DISSERTATIONES
INFORMATICAE
UNIVERSITATIS
TARTUENSIS
**67**

ALIREZA AKHAVI ZADEGAN

# A Multimodal Approach for Refining Mapping and Localization by Integrating Generative AI and Pedestrian-Centric Data

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

**67**

DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS
**67**

**ALIREZA AKHAVI ZADEGAN**

# A Multimodal Approach for Refining Mapping and Localization by Integrating Generative AI and Pedestrian-Centric Data

UNIVERSITY OF TARTU
Press
1632

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on May 6, 2025 by the Council of the Institute of Computer Science, University of Tartu.

*Supervisor*

| | |
|---|---|
| Assoc. Prof. | Amnir Hadachi<br>University of Tartu, Estonia |

*Opponents*

| | |
|---|---|
| Assoc. Prof. | Alain Kibangou<br>University Grenoble Alpes, France |
| Assis. Prof. | Salvatore Flavio Pileggi<br>University of Technology Sydney, Australia |

The public defense will take place on May 30th at 10:15 in Narva Rd. 18-1018.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print) ISSN 2806-2345 (pdf)
ISBN 978-9916-27-882-6 (print) ISBN 978-9916-27-883-3 (pdf)



University of Tartu Press
http://www.tyk.ee/

*To my loving parents*

# ABSTRACT

As urban environments continue to evolve and grow in complexity, accurate, up-to-date, and pedestrian-aware mapping has become a critical enabler for autonomous systems, smart mobility solutions, and urban planning. This thesis presents a comprehensive research effort that introduces original methodologies, platforms, and datasets designed to improve the granularity, adaptability, and contextual understanding of urban maps — with a particular focus on sidewalks, crosswalks, and pedestrian-centric infrastructure.

A key contribution of this work is the design and development of the DELTA platform — a fully customized multi-sensory data acquisition system built from scratch within the scope of this thesis. The platform integrates and synchronizes diverse sensing modalities, including high-resolution visual data, LiDAR, GNSS/IMU, and environmental audio, enabling the collection of rich, multi dimensional data along pedestrian routes. The resulting DELTA dataset, which also functioned as a benchmarking dataset, provides a valuable foundation for high-definition spatial analysis of often-overlooked pedestrian environments which also served for benchmarking

Building upon this dataset, the thesis proposes two novel methodological frameworks to advance urban mapping and localization, street2sat — a generative AI-based framework that leverages landmark segmentation to synthesize satellite-like views from ground-level imagery. This approach bridges the visual gap between street-level and aerial perspectives, improving localization accuracy and enhancing contextual awareness in complex urban scenarios. Street2GIS — an automated framework for generating GIS-ready shapefiles from monocular street-view images. By combining depth estimation, semantic segmentation, and cross-view synthesis, Street2GIS enables the rapid and scalable creation of georeferenced representations of key urban features, including roads, sidewalks, buildings, and vegetation. Both frameworks were evaluated using a combination of the DELTA dataset and captured data. The evaluation methodology involved quantitative metrics — such as localization accuracy, similarity measures, and segmentation performance — as well as qualitative analysis to assess robustness across diverse urban conditions. The results demonstrate the effectiveness of the proposed approaches in enhancing spatial accuracy, automation, and situational awareness, with significant potential benefits for autonomous systems, urban analytics.

Overall, this thesis contributes not only new algorithms and tools but also an end-to-end methodological framework for advancing pedestrian-centric urban sensing and mapping. The research outcomes aim to support the creation of safer, more accessible, and data-driven smart cities.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

## Acronyms

**BoW** Bag-of-Words. 23

**cGAN** Generative Adversarial Network. 82
**CNN** Convolutional Neural Network. 28

**DELTA** Digital Enhancement for Localization in Tartu Area. 30
**DoF** Degrees of Freedom. 25

**FFT** Fast Fourier Transform. 71
**FPN** Feature Pyramid Network. 80

**GANs** Generative Adversarial Networks. 93
**GIS** Geographic Information System. 76

**IBL** Image-Based Localization. 21
**IMU** Inertial Measurement Unit. 44

**LiFePO4** Lithium Iron Phosphate. 34

**MAE** Mean Absolute Error. 89
**mAP** mean Average Precision. 82
**MLP** Multilayer Perceptron. 24
**MVS** Multi-view Stereo. 21

**NCC** Normalized Cross-Correlation. 88
**NeRF** Neural Radiance Fields. 24
**NLP** Natural Language Processing. 63

**ORB** Oriented FAST and Rotated BRIEF. 102

**P3P** three-Point Perspective. 26
**PAN** Path Aggregation Network. 80
**PCA** Principal Component Analysis. 27
**PnP** Perspective-n-Point. 26
**PPK** Post-Processed Kinematics. 44

**RANSAC** Random Sample Consensus. 26
**RMSE** Root Mean Square Error. 89
**RTK** Real-Time Kinematic. 43

**SAM** Segment Anything Model. 67
**SATs** Summed-Area Tables. 86
**SfM** Structure from Motion. 21
**SLAM** Simultaneous Localization and Mapping. 21
**SSIM** Structural Similarity Index. 103

**TSDF** Truncated Signed Distance Function. 24

# LIST OF ORIGINAL PUBLICATIONS

## Publications in the scope of the thesis

I **Alireza Akhavi Zadegan**, Damien Vivet and Amnir Hadachi. "DELTA: integrating multimodal sensing with micromobility for enhanced sidewalk and pedestrian route understanding." Sensors 24.12 (2024): 3863. https://doi.org/10.3390/s24123863

**Scientific contribution:** We introduced the DELTA dataset, a resource designed to improve pedestrian zone analysis in urban areas. We developed a modular e-scooter platform that captures high-resolution data across various modalities, including 4K monocular and stereocameras for visual data, LiDAR for 3D point clouds, reflectivity, and range channel data, urban soundscapes for audio, and GNSS/IMU for positioning and motion tracking. The DELTA dataset includes sequences of stereo RGB images, depth maps, and point clouds generated from disparity maps for detailed visual representations, alongside sequences of 3D point clouds, reflectivity, and range channel data from LiDAR for precise spatial measurements. Additionally, it features raw GNSS data and IMU measurements for accurate positioning and motion tracking. We also developed three pedestrian route segmentation models tailored to each sensor: 4K camera, stereocamera, LiDAR. Additionally, the study explored audio event-based classification to link unique soundscapes with specific geolocations, enriching the spatial understanding of urban environments.

**Author's contributions:** The author designed, developed, and built the e-scooter-based multi-sensor data acquisition platform, carried out the data collection, calibration, synchronization, and filtering of data from all integrated sensors. The author also manually labeled the images for each data source, formulated the main algorithmic ideas and hypotheses, devised and implemented the algorithms, and defined the evaluation metrics and results and wrote the paper.

II **Alireza Akhavi Zadegan** and Amnir Hadachi. "Generative-AI based Map Representation and Localization." In: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances in Urban-AI 2024*. Pp.34-42 https://doi.org/10.1145/3681780.3697276

**Scientific contribution:** We introduced a novel framework that leverages generative AI to enhance map representation and localization processes. This approach aims to improve the accuracy and efficiency of geospatial data interpretation by utilizing advanced AI models to generate and refine map representations.

**Author's contributions:** The author developed the dataset, labeled and trained a custom landmark segmentation model, raised the main algorithm

ideas and hypotheses, devised the algorithms, defined the results to be calculated and evaluated, implemented the code, designed data visualization, wrote the paper.

III **Alireza Akhavi Zadegan**, Jose Medina and Amnir Hadachi. "Street2GIS: Multimodal Generative Framework for Pedestrian Infrastructure Mapping." In: *Proceedings of the 17th International Conference on Joint Urban Remote Sensing (JURSE) 2025*. Status: Accepted.

**Scientific contribution:** We introduced Street2GIS, an open-source multimodal framework for automating the generation of GIS shapefiles for pedestrian infrastructure. By integrating multiview and cross-cue data fusion, our approach improves the accuracy and completeness of urban mapping. Using hierarchical transformation networks, Street2GIS combines depth maps, semantic segmentation, and elevation data to extract and georeference urban features.

**Author's contributions:** The author developed the dataset, raised the main algorithm ideas and hypotheses, devised the algorithms, defined the results to be calculated and evaluated, implemented the code, designed data visualization, wrote the paper.

## Publications out of the scope of the thesis

IV **Alireza Akhavi Zadegan**, Damien Vivet and Amnir Hadachi. "Challenges and Advancements in Image-based 3D Reconstruction of Large-Scale Urban Environments: A Review of Deep learning and Classical methods." Frontiers in Computer Science. Status: under review.

**Scientific contribution:** We provide a comprehensive review of image-based 3D reconstruction techniques for large-scale urban environments, distinguishing between traditional and deep learning approaches. Our study systematically compares these methods, evaluates their performance, and identifies their strengths and limitations. Additionally, we highlight commonly used 3D datasets and performance metrics in this domain. By outlining key challenges and future research directions, this work contributes to the advancement of scalable and precise urban modeling solutions for smart city applications and digitalization.

**Author's contributions:** The author conceptualized the study, formulated the key hypotheses, and developed the classification framework, conducted the systematic review, analyzed existing image-based 3D reconstruction techniques, and evaluated their methodologies. Additionally, the author identified relevant datasets and performance metrics, synthesized findings, and outlined future research directions, wrote the paper and designed the data visualizations.

# 1. INTRODUCTION

## 1.1. Overview and motivation

The rapid pace of urbanization has significantly transformed urban mobility infrastructure, requires the development of intelligent systems capable of navigating complex environments. As cities become more densely populated, autonomous systems—including micromobility solutions like e-scooters and e-bikes, as well as delivery robots and self-driving vehicles are being integrated to enhance efficiency, reduce congestion, and improve the overall quality of life. However, critical gaps remain in how these systems perceive, interpret, and adapt to pedestrian-centric environments such as sidewalks, crosswalks, and shared paths [AG24]. Unlike vehicular roads, pedestrian environments are less structured more unpredictable, posing unique challenges for autonomous navigation. Factors such as unpredictable human behavior, varying sidewalk conditions, and the presence of diverse street elements introduce complexities that are difficult for autonomous systems to navigate. While vehicular networks have been extensively mapped, pedestrian spaces often lack comprehensive mapping, leading to navigation difficulties. Autonomous systems require advance perception capabilities to interpret complex pedestrian environments, including recognizing and predicting human behaviors and understanding social cues. Additionally, the lack of standardized regulations governing the operation of autonomous systems in pedestrian spaces creates uncertainties in system design and deployment, hindering widespread adoption [Mav+23].

Autonomous systems face significant challenges in pedestrian environments, with shortcomings that impact safety, efficiency, and urban planning. Inadequate perception and decision-making can lead to accidents, endangering both pedestrians and autonomous vehicles, while unreliable navigation limits the benefits of micromobility and autonomous delivery services. Furthermore, urban planners lack the data and tools needed to design spaces that seamlessly integrate human and autonomous traffic [MFD24; RL24]. A multidisciplinary approach is essential to overcome these challenges [Guo+21]. First, creating high-definition maps of pedestrian areas using crowdsourcing, LiDAR scanning, and computer vision can provide a critical data foundation. Second, advancing machine learning algorithms to better interpret complex human behaviors and unstructured environments will enhance system adaptability. Finally, collaboration among urban planners, technologists, and policymakers is key to redesigning pedestrian spaces and establishing clear guidelines and standards for the safe operation of autonomous systems. By addressing these issues through improved mapping, advanced perception algorithms, cooperative urban planning, and regulatory standardization, autonomous systems can more effectively navigate the complexities of pedestrian environments. This progress is vital for realizing the promise of safer, more efficient, and livable cities.

## 1.2. Research questions

To address the challenges outlined in section 1.1, this thesis explores the following research questions:

1. How can multimodal sensing technologies improve the granularity and contextual understanding of pedestrian environments for urban analytics?
2. What role can generative AI play in bridging different perspectives (ground and satellite) to facilitate accurate and adaptive mapping and localization in dynamic urban settings?
3. How can automated GIS data generation frameworks advance the mapping of pedestrian infrastructure to support autonomous navigation and urban planning?

## 1.3. Contributions

This thesis collectively addresses the limitations of traditional urban mapping and localization methods, proposing innovative, scalable, and data-driven solutions to enhance urban mobility systems. Hence, the main contributions can be summarized as follows:

1. The DELTA project introduces a novel multimodal sensing platform for comprehensive pedestrian route analysis. It leverages synchronized data from visual, LiDAR, audio, and GNSS/IMU sensors, along with dedicated preprocessing pipelines and specialized tools, to generate rich datasets tailored for sidewalk and pedestrian environments [AVH24].
2. The street2sat framework employs generative AI to bridge the observational gap between ground and satellite perspectives, enhancing localization and mapping accuracy in urban settings [ZH24].
3. The Street2GIS framework advances pedestrian infrastructure mapping by automating GIS data generation from street-view imagery, integrating cutting-edge techniques like semantic segmentation and cross-view synthesis to produce precise, georeferenced shapefiles [ZMH25].

## 1.4. Thesis structure

This PhD thesis unfolds across six constructed chapters, each contributing to a comprehensive exploration of key aspects of urban mapping, localization, and the integration of AI techniques The structure of the thesis is designed to guide the reader through a progressive understanding of the subject matter, from foundational concepts to applications and future research directions. The remainder of the manuscript is structured as follows.

Chapter 2 explores the current landscape of mapping and localization technologies, offering an in-depth analysis of map representations and image-based localization techniques. This chapter explores geometric, topological, feature-based, and implicit map representations, alongside classical and deep learning-based approaches to image localization, providing a comprehensive context for the research.

Chapter 3 focuses on the design and implementation of a versatile data collection platform. It details the hardware components, including data collection mechanisms, sensor interfaces, power management systems, and damping solutions, as well as software tools for sensor configuration and integration. This chapter emphasizes the technical innovation involved in building a robust multi-sensory platform for urban data acquisition.

Chapter 4 introduces the DELTA dataset, a comprehensive collection of multi-modal data tailored for urban analysis. It discusses the importance of pedestrian-centric data, sensor synchronization, and fusion methods. The chapter also explores data annotation and segmentation techniques for urban features, including sidewalks, roads, vegetation, and auditory events, showcasing the dataset's applicability in urban mobility research.

