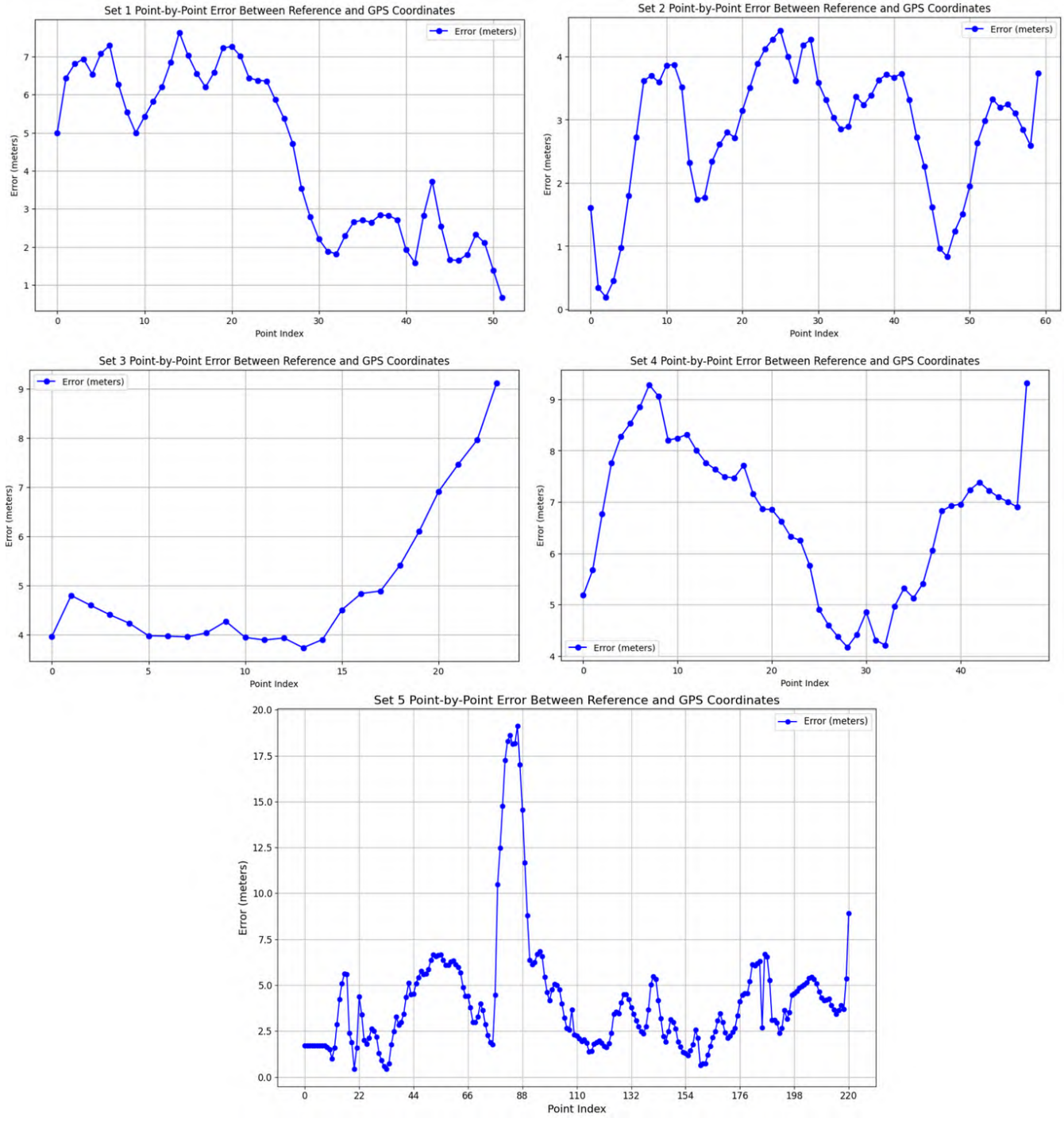

Figure 2.4: GPS Point-by-Point Error Between Reference and GPS Coordinates

2.6 Conclusion

In summary, this chapter has established the foundation for our experiments by presenting a range of datasets. We started with commonly used real-world datasets to provide a comparison baseline within the existing literature. We then introduced our proprietary dataset, emphasizing its distinctive features and the methods used in its creation.

Chapter 3

Defining our Visual Localization Methodology

3.1 Introduction

In the following chapter, we introduce our proposed approach to tackle the Visual Localization problem. We start with an overview of the pipeline to provide a high-level understanding of its structure. Then, we delve into each component in detail, offering in-depth insights into their specific functionalities.

3.2 Overall Pipeline

Before delving into the intricate details of our Ego-Localization pipeline, it is essential to outline the specifics of the inputs we provide and the desired outputs we expect from the system. The input consists of a ground-view camera frame from the reference vehicle and an initial rough localization obtained from a GPS sensor, which may have some inaccuracies, as discussed in the previous chapter. The ultimate goal is to achieve accurate vehicle localization.

An end-to-end approach, where a model directly learns to predict localization solely from the vehicle's camera feed, may not be the most suitable option. This method requires a large amount of training data, but our dataset includes fewer than 450 aerial views. Compared to state-of-the-art models trained on larger benchmark datasets, this is insufficient for training. Additionally, we lack labeled data for orientation or extrinsic camera parameters on a frame-by-frame basis.

Therefore, our strategy is to construct a multi-stage vision-based system. Our pipeline takes aerial and ground images as input and produces ego-localization, similar to an end-to-end approach. However, the key difference is that we employ a cascade of multiple steps rather than relying on a single model to directly predict localization.

The structural breakdown of our pipeline is as follows and can be observed in Figure 3.1.

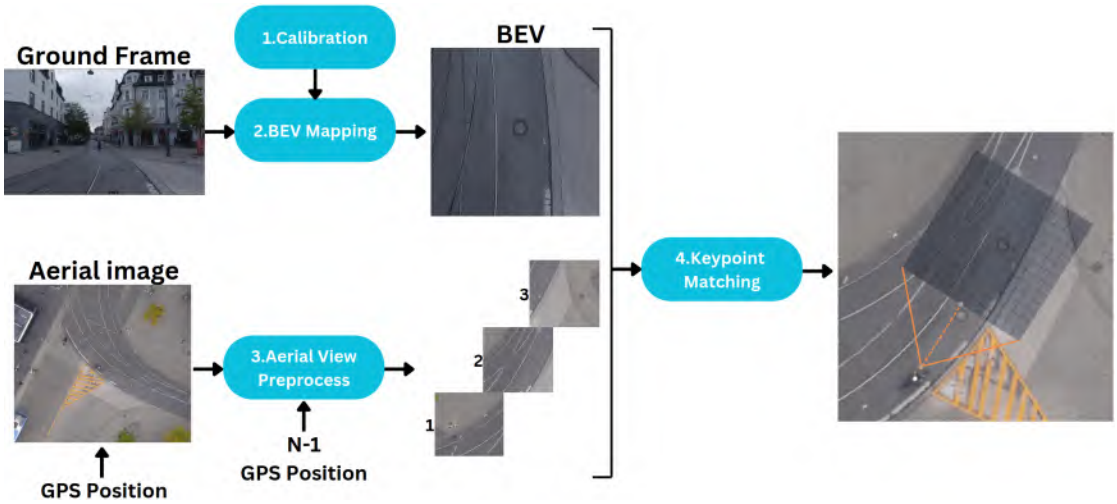

Figure 3.1: Overall Pipeline

3.3 Component's Details

3.3.1 Ground Frame Calibration

The objective of this initial step is to estimate certain intrinsic camera parameters, specifically the roll and pitch. The primary motivation for this step is to provide inputs for the subsequent stages, where we will explore their theoretical applications in greater depth. In this section, we aim to outline the calibration framework utilized and justify its selection.