Chapter 5 presents the core contributions of the thesis, detailing two innovative frameworks: street2sat and Street2GIS. The street2sat framework leverages generative AI and landmark segmentation for urban mapping and localization, employing template matching and tiled image-maps for precise geolocation. The Street2GIS framework automates GIS data generation, integrating depth estimation, semantic segmentation, and raster-to-polygon conversion to produce accurate urban feature maps. This chapter evaluates both frameworks, providing insights into their performance and addressing challenges through failure analysis.

Chapter 6 summarizes the thesis's key contributions to AI-driven urban mapping and localization, emphasizing their impact on spatial analysis and navigation. It outlines future directions such as dataset expansion and model refinement, while also addressing real-world deployment challenges, including computational constraints, data privacy concerns, and integration with municipal GIS systems.

# 2. STATE OF THE ART

## 2.1. Overview of map representations

In the domain of Image-Based Localization (IBL), especially within complex urban landscapes, the selection of an appropriate map representation is critical for accurately determining a camera's location from an image. This selection is influenced by the inherent challenges presented by variability in lighting, perspective, and dynamic elements like moving objects. To navigate these challenges, IBL utilizes a spectrum of map representations, each tailored to leverage different aspects of spatial information and image features. The comprehensive categorization of these map representations includes geometric, topological, and feature-based representations, each offering unique benefits and facing specific challenges.

### 2.1.1. Geometric representations

In the domain of IBL, the creation and utilization of 3D metric maps through structure-based localization represent a cornerstone technique, harnessing various sophisticated methodologies to accurately model the environment. These techniques encompass Structure from Motion (SfM) [Akb+06], which reconstructs the 3D structure from sequences of images without prior information about the scene; Multi-view Stereo (MVS) [PKF07], which estimates depth information for each pixel across multiple image views; and Simultaneous Localization and Mapping (SLAM) [CN11; Dav+07; Dis+01], a dynamic process that concurrently maps the environment while estimating the robot's or device's location and orientation within it. The enrichment of these maps is achieved through the generation of point clouds—dense aggregations of 3D points that outline the environment's surface—and mesh representations that connect these points to form structured surfaces. Additionally, feature markers like SIFT [Low04], SURF [BTV06], and ORB [Rub+11] are employed to identify and describe unique points in images, facilitating accurate matching and localization.

However, the deployment of 3D metric maps is not without challenges. The computational cost is significant, requiring substantial processing power and resources. Adapting to dynamic environments and addressing occlusions, where objects or lighting may obscure key features, further complicates the accuracy and reliability of these maps. Despite these hurdles, the applications of 3D metric maps are vast, ranging from indoor robot navigation within complex spaces like warehouses [BR13] to augmented reality experiences [MEM+20] that seamlessly superimpose virtual objects onto the physical world, and detailed 3D reconstruction and modeling projects.

Conversely, 2D maps offer a different perspective, emphasizing efficiency and scalability over the detailed depth information provided by their 3D counterparts. These maps include satellite images [Hu+18; HL20], which offer broad coverage with relatively lower resolution; aerial photography [Cai+19], providing detailed

views of smaller areas with higher resolution; and planimetric maps [SZC20; Seo+18], which are widely available and informative, albeit lacking in elevation data. The strengths of 2D maps lie in their efficiency, ease of visualization and sharing, and adaptability to changes in the environment, making them particularly suitable for applications like outdoor navigation, route planning and logistics, and large-scale environmental monitoring. However, they may fall short in delivering the precision required for specific localization tasks, especially in environments with repetitive textures or complex terrains.

In an effort to leverage the best of both worlds, hybrid approaches that combine elements of 3D and 2D map representations have emerged, offering a balanced solution between the high accuracy of detailed 3D models and the efficiency and scalability of 2D maps. The choice between these map representations, be it 3D, 2D, or a hybrid form, ultimately hinges on the specific needs and constraints of the application at hand, underscoring the importance of tailored approaches in the field of IBL.

### 2.1.2. Topological representations:

Topological maps [GO15; FEN07] present a unique approach to understanding environments by focusing on the connectivity and spatial relationships rather than capturing precise geometric details. These maps can be likened to simplified subway maps [Mou14] that highlight stations and their connections without delving into exact distances or platform shapes. The essence of topological mapping lies in its representation through nodes, which symbolize key locations such as landmarks and intersections, and edges that link these nodes, signifying possible movements between them. Unlike 3D metric maps, topological maps do not concern themselves with the physical characteristics of these connections, such as specific distances or angles.

The advantages of employing topological maps are manifold. They are notably efficient, boasting a smaller data size that facilitates faster processing and storage compared to their 3D counterparts. This efficiency also extends to scalability, allowing these maps to easily accommodate large environments without succumbing to overwhelming complexity. Additionally, topological maps are highly adaptable to dynamic environments, capable of adjusting to changes in appearance, such as furniture rearrangements, as long as the fundamental connectivity remains intact. Furthermore, they enable symbolic reasoning, supporting path planning and navigation through the use of graph search algorithms.

However, this mapping approach is not without its challenges. The lack of precise location information within a node can pose difficulties for tasks requiring fine-grained precision. The potential for geometric ambiguity also arises, making it challenging to distinguish between similar-looking nodes without supplementary information. The localization procedure is further complicated by the requirement for strong feature extraction and matching methods.

Several techniques have been developed to construct and utilize topological maps effectively. Landmark-based methods [Ghr+12] identify nodes with easily recognizable landmarks, with edges determined by visibility or proximity. Grid-based [JYM14] approaches partition the environment into a grid, placing nodes at grid intersections and connecting them based on adjacency. Another method involves the use of Voronoi diagrams [HYS19], where nodes are positioned at the centroids of regions closest to specific landmarks, with edges delineating the boundaries between these regions.

### 2.1.3. Feature-based representations

In the realm of IBL, maps that prioritize distinctive visual elements offer a strategic balance between computational efficiency and the necessary richness of detail for accurate localization. These maps focus on leveraging unique, recognizable features within the environment, such as building facades, statues, or specific signage, thus requiring less processing power and storage compared to exhaustive 3D metric mapping. This approach is adaptable to large and dynamically changing environments and exhibits robustness against variations in lighting and occlusions.

The Bag-of-Words (BoW) model [BMG11] stands out as a principal method in this context, abstracting images into collections of visual words. This method draws a parallel with document representation by text words, involving the extraction of local image features, clustering these features into "words" based on similarity, and then representing images as histograms of these word occurrences. The BoW model is noted for its efficiency in image retrieval within large databases and its relative robustness to partial occlusions. It manages to capture some spatial relationships between features, although challenges remain in the accuracy of feature detection and the need for meticulous vocabulary design. This model may overlook specific landmark details, necessitating careful consideration in its application. The BoW model, in particular, highlights the versatility and potential of IBL systems that focus on unique environmental features, contributing significantly to advancements in navigation, augmented reality, autonomous driving, and digital content retrieval. This streamlined approach aligns with the evolving needs of computational efficiency and detailed environmental understanding in the modern landscape of urban mobility and planning.

### 2.1.4. Implicit map representations

Implicit mapping has emerged as a transformative approach in robotics and computer vision, leveraging deep learning to navigate and understand complex environments directly from raw sensory inputs, such as images or videos. This methodology represents a paradigm shift from traditional navigation techniques that rely on explicit feature extraction and landmark identification, towards more adaptive and scalable solutions that can automatically infer spatial layouts and key features necessary for navigation.

The foundation of implicit mapping lies in data-driven models trained on extensive datasets of environmental imagery. These models, utilizing deep learning algorithms, learn to encode spatial information in a continuous, memory-efficient manner that is not predefined, enabling the system to identify crucial navigational cues solely based on visual inputs. This capability is particularly invaluable in environments lacking distinct landmarks or in dynamic settings, where traditional systems might falter.

Implicit maps facilitate global localization, allowing systems to determine their location within an environment from a single image, without the need for a known starting point or continuous location tracking. However, the dependency on large, scene-specific training datasets presents a notable challenge, requiring retraining or fine-tuning for new environments and thus posing scalability issues.

The integration of implicit mapping with traditional methods, such as surfel-based representations [BS18], triangle meshes [Gur+11], and octree-based occupancy representations [Ves+18], alongside advanced techniques like theTruncated Signed Distance Function (TSDF) [Par+19] and Neural Radiance Fields (NeRF) [Mil+21], highlights a growing trend towards utilizing neural representations for mapping and reconstruction. These neural approaches, capable of rendering novel views or scene geometries, signify a shift towards more dynamic mapping solutions. Recent developments in implicit neural mapping, including iMap [Suc+21], iSDF [Ort+22], and Neural RGBD [Azi+22], have shown promise in representing entire indoor scenes with a single Multilayer Perceptron (MLP). Yet, challenges such as the limited model capacity for larger environments and the high memory costs of dense voxel structures remain. Solutions like NICE-SLAM [Zhu+22] and Go-SURF [WBA22], which combine shallow MLPs with optimizable local feature grids, offer improvements in surface reconstruction for larger scenes but still grapple with scalability and memory efficiency.

Looking forward, the future of implicit mapping lies in enhancing the adaptability and robustness of these systems through hybrid approaches that merge the strengths of various methodologies. Increasing the interpretability of implicit map-based systems is essential for integration into safety-critical applications, where user trust is paramount. Furthermore, improving data efficiency through techniques like transfer learning, domain adaptation, and few-shot learning is critical for reducing reliance on extensive training datasets and expanding the applicability of implicit maps to a broader range of environments.

## 2.2. Overview of image based localization

In the evolving field of IBL, particularly within urban environments, the complexity of accurately identifying a location from an image is compounded by the inherent variability in lighting, perspective, and dynamic elements such as moving objects. This variability challenges traditional pixel-by-pixel image comparison methods, necessitating more sophisticated approaches that utilize both local and

global image features. These features not only provide a more robust basis for comparison but also have the advantage of requiring less storage space than full-resolution images.

Urban settings, with their dense architectural structures and frequent obstructions of satellite signals, present unique challenges and opportunities for IBL systems. Applications range from pedestrian localization, where buildings may block GPS signals, to augmented reality and robotics, which aim to enhance human navigation and provide autonomous vehicle guidance within these complex environments. The focus on urban spaces is further justified by the abundance of detailed datasets capturing cityscapes and road networks, making them more accessible for research and application development compared to rural areas.

IBL methodologies split into two primary categories: indirect and direct methods. Indirect methods treat localization as an image retrieval problem, providing an approximate location based on the closest match within a database of known locations. This approach contrasts with direct methods, which aim to precisely calculate the six Degrees of Freedom (DoF) pose of the camera or imaging system from the image itself. These methods differ fundamentally in their handling of imagery—focusing on scenes rather than objects—and in their evaluation criteria, prioritizing precision in identifying the exact location.

Extensive reviews and studies have shed light on the myriad approaches to IBL, highlighting the field's breadth and the specific challenges of city-scale localization. The significance of IBL in contemporary research is underscored by its frequent discussion in recent scholarly articles and its prominence in workshops and tutorials at major conferences. This attention reflects the critical role of IBL in advancing navigation and localization technologies, particularly as they apply to the complex and dynamic nature of urban landscapes.

In essence, a localization system is made up of three key components: the representation of features, the creation of a map, and the process of matching. The approach to localization can vary significantly based on the choice of mapping technique. The following sections explore global localization approaches, including classical and deep learning-based methods.

### 2.2.1. Classical approach

Structure-based localization methods [LLD17; JS11; Tan+21a] employ 3D models to establish correspondences between specific locations within these models and corresponding regions in images [Dav+07]. These correspondences are often established using feature markers, such as SIFT [Low04], SURF [BTV06], ORB [Rub+11], BRIEF [Cal+10], BRISK [LCS11], FREAK [AOV12].

The construction of 3D maps is typically accomplished through techniques such as SfM [Akb+06], multiview reconstruction [MP07], SLAM [CN11; Dav+07; Dis+01], or by augmenting existing 3D models with visual markers [ARS14; Sib+13]. Structure-based localization, utilizing these 3D models, employs a three-stage process:

- *Feature identification:* 2D features are extracted from the captured image.
- *Feature matching:* These 2D features are matched to corresponding 3D points within the model.
- *Camera pose estimation:* The camera's location and orientation are determined from the feature matches using the Perspective-n-Point (PnP) [Har+94; Gao+03] problem. The PnP problem typically requires a minimum of three correspondences between 2D image points and 3D model points to accurately estimate the camera's pose.

Recent advancements in PnP and three-Point Perspective (P3P) [HR11; ULH16] algorithms have significantly enhanced the speed and accuracy of camera pose estimation, a crucial aspect of structure-based localization. However, PnP algorithms are susceptible to errors introduced by outliers or incorrect data. To overcome this challenge, they are often employed within robust estimation techniques like the Random Sample Consensus (RANSAC) algorithm [FB81], which effectively filter out unreliable data and ensure that only valid matches contribute to the pose estimation process. In larger-scale environments, the extensive number of 3D points required for accurate localization models can pose significant challenges. The sheer volume of data can lead to increased memory consumption and hinder the performance of matching algorithms, potentially leading to ambiguous or erroneous results. To address these issues, recent research efforts have focused on developing techniques to minimize model size [LSH10; HQM14], accelerate matching processes [Che+19; Li+12], and enhance the reliability of matching algorithms [Lar+16; Alc+11]. These advancements aim to optimize the utilization of computational resources and ensure the robustness of localization systems in large-scale environments.

Retrieval-based localization stands in contrast to structure-based localization, which utilizes 3D models to establish correspondences between image features and specific locations within the environment. Retrieval-based localization employs a different approach, involving three key steps:

- *Feature extraction:* A comprehensive feature representation of the captured image is generated. This representation typically encompasses a variety of local and global features, capturing the distinctive characteristics of the image.
- *Feature matching:* The extracted features are matched to corresponding features within a database of reference images that have previously been geotagged. This matching process is typically performed using techniques such as BoW, which segments local features into discrete visual words and

represents each image with a sparse histogram of these words.

- *Location estimation:* The location of the captured image is determined based on the most similar reference image(s) found in the database. This estimation can be achieved by identifying the best match or by averaging the matches of the top-ranking reference images.

Various global feature representations have been employed in retrieval-based localization for tasks such as loop closure (re-identifying previously visited locations) and place recognition (identifying new locations as part of a known environment). Early methods utilized techniques like Principal Component Analysis (PCA) [Krö+01] and color histograms [HE08; UN00]. These approaches were later supplemented by crafted global descriptors like Gist [OT01; OT06], which demonstrated superior performance in certain scenarios.

In recent years, there has been a growing emphasis on combining local features to create a more comprehensive global image descriptor. Techniques such as WI-SURF [AKB08] and BRIEF-Gist [SP11] utilize local feature extraction from specific points across the image to capture more detailed information. However, the BoW approach [SZ03] has emerged as the most widely adopted method for aggregating local features into a singular global descriptor. BoW divides the array of local features into "visual words" and represents each image with a sparse histogram of these words. To further enhance the efficiency of the retrieval process, BoW is often paired with an inverted index, a technique that accelerates searches by indexing feature descriptors and storing reference image locations alongside the corresponding visual words. This indexing method is employed in FAB-MAP 2.0 [CN11], a prominent retrieval-based localization system.

### 2.2.2. Deep learning approach

Deep learning based approaches to global localization can be categorized into three main types, based on the nature of the inquiry data and the map used: 2D-to-2D, 2D-to-3D, and 3D-to-3D localization [Che+20]. In 2D-to-2D localization, the process involves estimating the camera pose of an image relative to a 2D map. This map can be either an explicit map, constructed using a geo-referenced database of images, or an implicit map, encoded within a neural network. Such methodologies are pivotal for achieving accurate global positioning across a variety of applications, including autonomous navigation and augmented reality.

Furthermore, 2D-to-3D localization establishes correspondences between the 2D pixels of images and the 3D points of a scene model, providing a more detailed understanding of the environment by leveraging the depth information. On the other hand, 3D-to-3D localization involves matching 3D scans with a pre-built 3D map, offering precise alignment and integration within three-dimensional space.

These approaches represent the cutting edge in leveraging deep learning for global localization, highlighting the diverse strategies employed to interpret and navigate known environments. By utilizing explicit databases of geo-referenced

images or implicit neural maps for 2D-to-2D queries, establishing pixel-to-point correspondences for 2D-to-3D localization, or matching 3D scans in 3D-to-3D localization, these methodologies underscore the flexibility and depth of current global localization techniques. Each method offers unique advantages, whether in terms of the simplicity and accessibility of 2D-to-2D localization, the depth and detail afforded by 2D-to-3D methods, or the comprehensive spatial understanding enabled by 3D-to-3D approaches.

**Explicit map-based localization:** Explicit map-based 2D-to-2D localization uses a database of geo-tagged images to represent the environment. The process is twofold: initially, image retrieval identifies the closest match within the database to the query image, and subsequently, pose regression calculates the camera's relative pose to these reference images. One challenge in this method is selecting effective image descriptors for retrieval. Deep learning advancements have led to the utilization of Convolutional Neural Network (CNN) models to extract image-level features, significantly improving image matching accuracy. For example, the introduction of the NetVLAD layer [Ara+16], a trainable component, enhances the feature extraction capability of CNN's, making them more adept at identifying similarities among images for retrieval purposes.

To refine pose estimation, techniques have evolved from relying on epipolar geometry, which depends on matching local descriptors between images, to directly regressing relative poses using deep learning [Zho+20; Mel+19]. Innovations such as NN-Net [Las+17] estimate relative poses between a query image and its nearest references, integrating these estimates with known 3D geometries to deduce the query's absolute pose. Approaches like Relocnet [BLP18] and CamNet [Din+19] further refine this process by optimizing global descriptors and applying a two-stage retrieval method to enhance pose accuracy. These reference-based methods offer scalability and adaptability across various scenarios without the need for retraining, balancing accuracy and scalability effectively.

**Implicit map-based localization:** In contrast, implicit map-based localization bypasses the need for a reference database, directly estimating camera poses from single images through deep neural networks. This approach is embodied by PoseNet [KGC15], which predicts camera poses end-to-end from RGB images. Despite its innovative use of ConvNet architectures, PoseNet's initial models struggled with overfitting and required extensive manual tuning. Subsequent enhancements have addressed these issues by incorporating LSTM units [Wal+17], synthetic data augmentation [WMH17; NB17; PZZ18], advanced neural architectures [Mel+17; CSR18], and geometry-aware loss functions [KC17].

Further developments include Atloc [Wan+19], which employs an attention mechanism to focus on relevant image features, and RVL [Hua+19], which uses a prior-guided dropout to mitigate uncertainties caused by dynamic elements in the scene. VidLoc [Cla+17] introduces temporal constraints to account for the sequence of images, enhancing the modeling of visual localization. Adding more external motion constraints can increase the consistency in predicting posture.

Implicit map-based localization leverages deep learning's feature extraction strengths, particularly in environments lacking distinct features. However, its dependency on scene-specific training limits its applicability to new, unseen environments without additional training. Despite these challenges, implicit methods continue to evolve, integrating semantic learning and pose regression to improve localization accuracy.

## 2.3. Conclusion

This chapter has presented a comprehensive review of the state-of-the-art techniques in map representations and image-based localization, with a specific focus on their applicability to complex, dynamic urban environments. Through the exploration of geometric, topological, feature-based, and implicit representations, we have highlighted the strengths and limitations of various mapping strategies, ranging from precise 3D metric maps to lightweight, memory-efficient neural representations. Similarly, the overview of classical and deep learning-based localization methods has revealed a shift from structure-based models and image retrieval systems toward scalable, data-driven approaches capable of handling diverse environmental conditions. Despite recent advances, existing methods often struggle with balancing accuracy, computational efficiency, and generalizability across varying urban morphologies. These insights underscore the need for multimodal, context-aware solutions that leverage complementary data sources and learning paradigms. The methodologies proposed in the following chapters are informed by these gaps, aiming to build upon the strengths of existing approaches while addressing their limitations through the integration of generative AI and pedestrian-centric sensing.

# 3. DESIGN AND DEVELOPMENT OF THE MULTI-SENSORY DATA ACQUISITION PLATFORM (DELTA PLATFORM)

The Digital Enhancement for Localization in Tartu Area (DELTA) platform was developed to address the challenges of acquiring detailed, multimodal data necessary for analyzing urban environments, particularly pedestrian pathways and micromobility infrastructure. Existing mapping methods often rely on limited data modalities, making them insufficient for capturing the spatial and contextual complexity required for precise localization, infrastructure planning, and urban navigation in dynamic settings. DELTA was designed to bridge this gap by integrating diverse sensors to provide a more comprehensive and nuanced understanding of urban landscapes.

A key shortcoming of conventional mapping techniques is their inability to effectively capture the dynamic and multi-faceted nature of urban environments. Urban spaces are not static but inherently dynamic, encompassing not just the physical layout but also elements like pedestrian flow, vehicular traffic, and a range of environmental factors [Mir+16]. Traditional mapping approaches might provide a basic structural layout but fall short in terms of spatial and temporal resolution, crucial for reflecting the ever-evolving nature of cities [Zha+18]. Changes in urban infrastructure, seasonal variations in vegetation, and fluctuating patterns of pedestrian and vehicular movement are just some examples of the dynamic aspects that conventional methods struggle to capture. As a result, these methods often yield a static and outdated view of the urban landscape, lacking the necessary detail and responsiveness to keep pace with rapid urban changes.

When designing the DELTA platform, the main goal was to build a system that could collect high-resolution, diverse data for urban planning, mobile robotics, and autonomous driving research. The focus was on making it modular, compact, and easy to replicate so it could be adapted and used in different settings. The following sections will provide a detailed overview of the DELTA platform. They describe the integrated hardware—comprising the data collection rig, sensor interfaces, power management, and vibration control—as well as the software tools used to configure and optimize sensor performance. Together, these sections outline the platform's design and its capabilities for multisensory urban data collection.

## 3.1. Overview of the platform

The following sections introduce the DELTA platform (Figure 1), the platform incorporates an array of sensors and recording devices that allow for accurate and reliable data collection. A central aspect of the design is the implementation of a vibration-damping system to minimize external noise and vibrations, ensuring

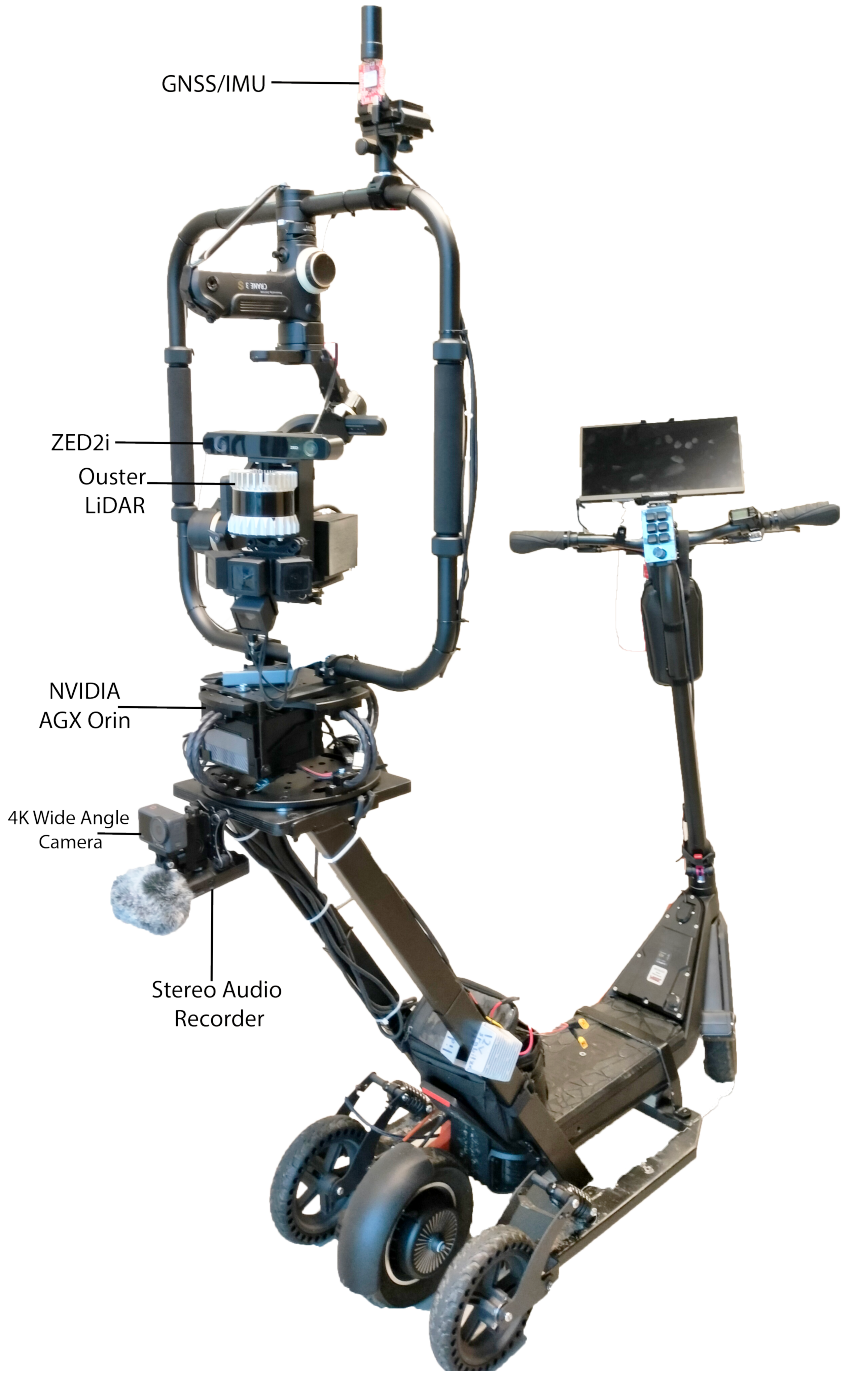



**Figure 1.** DELTA platform: a customized e-scooter multi-sensing platform.

the integrity of the collected data. performance, various isolation strategies were carefully integrated into the design. The scooter features a modular architecture, enabling straightforward customization to adapt to various research needs. Its in-

terface was designed to support real-time monitoring and control of the sensors, providing ease of use for operators regardless of their technical expertise. This ensures that the platform is both practical and efficient for collecting data in diverse urban scenarios.

## 3.2. DELTA platform hardware setup

The following sections detail the design and functionality of the DELTA platform, covering key components essential for urban data collection. It begins with the data collection setup, including the choice of vehicle, custom rig, and onboard computing system. Next, it explores sensor interfaces, describing how the stereo camera, LiDAR, GNSS, and IMU integrate with the main board. Power management strategies are then outlined, ensuring stable and efficient operation of all components. Finally, the section discusses the damping systems used to minimize vibrations and enhance data quality, ensuring reliable performance in diverse urban environments.

### 3.2.1. Design and setup of the e-scooter–based platform

A high-performance electric scooter was chosen for the data collection platform because it could easily carry the equipment while providing enough space for both the instruments and the operator. The Inokim Ox Hero was selected for this job, featuring a 1000W motor and a 13Ah battery. This setup allows it to handle rough terrain and steep inclines smoothly. The scooter can reach speeds of up to 45km/h, making it quick and efficient for urban data collection. On a single charge, it has a range of up to 60km (without the equipment), making it a reliable and durable option for long data collection sessions.

To facilitate the crucial task of data collection, the establishment of a robust platform is essential. For this purpose, a custom-built metal rig was meticulously engineered and constructed to accommodate the scooter (refer to Figure 2), ensuring seamless integration of sensors, the damping system, and the compute board. This rig features a duo of rear shock absorbers equipped with anti-puncture wheels, securely attached to the scooter's body using nuts and bolts. This configuration not only enhances the platform's stability but also ensures the safe housing of equipment during operation. Furthermore, a pair of angled beams is strategically welded to the rig's base, creating an optimized space for the placement of sensors, damping devices, and the compute board. This design consideration ensures that all components are securely mounted and positioned for optimal functionality, facilitating efficient data gathering in various urban environments.

To effectively manage and process the substantial volume of data generated by its sensors, the DELTA platform incorporates a Samsung 980 PRO 2 TB PCIe 4.0 NVMe M.2 Internal SSD. This choice is predicated on the SSD's speed, boasting read/write speeds of up to 7000/5000 MB/s, which is essential for the swift recording and retrieval of sensor data. By mitigating potential delays or bottlenecks. The


**Figure 2.** The platform equipment rig, where sensors, damping system, and the compute board are mounted.

Tascam audio recorder and the Osmo Action 3 camera independently store their data within their respective storage units, with 32GB and 128GB respectively.

The system uses the state-of-the-art NVIDIA Jetson AGX ORIN development board to handle the integration and analysis of inputs from its stereo camera, LiDAR, GNSS, and IMU modules. The Jetson AGX Orin board, is equipped with an Arm Cortex-A78AE CPU and an NVIDIA Ampere architecture GPU, along with the newest deep learning and image processing accelerators and improved video encode and decode features. This development board is a complete system-on-module, incorporating CPU, GPU, PMIC, DRAM, and flash storage, thereby streamlining development processes and reducing both time and cost, it also includes reference carrier board, power supply, and all necessary components.