Camera calibration involves estimating both intrinsic and extrinsic camera parameters, which are essential for most image-based 3D applications, such as 3D reconstruction, novel view synthesis, and perspective transformation. This problem has been extensively studied, with numerous tools based on 3D geometry available. These tools typically rely on sequences of images to track certain characteristics, such as flow variation and other geometric properties.

Recent advancements have addressed the challenge of single-image calibration using deep neural networks (DNNs) trained in a supervised manner. One such approach is GeoCalib [30], a DNN that incorporates universal rules of 3D geometry into its optimization process. GeoCalib [30]is trained end-to-end to estimate camera parameters and learns to extract meaningful visual cues from the data. Experiments on various benchmarks demonstrate that GeoCalib [30]is more robust and accurate compared to traditional and other learned methods.

In this study, we used a configuration for a pinhole camera with unknown parameters. We adopted GeoCalib[30] as an open-source framework to estimate the roll and pitch of a given image under this configuration.

3.3.2 Bird's-Eye-View Mapping

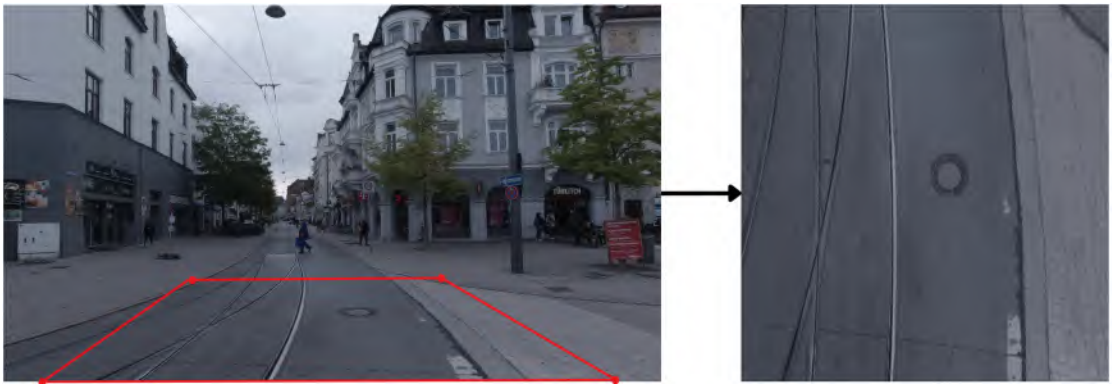

Figure 3.2: Bird's-Eye-View Mapping

A Bird's-Eye-View (BEV) perspective refers to a visual viewpoint that looks down on a scene from a high, elevated position, typically directly overhead or at a steep angle, resembling the view a bird would have while flying. In computer vision and mapping, a bird's-eye view provides a comprehensive, top-down representation of an area, it has been thought of as the preferred representation for many tasks including navigation or localization. Projectiong a given perspective to a Bird's-Eye View , this approach is referred to as an Inverse Perspective Mapping (IPM).

Recent methods leverage transformers to project feature maps from ground images onto a Bird's-Eye View . However, no existing approaches have explicitly used Inverse Perspective Mapping for BEV images in cross-view ego-localization. Such an approach would depend entirely on the coefficients of the homography matrix, which are derived from a thorough understanding of the camera's intrinsic and extrinsic parameters through a calibration process.

We observe that projecting ground images onto a bird's-eye view perspective can make aerial-based localization tasks more intuitive, much like how humans use maps

for navigation. To reduce the domain gap caused by the perspective difference between the aerial and ground views, applying a perspective transformation to one of the images can simplify the problem. In this context, we aimed to project the region of interest from the front view image onto the bird's-eye view. This step transforms the complex cross-view localization problem into a 2D image alignment task.

In Figure 3.2, the red area in the front-view image on the left represents the region of interest, while the transformed image on the right shows this area from a bird's-eye perspective.

The general assumption for the transformation to bird's-eye view images using inverse-perspective mapping is that the area of interest is assumed to be flatt, and the height of the road environment is effectively zero. This assumption is reasonable when it comes to performing IPM on an image of the road around the vehicle. In the proposed geometric IPM approach, we also assume that at least one of the camera's field of view (FOV) angles is known, as the other can be inferred from this angle and the image resolution. Additionally, we assume knowledge of the camera's pitch angle relative to the scene, whiche is the pitch, We also assume that the camerahas no roll against the scene at the given frame as this can be filtred using the previous step of our pipline. The proposed geometric IPM approach for BEV image projection is suited for usecases with limited knowledge of the camera's parameters, especialy where an estimation can derive the needed parameters without the need to the explicite caribration befor and during data acquisation, this show the importance of the step one of the pipline,

In this section, we describe in detail the geometric reasoning of IPM used and the reduced projection in terms of known camera's parameters. In Figure 3.3, f_{ov} represents the vertical field of view of the front view image, f represents the focal length

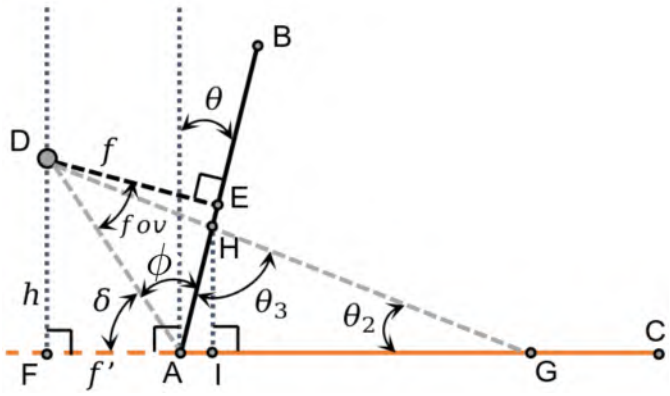

Figure 3.3: Bird's-Eye-View Geometric Reasoning

DE, h represents the distance from the camera optical center D to the bird's-eye projection plane, and θ represents the pitch angle that we estimated in the previous step otf the pipline. To facilitate the subsequent derivation, we also define δ to represent $\angle FAD$, ϕ to represent $\angle BAD$, θ_2 to represent $\angle AGD$, θ_3 to represent $\angle AHG$, f' to represent FA, and l_0 to represent AD. These variables are calculated as follows:

$$
\begin{aligned}
f &= \frac{H_f}{2\tan(f_{\text{ov}})} \\
\phi &= \frac{\pi}{2} - f_{\text{ov}} \\
\delta &= \frac{\pi}{2} - (\phi - \theta) \\
l_0 &= \sqrt{f^2 + \left(\frac{H_f}{2}\right)^2} \\
h &= l_0 \sin\delta
\end{aligned}
$$