Data acquisition and processing tasks are managed directly from the e-scooter, with all data viewable in real-time on a 14-inch 1080p FHD display (Figure 3(a)). The display is mounted on the front of the e-scooter for optimal viewing, and is powered by an external 20000mAh power bank, and connected to the main board via HDMI cable, allowing the operator to monitor the operation and outcomes of data collection as it happens. Additionally, a Bluetooth-enabled macro keyboard (Figure 3(b)) is mounted in front of the screen to allow the user to communicate

with the scooter's sensors. The user may enable and/or disable certain sensor functions while the sensor is collecting data using the dial and six programmable buttons on this module. It is powered by an integrated rechargeable battery, which greatly improves convenience of usage.


**Figure 3.** The platforms interface, (a) FHD display, (b) macro keyboard.

### 3.2.2. Sensors interfaces

Each sensor inside the system interfaces with the main board using various connection techniques, adapted to their operating requirements and data transfer demands.

The ZED2i stereo camera establishes its connection via a USB 3.0 port, ensuring high-speed data transfer capabilities essential for processing the rich visual information it captures. The Ouster LiDAR system consists of two components: the primary LiDAR unit and an interface box. The LiDAR unit connects to the interface box, which not only supplies the 24V DC power supply required for the LiDAR's functioning, but also serves as a data transmission channel. The interface box has a breakout point for the LiDAR's Ethernet connection, as well as an RJ45 connector, which makes it easier to connect the LiDAR to the main board via Ethernet cable. The GNSS module utilizes a USB-C connection to interface with the main board. The IMU module's interfacing is slightly more complex, requiring a serial converter module to facilitate communication with the main board. The IMU connects to this converter, which then links to the main board using a USB port. The necessity for the serial converter arises from the IMU's specific communication protocol, which is not natively supported by the main board's direct connections.

### 3.2.3. Power management

The power management architecture of the platform includes a rechargeable Ecoworthy 30Ah Lithium Iron Phosphate (LiFePO4) battery, which is to support the Jetson AGX Orin development board and the large sensor array. Because of its

30Ah rating, this battery is characterized by a high energy storage capacity, which means that it can sustain a certain current output for an extended period of time. This explains the LiFePO4 battery's flexibility in providing continuous power according to demand. It is theoretically possible to provide a constant current of 1 ampere for 30 hours, or 2 amperes for half that time. With this feature, the battery's operational usefulness can be assessed in a variety of settings, providing a reliable baseline for both its lifetime and energy delivery efficiency. Crucially, this battery contains a BMS, offering vital protection against potential damage due to over or under-voltage, over-current, excessive temperatures, or short circuits. By protecting the cell integrity, this BMS makes sure that the platform's power supply is safe and operational in a variety of scenarios. By reducing battery waste and prolonging the life of the power source, such a solution not only improves the energy supply's dependability and safety but also supports sustainability and environmental goals.

To keep the Jetson AGX ORIN board running smoothly and protect it from power surges, the system includes a 12V DC regulator. This is important because a fully charged LiFePO4 battery outputs 14.3V, which is higher than what many components are designed to handle. The regulator adjusts this to a steady 12V, ensuring a safe and reliable power supply for the Jetson board and preventing any damage from excessive voltage.

Additionally, the platform utilizes a DC-DC step-up converter to cater to the Ouster LiDAR's specific requirement of 24V. This converter adeptly raises the regulated voltage to meet the LiDAR's operational needs without compromising the integrity or performance of the system. By managing the voltage supplied to each component, the system ensures that the entire data collection apparatus, including the high-demand LiDAR sensor, operates efficiently and effectively. This comprehensive approach to power management is pivotal in maintaining the optimal performance of all onboard technological components.

As parts of the data collecting system, the Osmo Action 3 camera and the Tascam DR-07X audio recorder have separate battery systems that allow them to operate independently.

### 3.2.4. Damping systems

In order to reduce the effects of vibrations on cameras and sensitive devices, especially those connected by wires or cables, the integration of a 3-axis camera gimbal stabilizer (Figure 4 (a)), an anti-vibration plate (Figure 4 (b)), and particularly the vibration isolator wire mount (Figure 4 (c)) is essential. These systems use a dual strategy: they either physically separate the camera from the vibration source or directly reduce the vibrational energy. The use of flexible metal wires in a parallel arrangement is crucial in this setup. They absorb and disperse vibrational forces, preventing damage to the attached equipments.

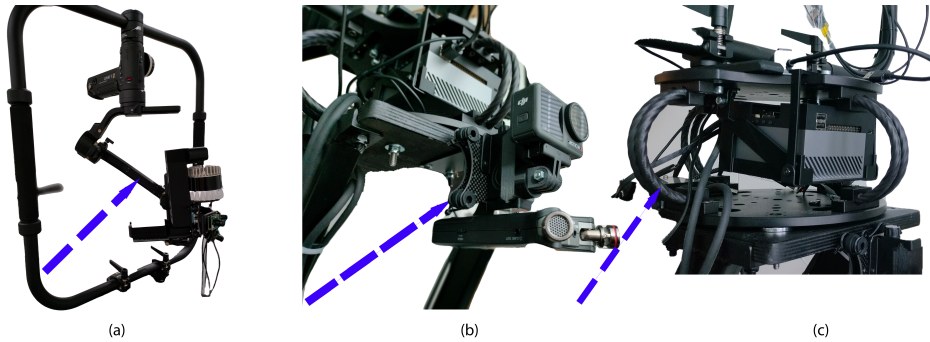



**Figure 4.** Damping systems: (a) 3-axis camera gimbal, (b) anti-vibration plate, (c) vibration isolator wire mount.

To maintain the stability of the stereo camera and LiDAR in dynamic environments, a 3-axis camera gimbal is used. This stabilizer actively counteracts external disturbances using a combination of brushless motors and sensors, which work together to reduce unwanted vibrations and rotational shifts. The system operates across three key axes: pitch (up/down), roll (forward/backward), and yaw (sideways). Each axis is controlled by a dedicated motor, guided by accelerometers and gyroscopes that continuously detect movement changes. When a disturbance is detected, the gimbal's control system processes the sensor data and sends real-time corrective signals to the motors. These motors then adjust their positions almost instantaneously, minimizing any tilting or shaking to keep the camera and LiDAR level and steady. To further reduce vibrations affecting the action camera and audio recorder, an anti-vibration plate equipped with rubber dampening balls is utilized. These rubber elements act as shock absorbers, isolating the devices from external movement and ensuring smooth video recording and clear audio signals. Finally, both the stereo camera and LiDAR are securely mounted on the gimbal stabilizer, which is fixed to a rectangular frame and further supported by a vibration isolator wire mount. This multi-layered stabilization system effectively minimizes any shakiness and sudden motions caused by the moving platform.

## 3.3. Software tools for sensors

In this section, I detail the software tools essential for configuring and optimizing sensor performance and system functionality. These tools enable precise calibration, customization, and correctt operation of the platform's sensors and peripherals. Except for the ZED SDK—which was installed on the Jetson board—all tools were run on a Windows operating system.

### 3.3.1. U-center: Configuring the ZED-F9P GNSS

U-center[1] created and developed by u-blox, is a complete software tool designed for the detailed examination, managing, and customization of u-blox GNSS modules. This software is critical for evaluating, configuring, and testing functions like the AssistNow A-GPS service, as well as conducting in-depth analysis of GNSS data streams. The program runs on a Windows platform and provides an interactive environment for communicating with u-blox devices via the UBX binary protocol or the NMEA-0183 standard. Its architecture offers real-time viewing of essential GNSS metrics such as location, velocity, and satellite trajectory. U-center comes equipped with a number of configuration utilities that enable for fine modifications to positioning modes, filters, and computational methods. Furthermore, it facilitates firmware administration tasks by providing means for downloading and uploading firmware revisions to u-blox modules. Notably, the program includes features such as antenna calibration, raw data stream access, and complete performance analysis tools, as well as logging and replay capabilities for thorough GNSS data study. The U-center was employed to configure the u-blox module, tailoring it to enable specific messages for output and disabling those that were unnecessary, thereby saving both time and memory. Configuring the module through the software was also crucial to facilitate the reception of raw GNSS messages, enhancing the efficiency and precision of data collection.

### 3.3.2. ZED SDK: Spatial perception framework

The ZED SDK[2] by Stereolabs provides tools for capturing and analyzing three-dimensional data from ZED stereo cameras. It enables real-time acquisition of depth, color, and inertial measurements, generating accurate depth maps and 3D environment representations. The SDK includes features for object detection, tracking, and motion analysis, along with visualization tools such as point clouds and depth fields. Its user-friendly API and real-time processing capabilities make it a practical choice for developing advanced computer vision applications.

### 3.3.3. WitMotion: Configuring the IMU

WitMotion[3] software supports efficient data collection and streaming, including real-time and delayed modes. It provides tools for visualizing sensor data in formats like charts, graphs, and 3D models, aiding analysis. The software includes calibration features to ensure sensor accuracy and reliability and allows customization of IMU settings to meet specific project requirements. Additionally, it enables data logging and export for further analysis and compatibility with other platforms or tools. The WitMotion software was used for calibrating the

[1] Available at: `https://www.u-blox.com/en/product/u-center`

[2] Available at: `https://www.stereolabs.com/en-ee/developers/release`

[3] Available at: `https://www.wit-motion.com/searchq.html`

IMU, targeting the gyroscope, accelerometer, and magnetometer for precise adjustments. This calibration process was complemented by fine-tuning various parameters, such as the baud rate and sampling rate, to enhance the sensor's performance in alignment with the specific needs of the project.

### 3.3.4. miniKeyBoard: Configuring the macro keypad

The mini KeyBoard software enables extensive customization of a programmable macro keyboard. Users can assign specific actions, such as launching applications or running scripts, to each of the 16 keys through a user-friendly interface. The software also supports multiple key assignment layers, allowing for flexible configurations tailored to different tasks or workflows. The miniKeyboard app was used to control the activation and deactivation of various sensors, including the stereo camera, LiDAR, GNSS, and the IMU. This setup enabled me to selectively turn these devices on or off, providing control over each sensor's operation as needed.

## 3.4. Conclusion

This chapter detailed the design and development of the DELTA platform, a custom-built, multimodal data acquisition system tailored for pedestrian-centric urban sensing. By integrating a diverse suite of sensors—including LiDAR, stereo and monocular cameras, GNSS, IMU, and a high-fidelity audio recorder—into a mobile e-scooter platform, the system enables rich and synchronized environmental data collection in real-world conditions. Special emphasis was placed on mechanical stability, sensor placement, and synchronization techniques to ensure accurate spatial and temporal alignment across modalities. The DELTA platform not only facilitates the capture of complex multimodal datasets but also serves as a research-grade testbed for experimentation with sensor fusion, segmentation, and localization techniques. Its modular and portable design supports adaptability across various use cases, including sidewalk analysis, soundscape classification, and urban mapping. The foundation laid in this chapter enables the data-driven methodologies explored in subsequent chapters and underscores the importance of tailored hardware solutions in advancing urban perception systems.

# 4. HANDLING AND PROCESSING RAW DATA FROM THE DELTA PLATFORM

The following sections detail the comprehensive approach taken to develop the DELTA dataset, which focuses on pedestrian-centric urban data collection. First, the need for such a dataset is discussed, highlighting the limitations of traditional vehicle-focused datasets and the importance of capturing detailed information about pedestrian infrastructures. Next, the concept of multimodal data collection is introduced, emphasizing the integration of various sensor types—such as GNSS, IMU, stereo cameras, LiDAR, audio, and high-resolution video—to capture a complete picture of urban environments. Subsequent sections explain how these diverse sensor inputs are synchronized and fused, ensuring that data from different sources are accurately aligned both in time and space. This is followed by an in-depth look at the calibration techniques, 3D-to-2D projection methods, and sensor fusion strategies employed to enhance the dataset's precision and reliability. Finally, the chapter outlines the data annotation and segmentation processes, which are critical for preparing the dataset for advanced machine learning and autonomous system applications.

## 4.1. Need for pedestrian-centric data

The evolution of urban environments and the integration of diverse mobility solutions, including micro-mobility devices and autonomous vehicles, have significantly transformed pedestrian pathways, sidewalks, and public spaces into dynamic components of urban infrastructure. This transformation underscores the need for a deeper understanding of pedestrian infrastructure's role in urban mobility and the advancement of autonomous and mobile robotic systems [FBC23; ALD21]. The shift towards dynamic pedestrian environments, accommodating electric scooters, e-bikes, and autonomous delivery robots alongside traditional foot traffic, raises crucial questions about safety, efficiency, and accessibility in these shared spaces.

Despite the proliferation of datasets aimed at supporting urban research, a discernible gap exists in data specifically pertaining to pedestrian-centric aspects of cities. Traditional datasets, often focused on vehicular perspectives and road networks, such as the KITTI vision benchmark suite [Gei+13], Cityscapes [Cor+15], and Waymo [Sun+20], which have primarily been developed with a focus on vehicular perspectives and roadways, emphasizing autonomous navigation and traffic management. However, these datasets tend to overlook the nuanced challenges specific to pedestrian environments. While UrbanLoco [Wen+20] and Hong Kong UrbanNav [Hsu+23] incorporate raw GNSS measurements for navigation research, they, too, focus predominantly on vehicle-centric urban areas. This oversight highlights the critical need for datasets that offer detailed insights

into pedestrian areas, which are indispensable for developing autonomous systems that are safe and efficient in both vehicular and pedestrian contexts.

A few existing datasets do address pedestrian dynamics, but often in constrained scenarios such as public space crowd behavior or indoor tracking. These specialized datasets, while useful, do not fully represent the complexity of pedestrian activity in urban settings. The need for quality data that captures the intricacies of pedestrian infrastructure is paramount, especially as advancements in deep learning continue to push the boundaries of computer vision techniques like semantic segmentation.

Recent research initiatives have pioneered a variety of innovative techniques for gathering data on pedestrian infrastructure, reflecting a dynamic shift in how urban environments are studied. For instance, Tile2Net [Hos+23] leverages semantic segmentation of aerial images to chart pedestrian pathways, offering a bird's-eye view of urban walkability. SideGuide [Par+20] takes a more nuanced approach by integrating insights from interviews with individuals who have mobility impairments, crafting a dataset that zeroes in on features critical to sidewalk accessibility. Another study [Wel+19] taps into Google Street View panoramas to evaluate the accessibility of sidewalks, showcasing the utility of leveraging readily available online images for urban studies.

The PESID project [Sun+19] contributes to this growing body of research by assembling a dataset from labeled photographs that illustrate a variety of sidewalk scenarios, emphasizing pedestrian safety. Similarly, a study [Nin+22] merges aerial and street-level imagery to construct a detailed representation of sidewalk networks, offering a holistic view of pedestrian infrastructure. CitySurfaces [Hos+22] employs computer vision technology to categorize sidewalk materials through analysis of street-level images, enriching datasets focused on material types.

In a creative twist, another research effort [Dec+22] utilizes the CARLA simulator's 3D virtual environment to simulate scenarios for electric wheelchairs on sidewalks, illustrating the innovative application of synthetic scenes in data collection for urban planning. These diverse approaches not only highlight the evolving landscape of data collection techniques but also underscore the growing recognition of the critical need for comprehensive datasets. Such datasets are instrumental in advancing our understanding and capabilities in urban planning, particularly in enhancing the development and safety of pedestrian infrastructure.

The DELTA dataset aims to compensate the crucial gap in urban mapping and analysis. By integrating high-fidelity audio-visual data, precise localization, and robust semantic annotations, it diverges from the conventional road-centric focus of most urban datasets. Instead, it aims to offer a comprehensive sensory and localization data collection, emphasizing pedestrian areas.

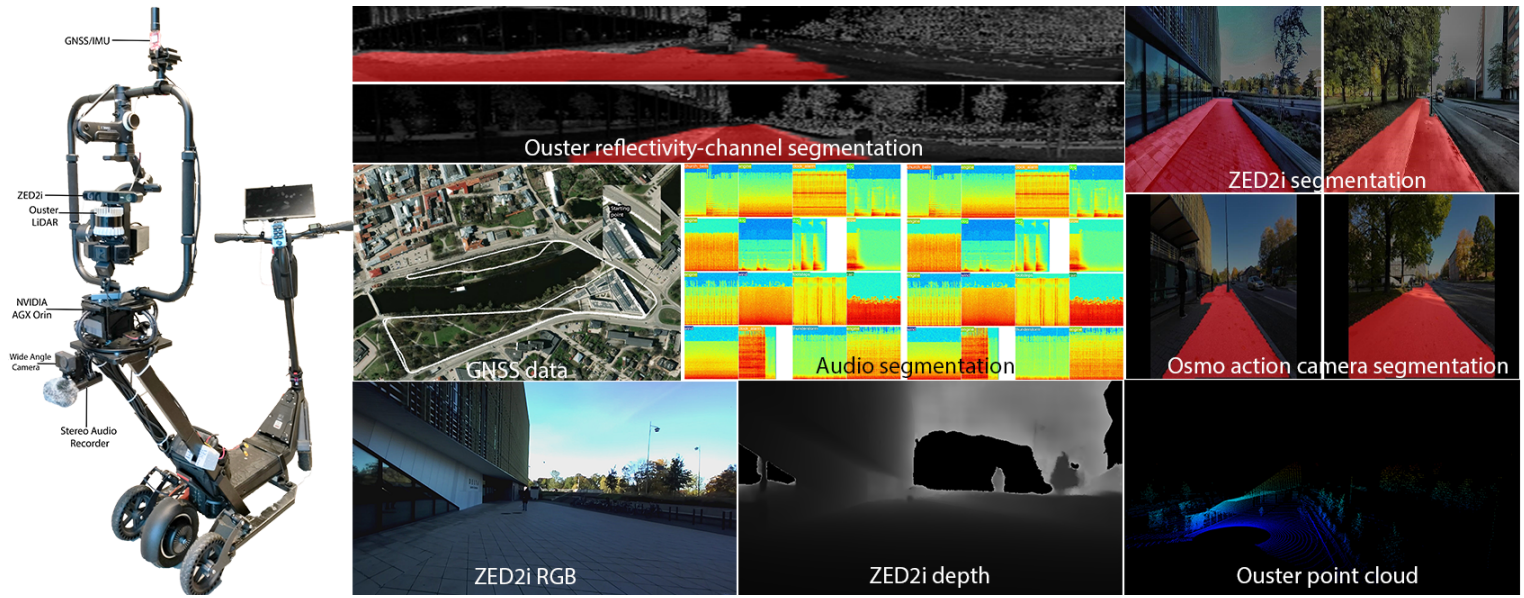



**Figure 5.** An overview of the platform, highlighting the multimodal dataset, sidewalk segmentation results, sound event classifications, and the study area in Tartu, Estonia.

## 4.2. Need for multimodal data collection

The necessity for multimodal data collection in urban mapping and analysis is rooted in the intricate and dynamic nature of urban environments. This scientific approach is vital for gaining a comprehensive understanding of these complex ecosystems. Urban areas are not just physical spaces; they encompass a blend of pedestrian traffic, vehicular movement, architectural features, and various environmental conditions [Jon14]. By employing multimodal data collection, which involves gathering information through a range of sensory inputs such as visual, auditory, spatial, and navigational data, a more holistic picture of urban landscapes emerges. Moreover, multimodal data collection enhances the accuracy and depth of urban analysis. Different sensors, each capturing unique aspects of the environment, contribute to a richer, more nuanced dataset. Visual sensors, like cameras, provide imagery and detail, LiDAR sensors offer precise spatial measurements, and audio sensors capture the ambient sounds of an area. The integration of these diverse data types are essential for training algorithms to safely and effectively navigate urban settings, enhancing their capability to understand and respond to the complexities of urban environments. As cities continue to evolve, facing new challenges such as the rise of micro-mobility solutions and autonomous delivery systems [ALD21; BAS19; ALD21], the adaptability offered by multimodal data collection becomes increasingly important. It ensures that the data reflects the current state of urban environments, facilitating the development of technologies and solutions that are relevant and effective in addressing contemporary urban challenges. Figure 5 presents an overview of the platform, showcasing the multisensor datasets, segmentation techniques, sound event classifications, and the area under study.

## 4.3. Multimodal integration in the DELTA dataset

The following sections describe how the DELTA dataset integrates multiple sensor modalities to capture a comprehensive view of urban pedestrian environments. Each section details a specific sensor—GNSS for precise positioning, IMU for motion and orientation, stereocameras for depth and spatial context, LiDAR for accurate 3D mapping, audio recorders for capturing urban soundscapes, and a 4K camera for high-resolution visual data—and explains its role, functionality, and how its data is synchronized and fused within the overall multimodal framework.

### 4.3.1. GNSS

GNSS represent a network of orbiting satellites that provide geo-spatial positioning with global coverage. This technology allows GNSS receivers to determine their location (longitude, latitude, and altitude) to high precision using the signals transmitted from these satellites. GNSS is widely used in various applications including navigation, surveying, geophysics, and more. It is the backbone of

**Table 1.** ZED-F9P GNSS module features

| Feature | Description |
|---|---|
| L1 and L2 GNSS Reception | Receives signals from L1 and L2 bands for RTK applications. |
| Communication Ports | Offers USB, I2C, UART/Serial, and SPI ports for data transfer and configuration. |
| USB Port | USB-C connector for configuration and NMEA sentence viewing. |
| I2C (DDC) Port | I2C port for reading NMEA sentences and handling RTCM data. |
| UART/Serial | UART pins for serial communication, UART2 for RTCM3 correction data. |
| SPI Port | SPI communication option, can disable UART1 and I2C. |
| Control Pins | Includes control pins for geofencing, RTK mode indication, etc. |
| Antenna | Compatible with U.FL connectors, supports multi-Band L1/L2 Helical Antenna. |

modern location-based services and technologies, enabling accurate and reliable geographic positioning information.

The GNSS data encompasses raw measurements from multiple satellite constellations, including GPS, GLONASS, Galileo, and BeiDou. This data includes pseudorange, which provides an initial estimate of the satellite-receiver distance, and carrier phase data, offering more refined measurements essential for applications needing heightened accuracy, such as differential GNSS. Doppler shift data is also included, indicating the rate of change in satellite-receiver distance, useful for determining the receiver's velocity.

Each satellite in the dataset is uniquely identified, and the data contains signal quality metrics like the carrier-to-noise ratio and lock time, which are indicative of signal strength and duration of continuous tracking, respectively. The UBX-NAV-PVT message within the dataset provides a detailed navigation solution, including position, velocity vectors, and temporal information, crucial for real-time navigation and geospatial data collection. Additionally, the NMEA-GNVTG message offers information on ground-relative course and speed, augmenting the dataset's applicability in pedestrian dynamics studies and urban planning. The ZED-F9P GNSS receiver board (Figure 6(a)) is a high-performance module tailored for Real-Time Kinematic (RTK) applications, offering advanced precision in positioning tasks. It stands out with dual-band GNSS reception, multiple commu-

nication ports including USB, I2C, UART/Serial, and SPI, and additional control pins for expanded functionality. Its capability to operate across L1 and L2 bands significantly enhances positioning accuracy.

The modules key features include a USB C port for straightforward connectivity to platforms like u-center, facilitating easy configuration and monitoring, alongside compatibility with single-board computers for versatile applications. The I2C port supports a wide range of operations from reading NMEA sentences to handling RTCM data, complemented by Qwiic connectors for easy integration with I2C devices. UART pins provide serial communication, particularly useful for RTCM3 correction data, while SPI communication offers additional data transfer options, albeit with specific configuration requirements.

The module also boasts control pins for geofencing, RTK mode indication, pulse-per-second outputs, and essential reset and safeboot functions, enhancing its control and monitoring capabilities. Antenna connectivity is ensured through U.FL connectors and an option for an SMA bulkhead, with a preference for a multi-band L1/L2 Helical Antenna to optimize signal reception. A detail description of the module specs are presented in Table 1.

The DELTA dataset incorporates a ZED-F9P GNSS module capable of providing raw GNSS data with a horizontal accuracy of approximately 1.1 meters (Figure 6(b)) without RTK corrections. While this accuracy is sufficient for many applications, the dataset also includes raw GNSS measurements, enabling researchers to apply Post-Processed Kinematics (PPK) for enhanced precision. PPK refines localization data to achieve centimeter-level accuracy by correcting GNSS signals during post-processing. Unlike RTK, PPK offers greater flexibility as it does not require immediate proximity to a base station during data collection.

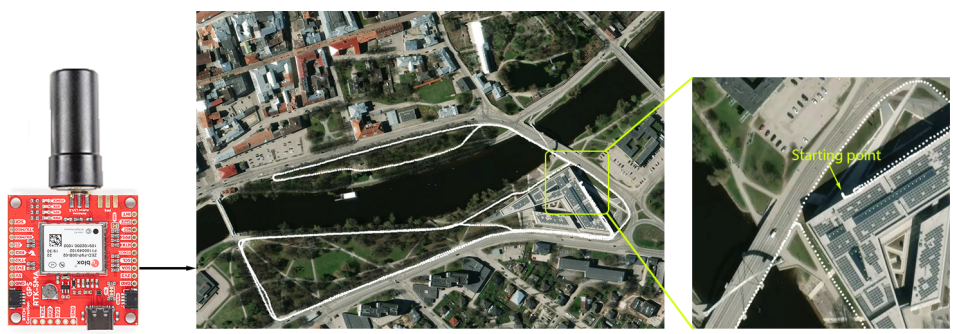



**Figure 6.** A visual representation of (a) GNSS module used and (b) the georegistered trajectory points overlaid on OpenStreetMap.

### 4.3.2. IMU

An Inertial Measurement Unit (IMU) is a key sensor in various navigation and tracking systems. It typically consists of accelerometers, gyroscopes, and magnetometers, which measure linear acceleration, angular rate, and magnetic field

orientation, respectively. These sensors allow the IMU to provide comprehensive motion data.

**Table 2.** HWT901B-RS485 MPU9250 9-axis IMU specifications

| **Parameter** | **Specification** |
|---|---|
| Working Voltage | TTL: 5V-36V |
| Current | $<$40mA |
| Size | 55mm x 36.8mm x 24mm |
| Data | Angle: X Y Z, 3-axis<br>Acceleration: X Y Z, 3-axis<br>Angular Velocity: X Y Z, 3-axis<br>Magnetic Field: X Y Z, 3-axis<br>Air Pressure: 1-Axis<br>Time, Quaternion |
| Output frequency | 0.2Hz–200Hz |
| Interface | Serial TTL level |
| Baud rate | 9600 (default, optional) |

**Table 3.** Measurement range and accuracy of the IMU

| **Sensor** | **Measurement Range** | **Accuracy/ Remark** |
|---|---|---|
| Accelerometer | X, Y, Z, 3-axis<br>$\pm 16g$ | Accuracy: 0.01g<br>Resolution: 16bit<br>Stability: 0.005g |
| Gyroscope | X, Y, Z, 3-axis<br>$\pm 20000^\circ/s$ | Resolution: 16bit<br>Stability: 0.05°/s |
| Magnetometer | X, Y, Z, 3-axis<br>$\pm 4900\mu T$ | $0.15\mu T$/LSB typ.<br>(16-bit) PNI RM3100 Magnetometer Chip |
| Angle/ Inclinometer | X, Y, Z, 3-axis<br>X, Z-axis: $\pm 180^\circ$<br>Y $\pm 90^\circ$<br>(Y-axis 90° is singular point) | Accuracy: X, Y-axis: 0.05°<br>Z-axis: 1° (after magnetic calibration) |
| Barometer | 1-axis | Accuracy: 1m |

The HWT901B-RS485 MPU9250 9-axis IMU (Figure 7(a)) is an advanced sensor module designed for precise measurement of various motion and environmental parameters, encapsulating a wide array of functionalities within a compact

form factor. This sensor module integrates a 9-axis system comprising a 3-axis accelerometer, a 3-axis gyroscope, and a 3-axis magnetometer, alongside an air pressure sensor. The accelerometer is responsible for detecting linear accelerations, such as g-forces, which are pivotal in understanding motion dynamics. The gyroscope measures angular velocity, essential for tracking rotation rates and orientations. The magnetometer, on the other hand, assesses magnetic field strength, enabling compass functionalities crucial for navigation applications. Additionally, the inclusion of a barometric pressure sensor allows for altitude estimation, expanding the sensor's utility in vertical positioning and environmental monitoring. One of the standout features of the HWT901B-RS485 MPU9250 is its high data output rate, capable of reaching up to 200Hz. This capability ensures the sensor can capture dynamic movements and rapid changes with high fidelity, making it suitable for applications requiring real-time monitoring and fast response times. A detailed description of the module is offered in Tables 2 and 3.