In figure 3.4 we ploted on the BEV plane gride the camera projection refered as F calculated as follow:

$$f' = l_0 \cos\delta. \tag{3.1}$$

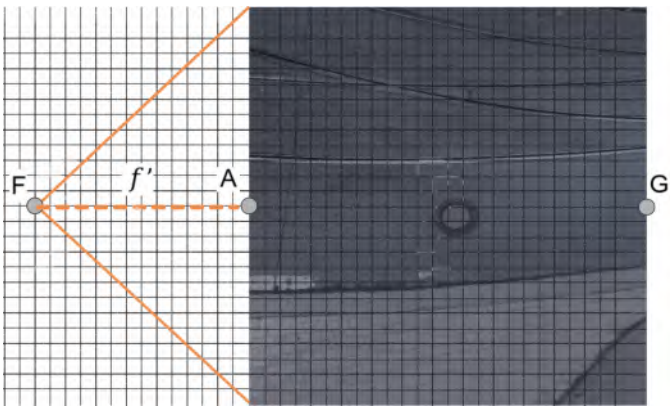

Figure 3.4: Camera Projection

We denote the corner coordinates on the bird's-eye view imaging plane as (u_b, v_b). The values of θ_2 and θ_3 can be calculated as follows:

$$\theta_2 = \arctan\left(\frac{h}{f' + H_b - v_b}\right)$$

$$\theta_3 = \frac{\pi}{2} + \theta - \theta_2.$$

Then, based on the sine and similarity triangle theorems, we can calculate the corresponding pixel coordinates (u_f, v_f) for each corner of the red shape on the front view image in figure 3.2. The calculation is as follows:

$$\frac{H_f - v_f}{H_b - v_b} = \frac{\sin\theta_2}{\sin\theta_3}$$

$$\frac{W_f/2 - u_f}{W_b/2 - u_b} = \frac{f' + (H_f - v_f)\sin\theta}{f' + H_f/2 - v_f}.$$

Solving for v_f and u_f, we get:

$$v_f = H_f - \frac{\sin\theta_2}{\sin\theta_3}(H_b - v_b)$$

$$u_f = \frac{W_f}{2} - \frac{f' + (H_f - v_f)\sin\theta}{f' + H_f/2 - v_f}\left(\frac{W_b}{2} - u_b\right).$$

Calculating the corners of our area of interest in the front view image allows us to directly apply the homography matrix, obtained from the projection process, to obtain the Bird's Eye View image.

To conclude, the proposed IPM approach can be utilized in scenarios where the camera's intrinsic parameters are not fully known or their acquisition is restricted. However, it is important to note that the accuracy of this transformation relies on the estimation of the extrinsic rotations.

3.3.3 Aerial View Preprocess

The primary goal of this approach was to make keypoint matching more efficient and effective by cropping the aerial image into three square windows. This maximizes the likelihood that the area of interest to be matched with the Bird's-Eye-View is contained within one of the cropped windows. This strategy proved to be more effective than directly feeding the model with the entire aerial view centered around the current GPS position.

As shown in Figure 3.5 the first step, we compute an arrow indicating the direction from the previous GPS point to the current one, represented by the blue arrow in Figure 3.5. This arrow simulates the direction in which the camera should be

oriented. We then rotate the aerial view by one of four possible angles: 0°, 90°, 180°, or 270°. We select the angle that maximizes the alignment of the arrow with the top of the image, ensuring that the camera's field of view is oriented optimally. This rotation helps reduce the angular difference between the two views, making the matching process easier.

Next, we crop three windows from the aerial image, focusing on the area above the GPS point. By validating this preprocessing step on the validation set (Set 3), we observed a significant improvement over matching directly with the centered aerial patch.

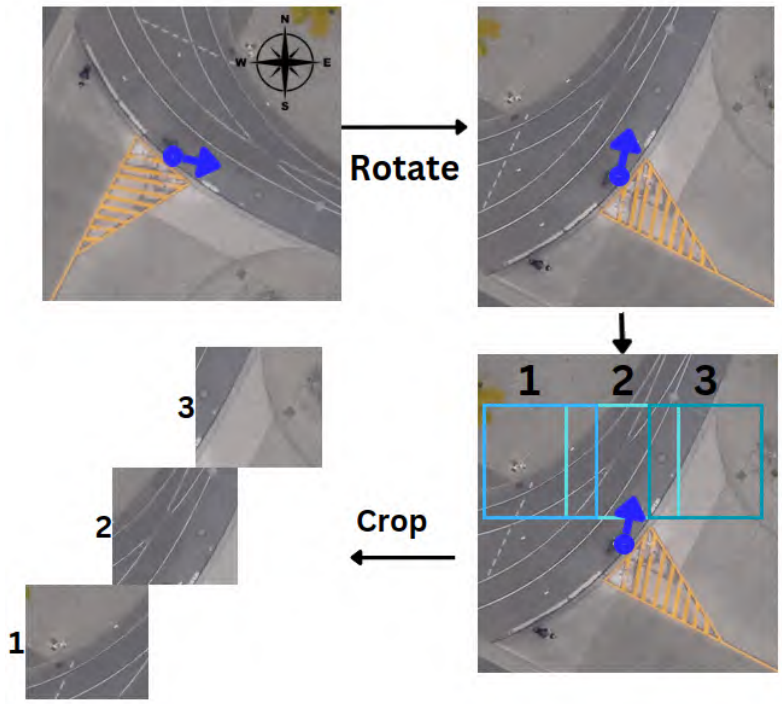

Figure 3.5: Aerial View Preprocess

3.3.4 Aerial to Bird's-Eye-View Keypoint Matching

After determining the camera's location on the bird's-eye view grid, our goal is to estimate its position within the aerial georeferenced patch to retrieve accurate UTM coordinates. This next step, Aerial to Bird's-Eye-View Matching, allows us to perform image registration, aligning the BEV over its true position on the aerial image. One approach to achieve this is through a keypoint matching framework.