The DELTA dataset captures tri-axial acceleration, gyroscope readings, angular orientation, and magnetometer data, that are all precisely timestamped (Figure 7(b)). When synchronized with GNSS coordinates, this IMU data becomes a powerful tool for enhancing localization accuracy.

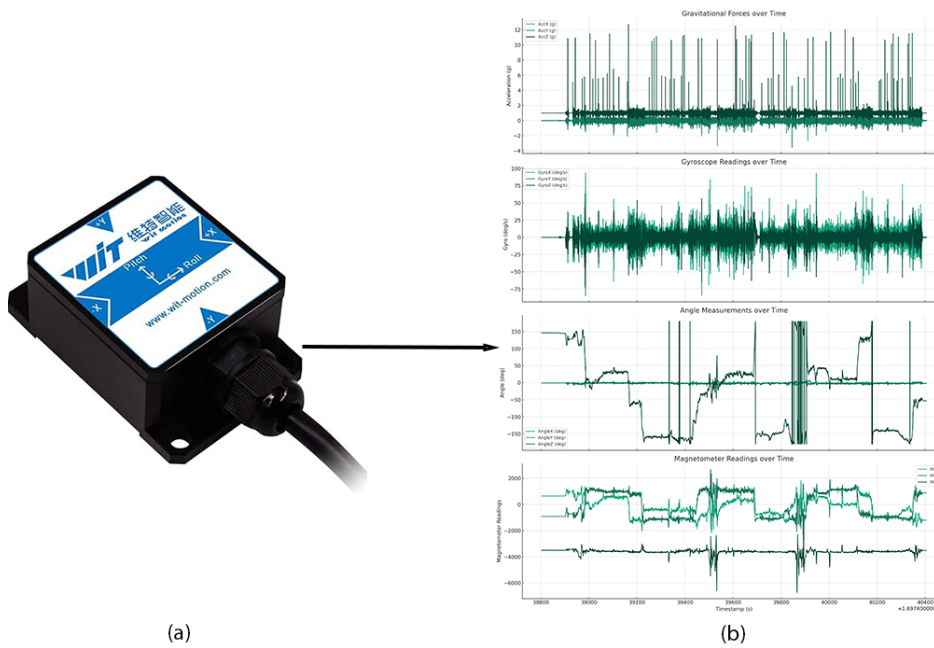



**Figure 7.** Example of the IMU module and the different types of data collected from the sensor.

### 4.3.3. Stereocamera

The ZED2i stereocamera (Figure 8(a)) offers versatile resolution options, including 2K at 15 fps, HD at 30 fps, and 720p at 60 fps, with a maximum of 100 fps for

**Table 4.** ZED2i stereocamera specifications details.

| **Camera Specifications** | |
|---|---|
| **Video Output** | 2.2K mode, 15 fps; resolution 4416 x 1242<br>1080p mode, 30/15 fps; resolution 3840 x 1080<br>720p mode, 60/30/15 fps; resolution 2560 x 720 (stereo passthrough mode)<br>WVGA mode, 100/60/30/15 fps; resolution 1344 x 376<br>Video recording, Native resolution video encoding in H.264, H.265, lossless format<br>Video streaming, IP (with ZED SDK)<br>ISP, New ISP tuned with machine learning for AI and vision tasks |
| **Depth** | Resolution, Native video (in ultra mode)<br>FPS, Up to 100Hz<br>Depth range, 20 cm to 20 m<br>Field of view, 110° horizontal, 70° vertical, 120° diagonal max.<br>Technology, Neural stereo depth sensing |
| **Motion** | Motion sensors, Accelerometer, gyroscope (data rate: 400Hz)<br>Pose update rate, Up to 100Hz<br>Position sensors, Barometer, magnetometer (data rate: 25/50Hz)<br>Technology, 6-DoF visual-inertial stereo SLAM with advanced sensor fusion and thermal compensation<br>Pose drift, 0.35% translation, 0.005°/m rotation (without loop correction) |
| **Lens** | Lens type, Wide-angle 8-element all-glass dual lens with optically corrected distortion<br>Aperture, f/1.8 |
| **Image Sensors** | Dual 4M pixel sensors with 2-micron pixels<br>Sensor format, Native 16:9 for a larger horizontal field of view<br>Sensor size, 1/3” BSI sensor with high low-light sensitivity<br>Shutter, Electronically synchronised rolling shutter<br>Camera controls, Adjust resolution, frame rate, brightness, contrast, saturation, gamma, sharpness, exposure, white balance |

fast-moving subjects. It features a 110° horizontal and 70° vertical field of view, and a depth range of 0.3 to 20 meters. Depth accuracy is within 1% up to 3 meters and 5% up to 15 meters. These capabilities enable the DELTA dataset to capture precise spatial data for analyzing complex urban environments. The DELTA dataset includes 91,411 timestamped stereo image sequences, enabling detailed temporal analysis. Using the ZED2i camera API, depth maps and point clouds were generated from the stereo image pairs, complementing RGB images (Figure 8). Depth maps provide object distance and shape information, while point clouds offer 3D representations of the environment, supporting comprehensive analysis of urban spaces. Table 4 summarizes the module’s specifications.

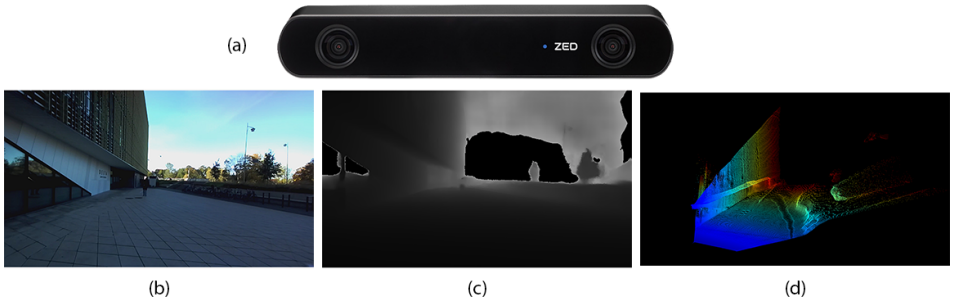



**Figure 8.** Example of (a) the ZED2i stereocamera (b) right-side RGB image of the stereocamera, (b) disparity map, (c) and the point cloud generated from the depth camera.

### 4.3.4. LiDAR

Lidar stands for Light Detection and Ranging. It is a remote sensing technology that uses laser light to measure distances in an environment. A typical lidar sensor emits pulsed light waves from a laser into the environment. These pulses bounce off surrounding objects and return to the sensor. The sensor uses the time it took for each pulse to return to the sensor to calculate the distance it traveled. This continuous cycle of pulse emission and detection creates a high-resolution 3D map—or point cloud—of the surroundings, offering distance accuracy that often exceeds that of traditional cameras, even those equipped with stereo vision.

For this study, the OS1 mid-range Ouster sensor was selected for distance measurement. The sensor features a 10Hz laser scanner with a range of 170 meters at over 90% detection probability under sunlight up to 100klx (for objects with 80% Lambertian reflectivity in 1024 @ 10Hz mode). For objects with 10% reflectivity, it achieves a 90-meter range under similar conditions. It offers a vertical resolution of 64 channels and configurable horizontal resolutions of 512, 1024, or 2048. With a 360° horizontal and 45° vertical field of view (+22.5° to -22.5°) and angular accuracy of ±0.01°, the sensor provides precise and comprehensive coverage. Its low false-positive rate of 1 in 10,000 enhances reliability. Details are provided in Table 5.

The LiDAR sensor (Figure 9(a)) collected 15,497 timestamped scans, including point clouds (Figure 9(b)), reflectivity channel (Figure 9(c)), and range data (Figure 9(d)). The LiDAR's reflectivity data provides critical information about the surface characteristics of objects, enabling differentiation between materials and surfaces based on their light reflection properties. This is particularly useful for identifying features such as road markings, building facades, or vegetation within an urban environment. The range data complements this by precisely measuring the distance between the sensor and objects, forming the foundation for accurate 3D spatial representation. Combining reflectivity and range data enables the creation of highly detailed and accurate 3D maps, which are essential for applications like autonomous navigation, urban planning, and environmental modeling, where understanding both the geometry and material composition of

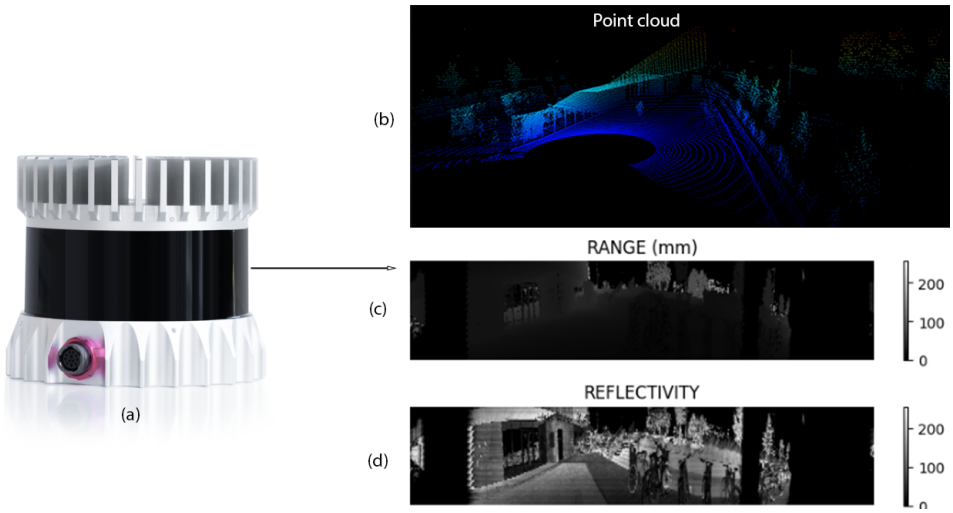



**Figure 9.** Example of (a) the LiDAR module, (b) point cloud, (c) range and, (d) reflectivity channel data collected from the sensor.

the environment is crucial.

### 4.3.5. Audio

In today's landscape of environmental assessments and soundscape evaluations, the significance of audio data is increasingly underscored. While urban sounds were conventionally scrutinized primarily for their contribution to noise pollution, recent perspectives have shifted towards acknowledging their subjective impact on human well-being [Sch+23]. Soundscape research has emerged as a pivotal field, combining in-situ surveys and audio recordings to comprehensively explore and manage sonic environments [PJ24]. This approach becomes essential when unraveling the auditory fabric of urban spaces, particularly in pedestrian-rich areas, where an intricate blend of human activity and environmental cues converge. The integration of auditory data into the DELTA dataset serves to enhance the development of nuanced, context-aware autonomous systems that can engage more effectively within dynamic pedestrian environments. By integrating this auditory layer with visual and spatial data, the dataset aims to provide a holistic view of urban environments, supporting the development of systems that navigate with an intuition akin to human perception.

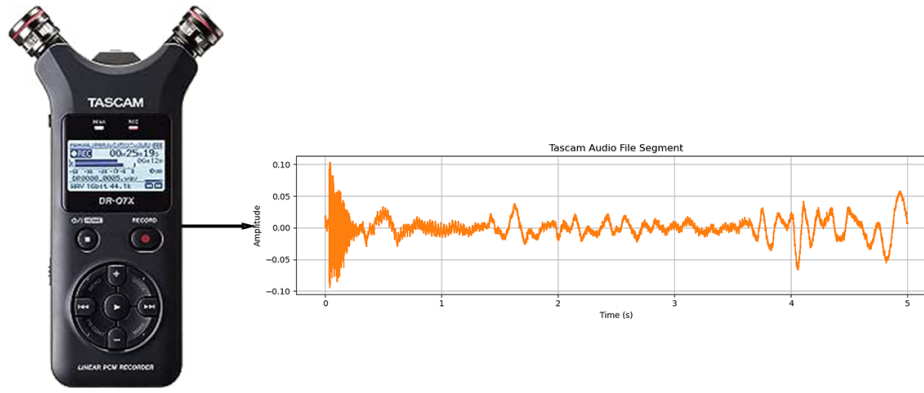
The Tascam DR-07X audio recorder (Figure 10(a)) was used for audio capture. It supports both stereo and mono recording (Figure 10(b)) and records high-resolution audio in WAV/BWF format at up to 96 kHz and 24-bit depth, ensuring accurate sound reproduction. Table 6 summarizes the module's specifications.

**Table 5.** LiDAR Ouster OS1 specifications.

| Specification | Value |
|---|---|
| Range (80% Lambertian reflectivity, 2048 @ 10Hz mode) | 100m @ >90% detection probability, 100klx sunlight<br>120m @ >50% detection probability, 100klx sunlight |
| Range (10% Lambertian reflectivity, 2048 @ 10Hz mode) | 45m @ >90% detection probability, 100klx sunlight<br>5m @ >50% detection probability, 100klx sunlight |
| Minimum Range | 0.3m for point cloud data |
| Range Accuracy | ±3cm for lambertian targets, ±10 cm for retroreflectors |
| Precision (10% Lambertian reflectivity, 2048 @ 10Hz mode, 1 standard deviation) | 0.3 - 1m: ±0.7cm<br>1 - 20m: ±1cm<br>20 - 50m: ±2cm<br>>50 m: ±5cm |
| Range Resolution | 0.3cm |
| Vertical Resolution | 64 channels |
| Horizontal Resolution | 512, 1024, or 2048 (configurable) |
| Field of View | Vertical: 45° (+22.5° to -22.5°)<br>Horizontal: 360° |
| Angular Sampling Accuracy | Vertical: ±0.01° / Horizontal: ±0.01° |

#### 4.3.6. 4K camera

A 4K camera is used for urban mapping and analysis to capture clear details. Its high resolution helps with object detection and classification, while the 155° wide field of view covers a large area of the environment. The video data works well with other sensors like LiDAR and GNSS to build a complete dataset for analysis. Lastly the built-in audio records the surrounding sounds. These features make the Osmo action camera 3 (Figure 11(a)) a useful tool for collecting data for urban studies and autonomous systems. Table 7 lists its specifications.