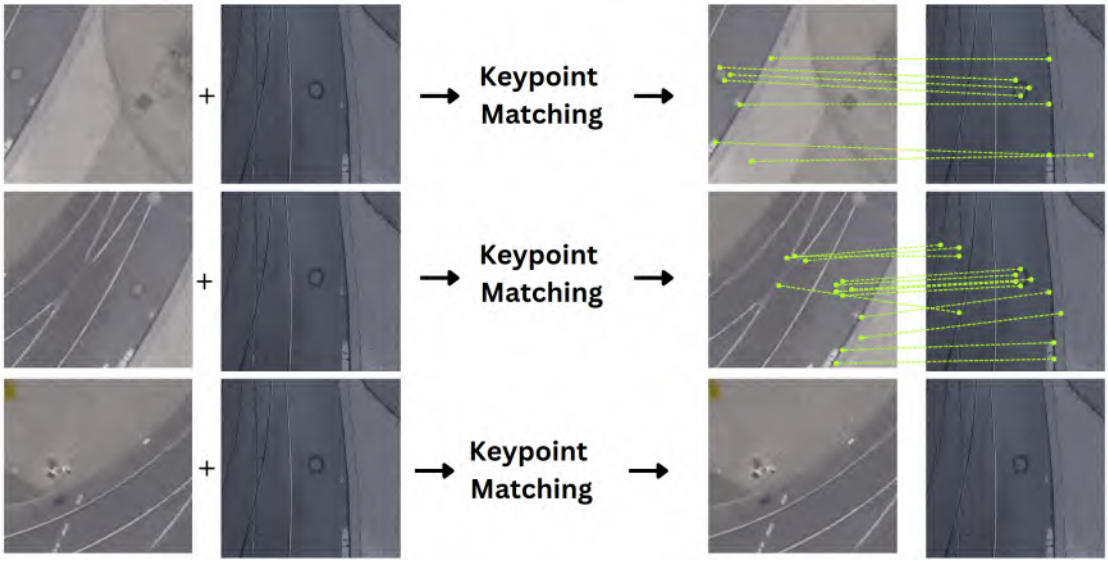

Figure 3.6: Aerial to BEV Keypoint Matching

In this process, the Bird's-Eye-View image and the aerial patch (selected based on the current GPS coordinates) serve as inputs. The framework identifies matching keypoints between the two images, providing two lists of corresponding keypoints that represent the shared features across the BEV and aerial perspectives. This registration step enables us to accurately position and align the BEV image on the aerial map, ultimately refining our localization precision.

Keypoint feature matching is a core technique in computer vision for identifying and aligning corresponding points across images. There are two main approaches: geometric methods and deep learning-based methods.

Geometric approaches use handcrafted descriptors, such as SIFT (Scale-Invariant Feature Transform) and ORB (Oriented FAST and Rotated BRIEF), to detect and describe keypoints based on geometric characteristics like edges, corners, and textures. These methods rely on predefined algorithms to capture visual features, offering consistency across simpler conditions.

Deep learning-based approaches, on the other hand, employ neural networks to learn features directly from large datasets, providing greater robustness to changes in scale, orientation, lighting, and perspective. Frameworks like SuperGlue and D2-Net are specifically designed to detect and match features by leveraging extensive training data, which allows them to generalize more effectively in complex or variable environments.

Each approach offers distinct advantages; however, a recent study[31] where they presented a comparative analysis of advanced feature matching algorithms in challenging high spatial resolution optical satellite scenarios found that the combination of SuperPoint [32] + LightGlue[33] stands out. This pairing demonstrated a strong balance of robustness and accuracy in complex optical satellite environments. Given the similarity of our scenario, we used these findings as a basis for testing various algorithms and ultimately selected the SuperPoint[32] + LightGlue[33] combination as our deep learning framework for keypoint matching.

SuperPoint[32] (SP) It is a key point extraction algorithm, which realizes self-supervised point of interest detection and descriptor generation through unsupervised training.

LightGlue [33] (LG) is a recent deep learning-based framework designed for efficient and accurate keypoint matching between images, it employs an attention-based mechanism to compare and match keypoints across images. The network uses both self-attention and cross-attention layers.

In our specific case, SuperPoint's [32] extracted keypoints and descriptors aim to capture shared and consistent features across both images (ground and aerial perspectives). This focus on image region rich in salient features (such as prominent landmarks or structural components), increases the likelihood of finding reliable matches even when the two perspectives vary in appearance. Once SuperPoint [32] has extracted the keypoints and descriptors, LightGlue[33] performs the matching process. Figure 3.6 shows examples of successfully matched BEV and the three preprocessed aerial windows.

In summary, SuperPoint [32] and LightGlue[33] together offer an efficient and robust approach for keypoint matching in our cross-view localization tasks. The final output from this component as mentioned before is geting the tow list of keypoint that match in both images.

3.3.5 Bird's-Eye-View to Aerial Registration

The final step before predicting the camera position on the georeferenced aerial image involves estimating a transformation matrix **H**, which calculates the necessary transformations to align the Bird's Eye View with the aerial view based on the list of keypoint that match in both images calculated in the step befor.

A homography is a 3×3 transformation matrix **H** that relates two planes in projective

space. It transforms points between two images or planes such that:

$$s\mathbf{x}' = \mathbf{H}\mathbf{x}$$

where:

- $\mathbf{x} = [x, y, 1]^T$ is a point in the first image/plane (in homogeneous coordinates),
- $\mathbf{x}' = [x', y', 1]^T$ is the corresponding point in the second image/plane,
- s is a scale factor (to normalize the homogeneous coordinate),
- $\mathbf{H}$ is the 3×3 homography matrix.

To estimate the homography matrix $\mathbf{H}$, we need at least four pairs of corresponding points $(\mathbf{x}, \mathbf{x}')$ because the matrix has 8 degrees of freedom. The estimation process involves formulating a linear system of equations for each pair of points based on the homography transformation. These equations are stacked into a matrix $\mathbf{A}$, and $\mathbf{H}$ is computed using Singular Value Decomposition (SVD) by solving $\mathbf{A}\mathbf{h} = 0$. To ensure numerical stability, points are normalized by translating and scaling them. In practice, noise and outliers in the correspondences can distort the estimation. To address this, we apply the RANSAC algorithm, which identifies the largest set of inliers and robustly estimates $\mathbf{H}$, ensuring reliable performance even in noisy conditions.

This transformation encompasses scaling, rotation, and translation. Since the warping from the ground view to the Bird's-Eye-View has already been performed in the BEV Mapping section, we proceed in the same manner to complete the transformation in this step. We use the estimateHomography function from OpenCV, followed by the warpPerspective function, to generate the overlay shown in Figure 3.7. The

camera location, computed in Equation 3.1, is then transformed from the BEV grid to the aerial patch using the same transformation matrix. This results in the final estimation in our cross-view pipeline, providing a more accurate visual representation of the camera's position.

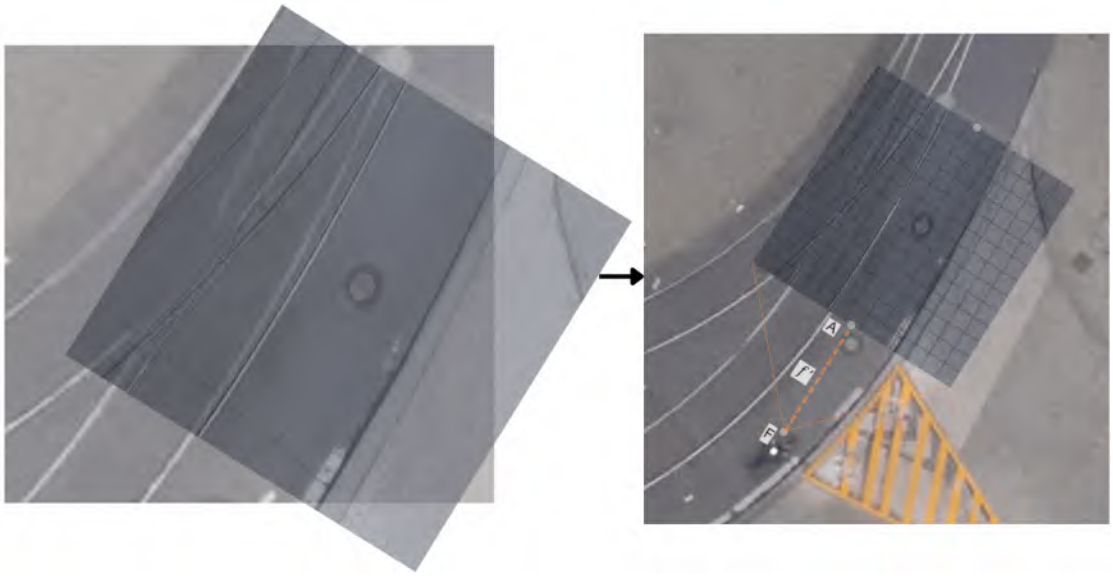

Figure 3.7: Bird's-Eye-View to Aerial Registration

3.4 Conclusion

In conclusion, this chapter has thoroughly examined the intricacies of our proposed methodology for addressing the Ego-Localization problem. We began with a comprehensive overview of the pipeline, providing readers with a foundational understanding of its structure. Subsequently, we delved into the finer details, offering a detailed explanation of each component and highlighting their roles and interactions within the framework. The insights presented in this chapter will serve as a guiding reference for our subsequent experimentation, enabling a deeper understanding of the

effectiveness and nuances of our approach to Ego-Localization.

Chapter 4

Testing our Methodology

4.1 Introduction

In the forthcoming chapter, we will present and analyze the results obtained from evaluating the various components of our methodology. Specifically, we will examine the performance of each component by testing them on our previously mentioned dataset and analyzing the outcomes of each step. This progression will ultimately lead us to our primary objective: visual ego-localization. Finally, we will discuss the overall results and their implications.

4.2 Bird's-Eye-View Mapping

As discussed in the preceding chapter, the Bird's-Eye-View Mapping component utilizes an Inverse Perspective Mapping (IPM) technique to bridge the gap between ground-level and aerial views, enhancing the effectiveness and robustness of the matching process. Figure 4.1 illustrates the results of several ground-view frames mapped into BEV images. These frames were sampled from the test sets to facilitate an analysis of the component's performance. The analysis is as follows:

1. Image **(a)** highlights the differences between the input and the output of the IPM process. Notably, the region of the road directly in front of the camera is mapped into a top-down view, resulting in clearer features that align more closely with the aerial view. Additionally, elements such as buildings, the sky, and distant objects are effectively removed. This reduction of irrelevant elements is advantageous, as such features are not visible in aerial views, thereby decreasing ambiguity and improving the mapping's relevance.

2. Images **(b)**, **(c)** and **(d)** demonstrate the presence of dynamic objects, such as cars and cyclists, which violate the assumption of a flat surface required for the IPM process. These moving objects are mapped into distorted shapes and often occlude parts of the road surface, making certain areas invisible. The impact of these distortions on the matching process will be analyzed in the subsequent section, as this aspect is critical to understanding the robustness of the overall system.

3. Images **(e)** and **(f)** exhibit a lack of road features, such as lane markings, sidewalks, and other distinctive elements. These images, referred to as featureless images, present a significant limitation. Even for a human observer, it is challenging to find matches between ground-view and aerial images in such cases. Moreover, localizing oneself accurately using only such ground frames and their corresponding aerial views is exceedingly difficult, highlighting a fundamental challenge in the process.

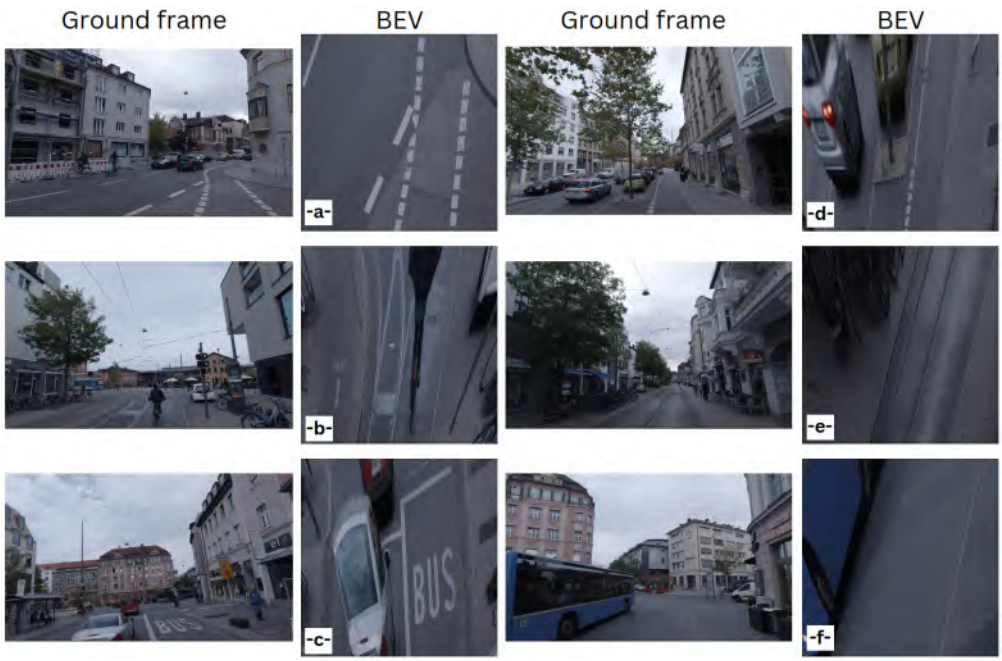

Figure 4.1: BEV Mapping Results

4.3 Aerial to Bird's-Eye-View Keypoint Matching

Moving on to Stage 2, the keypoint matching model, **LightGlue**, will attempt to identify matches between the aerial window cropped from georeferenced aerial images (based on GPS location and the preprocessing steps previously described) and the transformed BEV image. In Figure 4.2, we observe the output of LightGlue. For each pair of images, the left image represents the aerial window, while the right shows the BEV image. Detected points are connected by green lines, illustrating the matches.

From a preliminary analysis, it is evident that the chosen model effectively matches road features, such as lane markings, sidewalks, and other distinctive elements,

between both images in most cases. Diving deeper into the analysis:

1. In images **(a)**, **(b)**, and **(e)**, moving objects (e.g., cyclists or cars) appear directly in front of the ground camera. These objects are transformed into distorted shapes during the mapping process. As explained earlier, it is noteworthy that these distortions do not significantly affect the matching process, as the model still successfully matches most features.

2. In image **(d)**, the matching is nearly perfect despite the significant angular difference between features. However, it is important to note that deep learning-based descriptors are not inherently rotation-invariant. This implies that in extreme cases, the model might fail to match features due to rotational discrepancies.

3. Images **(c)** and **(f)** illustrate cases where BEV images contain fewer features. In **(f)**, only one matched keypoint is detected, which is insufficient for solving rotation, translation, and scaling through homography estimation, as this process requires a minimum of four points. Consequently, such images must be ignored in the pipeline, as they do not support reliable visual localization.