(a) (b)

**Figure 10.** Example of (a) Tascam audio recorder, and a (b) sample of the captured audio.

**Table 6.** Tascam-DR-07X specifications

| | |
|---|---|
| **WAV / BWF** | |
| Sampling Frequency | 44.1k / 48k / 96kHz |
| Bit-depth | 16 / 24-bit |
| **MP3** | |
| Sampling Frequency | 44.1k / 48Hz |
| Bit-rate for recording | 32k / 64k / 96k / 128k / 192k / 256k / 320k bps |
| Playback Bit-rate | 32k to 320k bps, VBR, ID3TAG Ver. 2.4 |
| **Number of Channels** | 2-channel (Stereo) / 1-channel (Mono) |
| **Playback Speed Control** | 0.5 to 2.0 times (in 0.1 increments) * 44.1k/48kHz only |
| **Built-in Microphone** | Unidirectional, Stereo (switchable between A-B and X-Y positions) |
| **MIC/EXT IN** | |
| Connector | 1/8" (3.5mm) Stereo Mini jack (Unbalanced, Plug-in Power) |
| Nominal Input Level | -20dBV |
| Maximum Input Level | -4dBV |
| Input Impedance | 18kΩ or more (PLUG IN PWR: OFF) / 2kΩ or more (PLUG IN PWR: ON) |

## 4.4. Experimentation setup

As part of the data collection process, I used the DELTA e-scooter platform to cover a 3.75-kilometer route around the Delta building in Tartu, Estonia. Throughout the ride, the platform captured a wide range of data from its onboard sensors,

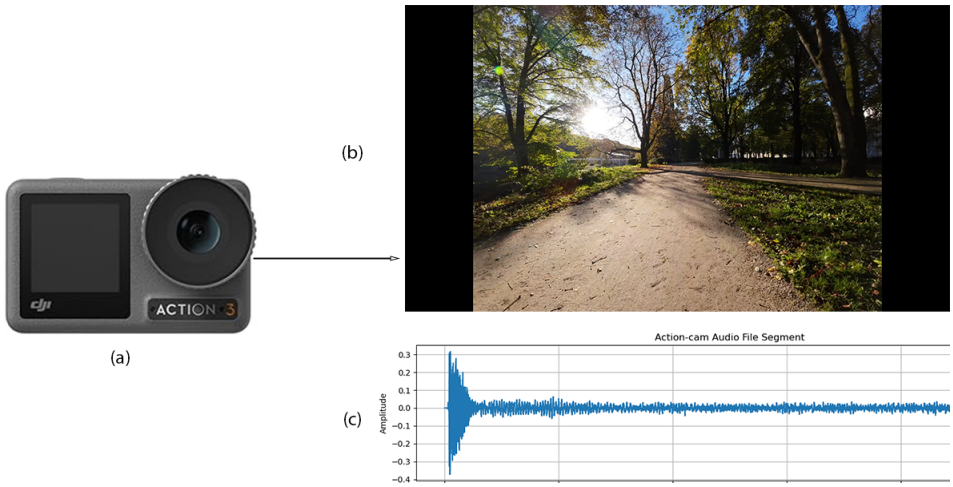



**Figure 11.** Example of (a) the Osmo action camera 3, (b) an RGB image sample, (c) an audio sample from the camera.

**Table 7.** Osmo action camera 3 specifications

| **Camera Specifications** | |
|---|---|
| Sensor | 1/1.7-inch CMOS |
| Lens | FOV: 155°<br>Aperture: f/2.8<br>Focus Range: 0.3 m to infinity |
| Audio Output | 48kHz; AAC |
| Standard Recording | 4K (4:3): 4096x3072@24/25/30/48/50/60fps<br>4K (16:9): 3840x2160@100/120fps<br>4K (16:9): 3840x2160@24/25/30/48/50/60fps<br>2.7K (4:3): 2688x2016@24/25/30/48/50/60fps<br>2.7K (16:9): 2688x1512@100/120fps<br>2.7K (16:9): 2688x1512@24/25/30/48/50/60fps<br>1080p (16:9): 1920x1080@100/120/200/240fps<br>1080p (16:9): 1920x1080@24/25/30/48/50/60fps |

including over 15,000 LiDAR scans, more than 91,000 stereo image sequences, and nearly 46,000 high-resolution 4K images. I also recorded 25 minutes of ambient audio and gathered around 1,600 GNSS location points. This dataset offers a detailed snapshot of a real urban environment and plays a key role in benchmarking my work on mapping, localization, and pedestrian-focused analysis.

## 4.5. Sensor synchronization

### 4.5.1. Overview of sensor synchronization

Sensor synchronization refers to the process of aligning the timing and data acquisition of multiple sensors to ensure that their measurements are temporally correlated. This is crucial for various applications, such as robotics and autonomous systems, where accurate sensor integration is required for reliable perception, decision-making, and real-time interaction with the environment. Achieving synchronization can involve methods such as aligning sampling instants, compensating for clock drift, and ensuring that all sensor measurements reference a common time frame. Various approaches are available, including hardware-based techniques (e.g., master-slave configurations and shared clock sources), external methods (e.g., trigger signals, GNSS time references, and network synchronization protocols), and software-based algorithms. The following sections provides more detailed exploration of these topics.

*Internal synchronization:.* Internal synchronization commonly employs two approaches: the master-slave configuration and the shared clock source method, each offering unique mechanisms and implications for system design. The master-slave configuration is a hierarchical approach where one sensor operates as the master, dictating the timing for data sampling to the other sensors, designated as slaves. This method hinges on synchronization signals sent from the master to the slave sensors, ensuring that all sensors in the network sample data concurrently. While this approach is relatively straightforward to implement, it is somewhat rigid due to its reliance on a single master device to coordinate the timing of all data collection activities. This configuration can introduce a single point of failure in the system; if the master sensor encounters an issue, the entire network's synchronization can be compromised. Additionally, this method may be less scalable, as adding more sensors necessitates more complex coordination and communication from the master device. Alternatively, the shared clock source method achieves synchronization by connecting all sensors in the network to a common clock signal. This ensures that each sensor samples data in unison, based on the same temporal reference. The primary advantage of this approach is its potential for high accuracy, as all sensors are directly tied to a singular, precise timing source. In this setup each sensor operates independently in terms of timing, reducing the risk associated with a single point of failure and facilitating easier expansion of the sensor network.

*External synchronization:.* External synchronization can be achieved through three advanced methods: trigger signals, time reference signals, and network synchronization protocols, each tailored to suit specific scenarios and applications.

A trigger signal method involves an external signal that prompts all sensors in the system to commence sampling simultaneously. This approach is particularly effective for capturing synchronized data snapshots in response to specific events. The primary challenge here lies in the need for precise timing control of

the trigger source to ensure all sensors are activated at the exact same moment. This method is often employed in applications where event-based data collection is critical, such as in experimental physics, high-speed photography, or event detection in security systems. The accuracy of synchronization depends heavily on the reliability and precision of the trigger mechanism.

The use of a dedicated time reference signal, such as the time provided by the GNSS, offers another strategy for sensor synchronization. By supplying a common timestamp derived from GNSS time signals, sensors distributed over large geographical areas can achieve precise time alignment. This method is invaluable for applications that require consistent time stamps across extensive distances, including environmental monitoring, geological surveys. The GNSS time signal provides a universally accurate time source, ensuring that data from disparate sensors can be accurately aligned and compared.

Network synchronization protocols, such as the Network Time Protocol (NTP) and the Precision Time Protocol (PTP), are designed to synchronize the clocks of devices across a network. NTP, being one of the oldest and most widely used protocols, offers millisecond-level accuracy, making it suitable for general data consistency purposes in computing environments. PTP, on the other hand, can provide sub-microsecond accuracy, catering to applications that demand higher precision, such as in telecommunications, power grid control, and financial trading systems. These protocols ensure that all devices in a network operate on a consistent timeline, facilitating coordinated data collection and analysis in a controlled network environment.

*Software-based synchronization:.* When real-time synchronization is not feasible or implemented, two widely used methods for software-based synchronization are post-processing alignment and event-based synchronization. Both approaches offer distinct applications, benefits, and challenges.

Post-processing alignment involves the independent collection of data from various sensors, followed by the alignment of this data using software algorithms during the analysis phase. This method is particularly useful in scenarios where sensors operate on different sampling rates or when real-time synchronization is not possible. The alignment process typically involves time-stamping data at the point of collection and subsequently using these timestamps to align the datasets accurately. The primary advantage of post-processing alignment is its flexibility, as it allows for the integration of data from a wide variety of sources without the need for complex hardware or synchronization protocols during data collection. However, this method can be computationally intensive, requiring significant processing power and sophisticated algorithms to match data points accurately across time. This can introduce delays in analysis and may not be suitable for applications requiring immediate data interpretation.

Event-based synchronization, on the other hand, relies on the occurrence of common events as reference points for data alignment. Sensors are configured to detect specific stimuli, such as sound or light pulses, which serve as markers for

synchronizing data during post-processing. This approach ensures that despite the independent operation of each sensor, the recorded data can be precisely aligned based on the timestamps of these detected events.

Implementing event-based synchronization necessitates precise control over the generation of synchronization events and the accurate detection of these events by all involved sensors. This method is particularly effective in controlled environments where the timing and characteristics of the synchronizing events can be tightly managed. While it reduces the computational load compared to arbitrary post-processing alignment, it requires careful setup and calibration of the sensors to ensure the reliable detection of synchronization events.

### 4.5.2. Synchronization of multisensory data on DELTA dataset

In the DELTA project, various post-processing methods were employed for sensor synchronization using the Jetson board's internal clock as a common time reference. The following sections describe how image files from the stereocamera and action camera — despite lacking a shared clock — were manually matched, how sensor timestamps were automatically aligned using a KD-tree algorithm, and how audio recordings from the dedicated audio recorder were synchronized with the action camera's audio file to ensure all data sources were properly aligned.

**Post-processing method (Manual image file synchronization):** To manually sync images, I built a simple GUI that makes it easy to align photos from the 4K camera with reference images from the stereocamera. It is possible to scroll through each dataset using the arrow keys and press 's' to save an image when it closely matches the stereocamera reference. While this process takes time, it was necessary to ensure the 4K images lined up correctly with the other sensor data. This hands-on approach helped maintain accuracy across different data sources, making the dataset more reliable for analysis. Figure 12 shows the GUI in action.

**Post-processing method (Proximity-based synchronization):** In this approach, each sensor reading is recorded with an exact timestamp. After data collection, a KD-tree algorithm [Ben75] is used to match each reading to the nearest timestamp from the ZED2i stereo camera, which acts as the reference. This method is applied to every sensor, aligning their data in time. Even though sensors capture data at different moments, the KD-tree ensures a synchronized dataset for reliable cross-sensor analysis.

**Post-processing method (Cross-correlation based audio file synchronization):** The subsequent sections detail the methodical approach undertaken to synchronize audio recordings from both the action camera and audio recorder.

*Cross-correlation for time lag detection:* The cornerstone of the audio synchronization process is the use of cross-correlation, a mathematical operation that measures the similarity between two signals as a function of the displacement of one relative to the other. By computing the cross-correlation between two audio signals, the algorithm identifies the point at which the signals exhibit the high-

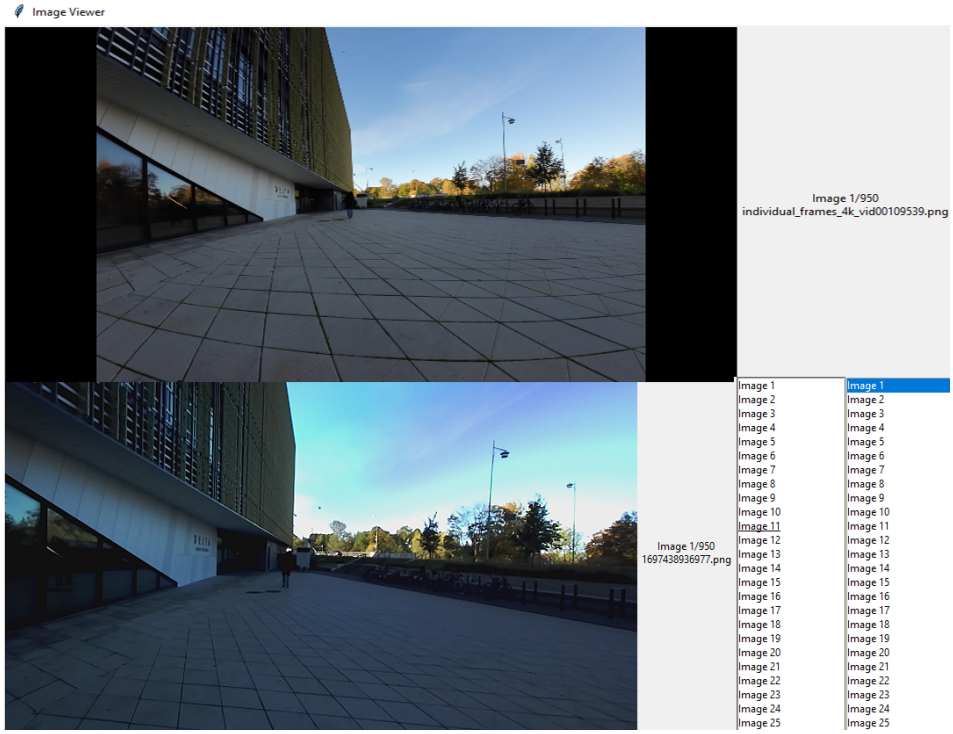



**Figure 12.** A screenshot of the GUI interface used for manually selecting the matching images from the stereocamera and the action camera for the purpose of synchronization.

est degree of similarity. This point corresponds to the optimal time lag or shift required to align the signals in time. The time lag is extracted by finding the maximum value in the correlation array, indicating the most aligned position between the two audio streams.

*Sample rate consistency:* A prerequisite for accurate cross-correlation and, consequently, effective synchronization is the consistency of sample rates across audio signals. Sample rate, measured in hertz, represents the number of samples of audio carried per second. If the audio signals have different sample rates, they are resampled to the higher of the two rates. This step ensures that the temporal characteristics of the signals are preserved and comparable, allowing for a meaningful cross-correlation analysis.

*Signal alignment:* Signal alignment is the process of adjusting the time positions of audio signals so that corresponding events in each signal occur simultaneously. To achieve this alignment, I use a technique called padding for time alignment, which involves adding silent audio samples to the beginning or end of a signal. This padding effectively shifts the signal in time without altering its content, ensuring that even if recordings start at different moments, they can be synchronized accurately.