4. In image **(c)**, although four keypoints are present, relying on the minimal number of points for registration is not robust. This approach risks focusing on a specific region of the image while ignoring the rest. Given potential errors propagated through the IPM process from inaccuracies in camera intrinsic estimation, setting a threshold on the minimum number of matched keypoints is critical for reliable registration.

To conclude, our matching process inherently filters certain inputs, primarily due to a lack of sufficient features. Using the current pipeline, visual localization is not

feasible in such cases. Establishing thresholds ensures a more robust registration process, minimizing errors due to insufficient keypoints or feature quality.Nonetheless, the matching output is mostly efficient even in the presence of distorted shapes, demonstrating the robustness of the LightGlue model in handling challenging conditions.

Figure 4.2: Keypoint Matching Results

4.4 Aerial to Bird's-Eye-View Registration

Now moving on to the Registration Module, it is important to note that while registration is not our ultimate goal, it is a necessary step for estimating the camera's

projection location on the BEV grid. This projection is aligned with the georeferenced aerial view, enabling an estimation of the camera's location. In this section, we present the visual output of registering the BEV image onto the aerial window. Figure 4.3 illustrates the results of this registration process.

As observed, the road features matched in the previous step align well with their counterparts in the aerial view. Moreover, even unmatched features are mostly registered correctly due to their shared surface plane, which ensures consistent alignment. Theoretically this characteristic also applies to the estimation of the camera's projection location on the aerial view.

On the other hand, distorted objects are not registered effectively. For instance, the car in image **(a)** and the cyclist in image **(b)** demonstrate this limitation. These objects do not conform to the assumed flat-plane geometry, leading to inaccuracies in their registration.

Figure 4.3: Registration Results

4.5 Visual Localization Results

Guided step by step, we have now reached the final and most critical stage: localization prediction. In this section, we present the results exclusively for Sets 1, 2, 4, and 5. Set 3 is excluded as it was previously used as a validation set. During validation, we fixed the threshold for matched keypoints and optimized the configuration of LightGlue.

Table 4.1 provides a comparative analysis of the prediction and GPS errors across multiple test sets, highlighting the significant improvements achieved by the prediction method. The table reports the mean and median errors for both prediction and GPS measurements (in meters) for each set, along with the corresponding improvement percentages.

The **mean improvement percentage** is calculated as the reduction in mean error of the prediction method compared to GPS, expressed as a percentage of the GPS mean error. Similarly, the **median improvement percentage** reflects the reduction in median error. These metrics illustrate the effectiveness of the prediction approach in enhancing accuracy.

For instance, in Set 1, the prediction method reduced the mean error from 4.35m (GPS) to 1.38m, resulting in a mean improvement of 68.28%. The median error improvement was even more pronounced, reducing from 4.16m to 0.78m, corresponding to an improvement of 81.25%. Across all sets, the prediction method consistently outperformed GPS, with total mean and median improvement percentages of 57.08% and 73.11%, respectively. These results demonstrate the robustness of the prediction model in minimizing localization errors and outperforming GPS measurements.

Table 4.1: Mean and Median Errors for Prediction and GPS across Test Sets (in meters)

Set	**Prediction Errors**		**GPS Errors**		**Improvement Percentage**	
	Mean	**Median**	**Mean**	**Median**	**Mean**	**Median**
Set 1	1.38	0.78	4.35	4.16	68.28%	81.25%
Set 2	1.63	0.90	1.66	1.49	1.81%	39.60%
Set 4	0.66	0.36	5.33	5.28	87.61%	93.18%
Set 5	1.01	0.73	3.44	3.38	70.63%	78.40%
Total	**1.17**	**0.69**	**3.69**	**3.57**	**57.08%**	**73.10%**

Figure 4.4 illustrates the **Prediction and GPS Errors** (in meters) across four test sets: Set 1, Set 2, Set 4, and Set 5. Each subplot visualizes the prediction errors (green line) and GPS errors (blue line) for individual images within each set, providing a detailed comparison of localization accuracy.

- **Set 1**: The prediction method consistently outperforms GPS, maintaining significantly lower errors across most image indices. Notably, during the first five predictions, the GPS error exceeded 5 meters, while the prediction error remained stable at approximately 1 meter. This demonstrates the robustness of the prediction method in maintaining accuracy despite significant variations in GPS errors.
- **Set 2**: This set proved to be the most challenging, as the mean GPS error was the smallest at **1.66 meters**, leaving limited range for improvement. Despite this, the our prediction managed to slightly outperform GPS, achieving a **1.81% improvement** in mean error.
- **Set 4**: The prediction errors are remarkably stable and significantly lower than GPS errors, which remain consistently high throughout the set.
- **Set 5**: Although there is variability in errors, our prediction consistently outperforms GPS, achieving lower errors across most image indices and confirming its effectiveness.

Overall, these plots emphasize our method superiority over GPS in terms of localization accuracy, even in challenging conditions. Our prediction consistently demonstrates its capacity to reduce both mean and median errors across all test sets in our Dataset.

4.6 Limitations

A critical step in our pipeline is the matching process, which presents a significant challenge when the bird's-eye view image lacks features, as previously mentioned. In such cases, we are unable to proceed with the visual localization process.

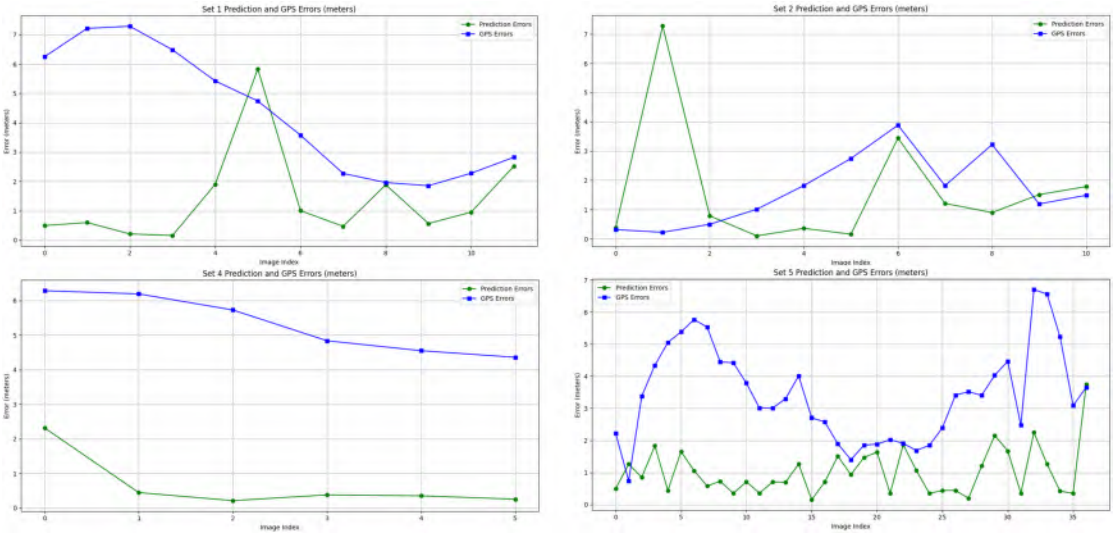

Figure 4.4: Prediction and GPS Errors

The table 4.2 above illustrates a limitation of our method. There are several factors contributing to this limitation. One significant issue is the presence of many featureless images. In order to localize these images in the aerial view, more than just front-view images are needed. This limitation is not exclusive to our setup, as many other works face similar challenges.

Additionally, our method incorporates a selective approach. We began by filtering out frames with a roll angle outside the range of -4 to 4 degrees. This filtering ensures that the bird's-eye view mapping produces acceptable outputs, as frames with significant roll angles might distort the BEV representation. However, this also reduces the number of frames that can be processed. Furthermore, in some cases, even when a BEV image contains features, LightGlue encounters difficulties in matching them, which halts further progression in our visual localization pipeline. This highlights some of the challenges and limitations of our approach, where featureless images and limitations in feature matching impact the effectiveness of the localization process.

Table 4.2: Frame Counts per Set

Set Number	**Total Frames**	**Predicted Frames**
Set 1	51	12
Set 2	59	11
Set 4	47	6
Set 5	220	37

4.7 Conclusion

In conclusion, the comprehensive evaluation of the different components within our method has provided valuable insights into their respective performances. Through meticulous testing on test sets, we gained a nuanced understanding of the strengths and limitations of the first and second components. Furthermore, our examination of the overall visual localization pipline in its distinct stages has offered a detailed perspective on its capabilities.

Chapter 5

Conclusion and Perspectives

5.1 Future Perspectives

The next phase of our research aims to explore specific refinements and expansions, shaping the future of our methodology. In this chapter, we will discuss both short-term and long-term perspectives for this work. In the short term, we have tested the effectiveness of certain refinements, such as adding a visual odometry component to our pipeline, which could help overcome some of its current limitations. Additionally, we propose an alternative methodology based on semantic segmentation matching, which could offer greater efficiency. The following two sections will present the potential impact of these refinements and expansions.

5.1.1 Visual Odometry Component

Visual Odometry (VO) is a technique used in robotics and computer vision to estimate the motion of a camera or vehicle by analyzing the sequence of images it captures. VO works by tracking the movement of visual features in consecutive frames, and by comparing the changes in their positions, it computes the relative displacement of

the camera. This allows the system to estimate the path or trajectory of the camera over time.

VO does not rely on external sensors like GPS but instead uses the camera's own images, making it particularly useful in environments where GPS signals are unavailable or unreliable (e.g., indoors or in dense urban areas). It can be used for applications such as autonomous driving, robot navigation, and augmented reality, often as a part of a larger localization and mapping system. To demonstrate the potential of this component, we tested a state-of-the-art visual odometry approach and reported the results for the Average Trajectory Error (ATE) in meters, as shown in Table 5.1. As we can observe, the Average Trajectory Error (ATE) for all the short sequences is approximately 1 meter or less. This demonstrates the potential of using such outputs to reconstruct the trajectory after performing cross-view matching. By leveraging this approach, we could predict the positions of all frames. However, this remains an open question regarding its effectiveness and implementation details. For more complex and longer sequences, the error increased significantly to 15 meters as we can see for set 5, highlighting the limitations of this method and the need for further research and development in this area.

Table 5.1: VO Absolute Trajectory Error (ATE) for Different Sets

Set Number	**VO ATE [m]**
Set 1	0.29
Set 2	1.16
Set 3	0.48
Set 4	0.37
Set 5	15.31

5.1.2 Segmentation Based Ego-localisation

For a long-term perspective, alternative approaches could potentially yield similar or even better results while addressing the limitations we currently face. One promising idea that emerged is the possibility of a segmentation-based cross-view matching approach. Instead of relying on keypoint matching between ground and aerial perspectives, this method utilizes semantic segmentation for matching. In scenarios where small, unique features are absent, the global structure of certain classes can assist in the matching process. However, this approach requires robust segmentation capabilities for both aerial and ground views to ensure its effectiveness.

As we successfully tackled our challenges with prepared data, it becomes evident that a labeled dataset is essential to initiate research in this area. Unfortunately, the currently available datasets lack the necessary specifications. To address this limitation, alternative solutions, such as using a synthetic simulator to generate the required data, could pave the way for advancing research and opening new possibilities in this field.

5.2 Report Summary

In this research endeavor, we conducted a comprehensive exploration of ego-localization through the perspective of object matching in aerial and ground view images. Our journey began with a thorough examination of the broader context and the complexities of the problem landscape. We delved into the academic framework of our internship, outlined the roles of the host organizations, and highlighted the challenges that served as the driving force behind this project. Alongside this, we reviewed prior efforts to tackle these challenges, laying the groundwork for our unique contributions.

Our first contribution, detailed in the chapter titled *"Kokovi Dataset,"* focused on refining the output of the acquisition campaign into a ready-to-use dataset tailored for the challenge at hand. The unique specifications of the dataset played a pivotal role in shaping our results and achievements. This highlights not only the significance of our contribution but also the dedication and commitment of the host organization to this field, as demonstrated by their investment and support for this project.

The subsequent chapter, "Defining Our Localization Methodology," takes a step-by-step approach to exploring the theoretical frameworks employed and the rationale behind each choice. This thorough examination was designed to maximize the likelihood of achieving our final goal while ensuring the methodology could generalize effectively to address the broader problem.

The final phase of our exploration is detailed in "Testing Our Methodology," a chapter devoted to the empirical validation of each component. Through rigorous testing on our Dataset, the results demonstrated significant performance, while also shedding light on the key limitations we encountered along the way.

In summary, our journey through Ego-Localization has culminated in the development of a novel methodology composed of multifaceted components. From contributing a unique dataset to designing an innovative approach and conducting rigorous testing, our work brings both depth and breadth to the ongoing discussions in this field. As we reflect in this concluding chapter, the challenges overcome, lessons learned, and potential for future advancements collectively illustrate the richness of the path we have navigated.

References

[1] DLR. *DLR Introduction*. URL: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-5277/8858_read-15912/ (cit. on pp. 2, 17).

[2] CRNS. *CRNS Introduction*. URL: https://crns.rnrt.tn/ (cit. on pp. 2, 5).

[3] DLR. *IMF Introduction*. URL: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-5279/8913_read-16239/ (cit. on pp. 3, 17).

[4] DLR. *Photogrammetry and Image Analysis Department Introduction*. URL: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-5297/8940_read-16255/ (cit. on pp. 4, 17).

[5] DLR. *Traffic Monitoring Team Introduction*. URL: https://www.dlr.de/eoc/en/desktopdefault.aspx/tabid-19231/30774_read-84310/ (cit. on pp. 5, 17).

[6] Witold Kazimierski Andrzej Stateczny and Pawel Burdziakowski. 'Sensors and System for Vehicle Navigation'. In: *PMC* (2022). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8914871/ (cit. on p. 6).

[7] STEVEN ASHLEY. 'Centimeter-accurate GPS for self-driving vehicles'. In: *SAE International* (2016). URL: https://www.sae.org/news/2016/10/centimeter-accurate-gps-for-self-driving-vehicles (cit. on p. 6).

[8] Michael Wing. 'Consumer-Grade Global Positioning Systems (GPS) Receiver Performance'. In: *Journal of Forestry* 106 (2008), pp. 185–190 (cit. on p. 7).

[9] V Akgul et al. 'Effects of the High-Order Ionospheric Delay on GPS-Based Tropospheric Parameter Estimations in Turkey'. In: *Remote Sensing* 12.21 (2020), p. 3569. URL: https://www.mdpi.com/2072-4292/12/21/3569 (cit. on p. 7).

[10] Thomas Windholz Kindra Serr and Keith Weber. 'Comparing GPS Receivers: A Field Study'. In: *ISU's GIS TReC - Idaho State University* (2006).

URL: https://giscenter.isu.edu/research/projects/comparinggps_urisavol18no2.pdf (cit. on p. 7).

[11] Mohammad Reza Mosavi Amir Tabatabaei. 'MP Mitigation in Urban Canyons using GPS-combined-GLONASS Weighted Vectorized Receiver'. In: *IET Signal Processing* (2017). URL: https://ietresearch.onlinelibrary.wiley.com/doi/full/10.1049/iet-spr.2016.0372 (cit. on p. 7).

[12] Simegnew Yihunie Alaba. 'GPS-IMU Sensor Fusion for Reliable Autonomous Vehicle Position Estimation'. In: *Mississippi State University* (2024). URL: https://arxiv.org/abs/2405.08119 (cit. on p. 8).

[13] Hendrik Hellmers and Andreas Eichhorn. 'Indoor localisation for wheeled platforms based on IMU and artificially generated magnetic field'. In: *Ubiquitous Positioning Indoor Navigation and Location Based Service (UPINLBS)* (2014). URL: https://ieeexplore.ieee.org/abstract/document/7033735 (cit. on p. 8).

[14] Alberto Y. Hata and Denis Fernando Wolf. 'Feature Detection for Vehicle Localization in Urban Environments Using a Multilayer LIDAR'. In: *IEEE Transactions on Intelligent Transportation Systems* (2016). URL: https://ieeexplore.ieee.org/abstract/document/7279128 (cit. on p. 9).

[15] Bochun Yang and Zijun Li. 'High-Precision Positioning, Perception and Safe Navigation for Automated Heavy-Duty Mining Trucks'. In: *Computer Vision and Pattern Recognition* (2024). URL: https://openaccess.thecvf.com/content/CVPR2024/html/Yang_LiSA_LiDAR_Localization_with_Semantic_Awareness_CVPR_2024_paper.html (cit. on p. 9).

[16] Ryan W. Wolcott and Ryan M. Eustice. 'Visual localization within LIDAR maps for automated urban driving'. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2014). URL: https://ieeexplore.ieee.org/document/6942558 (cit. on p. 9).

[17] Hang Wu and Zhenghao Zhang. 'MapLocNet: Coarse-to-Fine Feature Registration for Visual Re-Localization in Navigation Maps'. In: *Computer Vision and Pattern Recognition* (2014). URL: https://arxiv.org/abs/2407.08561 (cit. on p. 9).

[18] Long Chen and Yuchen Li. 'High-Precision Positioning, Perception and Safe Navigation for Automated Heavy-Duty Mining Trucks.' In: *IEEE Transac-*

tions on Intelligent Vehicles (2014). URL: https://ieeexplore.ieee.org/abstract/document/10465630 (cit. on p. 9).

[19] Matan Friedmann Eli Brosh and Ilan Kadar. 'Accurate visual localization for automotive applications'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* (2019). URL: https://arxiv.org/abs/1905.03706 (cit. on p. 9).

[20] Amir Roshan Zamir and Mubarak Shah. 'Accurate image localization based on google maps street view'. In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision* (2010). URL: https://web.stanford.edu/class/cs231m/lectures/lecture-13-geo.pdf (cit. on p. 9).

[21] Taojiannan Yang Sijie Zhu and Chen Chen. 'Vigor: Cross-view image geo-localization beyond one-to-one retrieval'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021). URL: https://arxiv.org/abs/2011.12172 (cit. on p. 9).

[22] Olaf Booij Zimin Xia and Julian F. P. Kooij. 'Convolutional cross-view pose estimation'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023). URL: https://arxiv.org/abs/2303.05915 (cit. on p. 9).

[23] Xiaolong Wang et al. 'Fine-Grained Cross-View Geo-Localization Using a Correlation-Aware Homography Estimator'. In: *Advances in Neural Information Processing Systems* (2024). URL: https://arxiv.org/pdf/2308.16906 (cit. on p. 10).

[24] Shan Wang et al. 'View Consistent Purification for Accurate Cross-View Localization'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023). URL: https://arxiv.org/abs/2308.08110 (cit. on p. 10).

[25] Christoph Stiller Andreas Geiger Philip Lenz and Raquel Urtasun. 'Vision meets robotics: The KITTI dataset'. In: *IJRR* (2013). URL: https://www.cvlibs.net/publications/Geiger2013IJRR.pdf (cit. on p. 14).

[26] Yujiao Shi and Hongdong Li. 'Beyond Cross-view Image Retrieval: Highly Accurate Vehicle Localization Using Satellite Image'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2022). URL: https://arxiv.org/pdf/2204.04752 (cit. on p. 14).

[27] Shan Wang et al. 'View Consistent Purification for Accurate Cross-View Localization'. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023). URL: https://arxiv.org/abs/2308.08110 (cit. on p. 15).

[28] Mapillary. *Mapillary: Street-Level Imagery Platform*. https://www.mapillary.com/ (cit. on p. 19).

[29] F. Kurz et al. 'GENERATION OF REFERENCE VEHICLE TRAJECTORIES IN REAL-WORLD SITUATIONS USING AERIAL IMAGERY FROM A HELICOPTER'. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-1-2022 (2022), pp. 221–226. DOI: 10.5194/isprs-annals-V-1-2022-221-2022. URL: https://isprs-annals.copernicus.org/articles/V-1-2022/221/2022/ (cit. on p. 21).

[30] Alexander Veicht et al. *GeoCalib: Learning Single-image Calibration with Geometric Optimization*. 2024. arXiv: 2409.06704 [cs.CV]. URL: https://arxiv.org/abs/2409.06704 (cit. on p. 30).

[31] Qiyan Luo et al. *Comparative Analysis of Advanced Feature Matching Algorithms in Challenging High Spatial Resolution Optical Satellite Stereo Scenarios*. 2024. arXiv: 2405.06246 [cs.CV]. URL: https://arxiv.org/abs/2405.06246 (cit. on p. 38).

[32] Daniel DeTone, Tomasz Malisiewicz and Andrew Rabinovich. *SuperPoint: Self-Supervised Interest Point Detection and Description*. 2018. arXiv: 1712.07629 [cs.CV]. URL: https://arxiv.org/abs/1712.07629 (cit. on pp. 38, 39).

[33] Philipp Lindenberger, Paul-Edouard Sarlin and Marc Pollefeys. *LightGlue: Local Feature Matching at Light Speed*. 2023. arXiv: 2306.13643 [cs.CV]. URL: https://arxiv.org/abs/2306.13643 (cit. on pp. 38, 39).