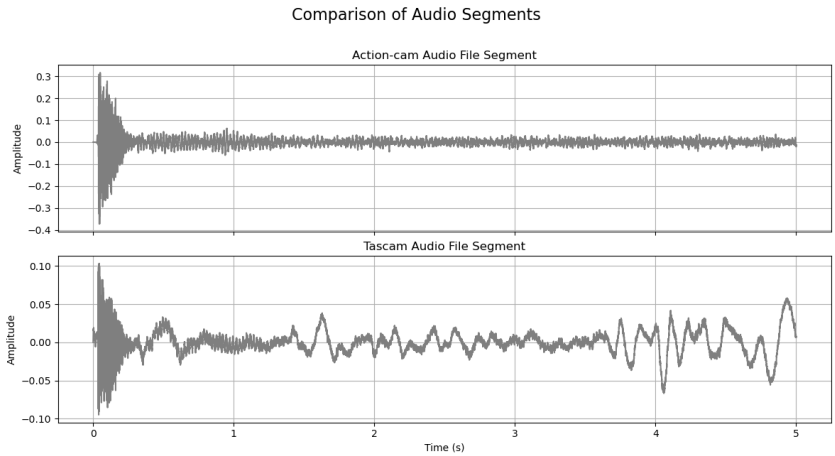



**Figure 13.** Example showcasing the aligned audio waveforms from the action camera (top) and the Tascam recorder (bottom), demonstrating synchronization over a 5s interval with corresponding time(s) and amplitude variations.

Figure 13 illustrates a segment where the two audio files have been successfully synchronized, demonstrating the effectiveness of this adjustment in achieving temporal coherence between different audio sources.

## 4.6. Sensor fusion

Sensor fusion, also known as multisensor fusion, is a crucial technique in various fields, including robotics, autonomous vehicles, and navigation systems. It entails combining data from multiple sensors to enhance the accuracy, robustness, and overall performance of a system. By leveraging the strengths of diverse sensors, sensor fusion addresses the limitations of individual sensors, yielding more comprehensive and reliable information.

### 4.6.1. Overview of sensor fusion

Sensor fusion can be categorized into various approaches based on the level of fusion and the type of sensors involved. Here's a breakdown of the prominent variations:

**Data-level fusion:** This approach involves directly merging raw data collected from various sensors. The process does not account for the spatial or temporal relationships between the data points. Instead, it focuses on acquiring relevant elements from each sensor's output and combining them into a single, cohesive data representation. This form of fusion is particularly useful in creating a broad overview of the sensed environment.

**Feature-level fusion:** In this method, the focus is on utilizing specific features extracted from each sensor's data. These features could include distinctive points,

edges, or textures, which are then used to establish correspondences between different sensor readings. By aligning and matching these features across various sensors, the system can identify corresponding objects or regions in the environment. This approach enhances the system's ability to interpret and understand complex scenarios by providing a more nuanced view of the surroundings.

**Decision-level fusion:** This approach takes a more holistic view, integrating sensor data at the crucial stage of decision-making. Here, the information derived from different sensors is synthesized at a higher level, where it informs key decisions. These decisions could range from obstacle avoidance in autonomous vehicles to route planning in navigation systems. By combining sensor inputs at this stage, the system can make more informed, accurate, and reliable decisions based on a comprehensive understanding of its environment.

Sensor fusion algorithms try to combine different sensor data into an overall view of the world. The topic of sensor fusion includes a variety of algorithms, each tailored to specific scenarios and system requirements. Among the most well-known and established are the Kalman filter [WB+95], particle filter [Dju+03], and Bayesian networks [HW95]. The Kalman filter has been recognized for its effectiveness in cases where system dynamics and noise are linear and Gaussian. It uses a two-step model: *prediction* and *update*. The prediction step projects the current condition forward in time, while the update step modifies the projection depending on new measurements. It is an optimal estimator under its operational assumptions and is renowned for its computing efficiency, making it ideal for applications such as vehicle tracking and navigation. However, the Kalman filter has its constraints, especially when dealing with non-linear dynamics or non-Gaussian noise. Here, the particle filter shines, offering a robust alternative for non-linear and non-Gaussian processes. Representing the state distribution with a swarm of weighted particles, it adapts to new measurements by updating the weights and resampling to focus on probable states. The particle filter is versatile and powerful but can suffer from computational intensity and challenges with high-dimensional state spaces.

Given the complexity and non-linearity of real-world environments, it is crucial to choose a sensor fusion algorithm that can handle such conditions effectively. This leads us to the technique of 3D to 2D projection, a method that involves transforming three-dimensional LiDAR data into a two-dimensional image plane. This technique is particularly compelling because it directly leverages the high-resolution and rich color data from cameras, and the precise depth information from LiDAR, thus capitalizing on the strengths of both sensing modalities.

In this work, a custom calibration and projection methodology was designed to achieve data-level fusion between LiDAR and camera data. The technique places raw 3D LiDAR outputs directly onto 2D camera images by mapping each point into the camera's coordinate system, making it possible to overlay LiDAR data onto the camera image. Consequently, this method is categorized as data-level fusion.

### 4.6.2. LiDAR-to-camera calibration and projection

Fusing LiDAR depth data with the color and texture details of a camera provides a rich, multidimensional view of the environment. Achieving this fusion involves carefully aligning coordinate systems and calibrating both intrinsic and extrinsic camera parameters so that 3D LiDAR data maps accurately onto 2D images.

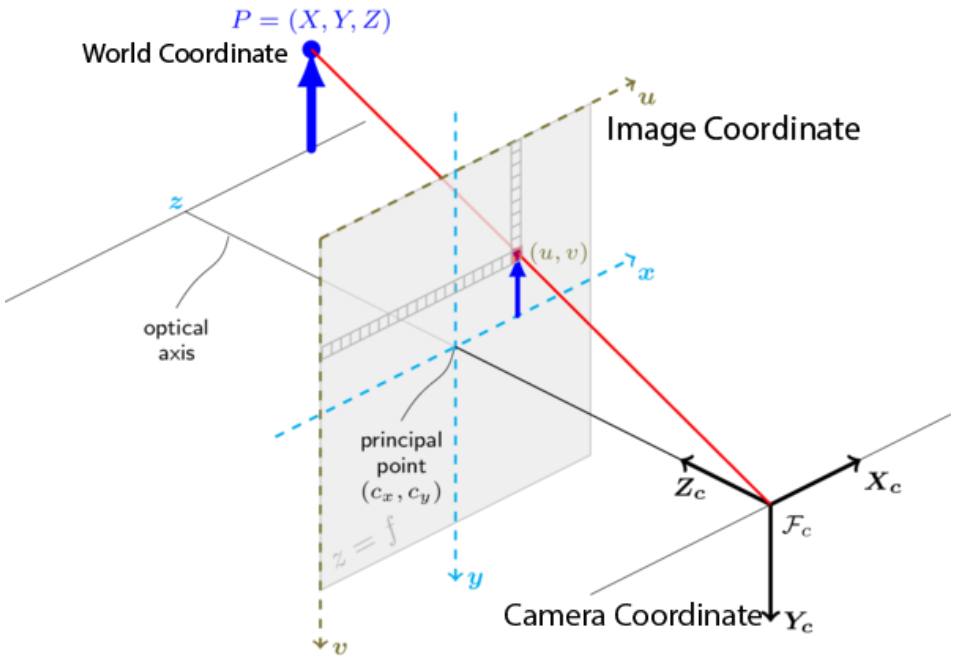



**Figure 14.** Coordinate systems involved in camera projection (sourced from [24]).

**Overview of 3D-to-2D projection.** Points in the *world coordinate system* are represented by $(X_w, Y_w, Z_w)$ and have a defined origin, orientation, and scale (Figure 14). These points are transformed into the *camera coordinate system*, $(X_c, Y_c, Z_c)$, according to the camera's location and orientation. Finally, the *pixel coordinate system*, $(u, v)$, describes how the camera perceives these points on a 2D image plane. Intrinsic parameters—focal length $(f_x, f_y)$, principal point $(C_x, C_y)$, and skew $s$—populate the projection matrix $P$, typically of size $3 \times 4$. Homogeneous coordinates manage rotations, translations, and scaling. The relationship between these systems can be written as:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = P \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix}. \tag{4.1}$$

Depth normalization is then applied by dividing each $(X, Y, Z)$ by $Z$, ensuring correct depth representation before projecting the points to $(u, v)$ on the image.

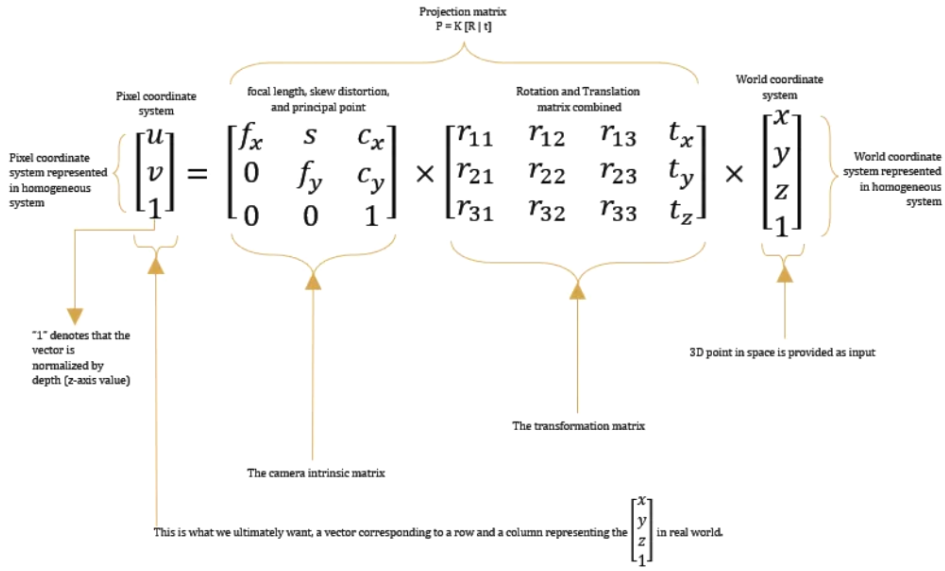



**Figure 15.** Overview of the 3D-to-2D projection mathematical concept.

**Calibration setup and data collection.** A *target-based calibration* approach was used to align LiDAR data with camera imagery. A black square target, shown in Figure 16, was placed at a known distance and orientation. The LiDAR and camera were mounted on a stable frame in a vertical orientation to capture this static scene. Because LiDAR resolution is limited in the vertical axis, specific attention was directed to target size and distance, ensuring edges and corners remained clearly visible to both sensors.

**Data preprocessing.** Raw point clouds and camera images were collected for calibration. During preprocessing, the LiDAR point clouds underwent manual filtering based on depth and azimuth criteria to remove irrelevant sections and retain only the target. Figure 17 illustrates the result of isolating the black target in 3D space.

**Correspondence establishment.** Eight key 3D points (corners and edges) were identified on the target after filtering. These points were manually matched to their 2D locations in the camera images by means of a custom GUI (Figure 18). Transformation matrix parameters, including rotation and translation, were adjusted with sliders in the interface, ensuring precise alignment between the 3D LiDAR points and the camera's 2D features.

**Calibration parameter estimation.** The matched points formed the basis for estimating extrinsic calibration parameters. OpenCV's `solvePnP` function was employed to compute rotation ($R$) and translation ($t$). Intrinsic camera parameters (e.g., $f_x, f_y, c_x, c_y$) and distortion coefficients were also considered. This step determines how the 3D coordinates of the target map onto the camera's 2D plane,

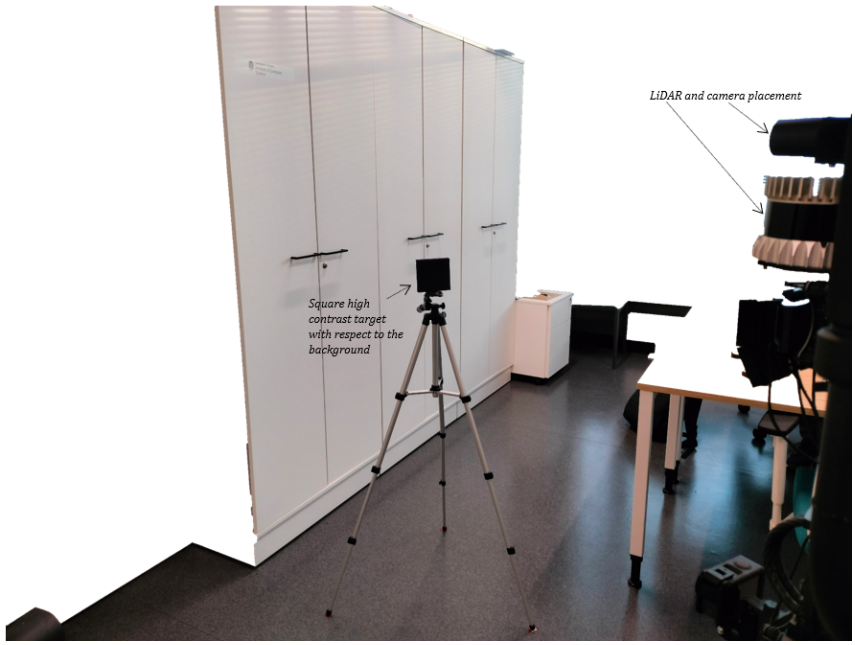



**Figure 16.** Target-based calibration setup, ensuring both LiDAR and camera have a clear view of the planar surface.

expressed by:

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K[R \mid t] \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}. \tag{4.2}$$

**Projection.** Once the rotation and translation vectors were estimated, I used OpenCV's projectPoints function to re-project the LiDAR point cloud onto the camera image. This step applies the extrinsic calibration parameters—namely, the rotation and translation—to accurately place the 3D data within the 2D camera view. Figure 19 illustrates the entire workflow and shows the final outcome of projecting the 3D points onto the target image.

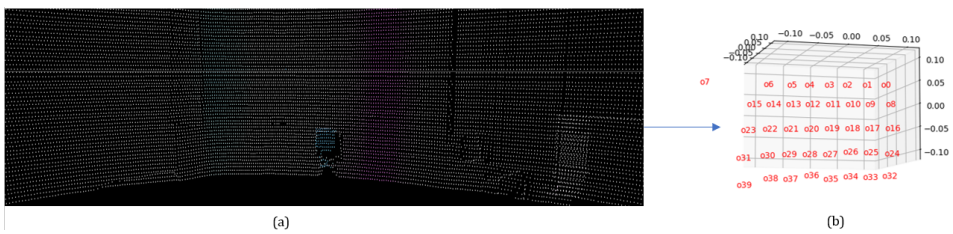



**Figure 17.** (a) Full scene LiDAR point cloud, (b) isolated point cloud of the black target.

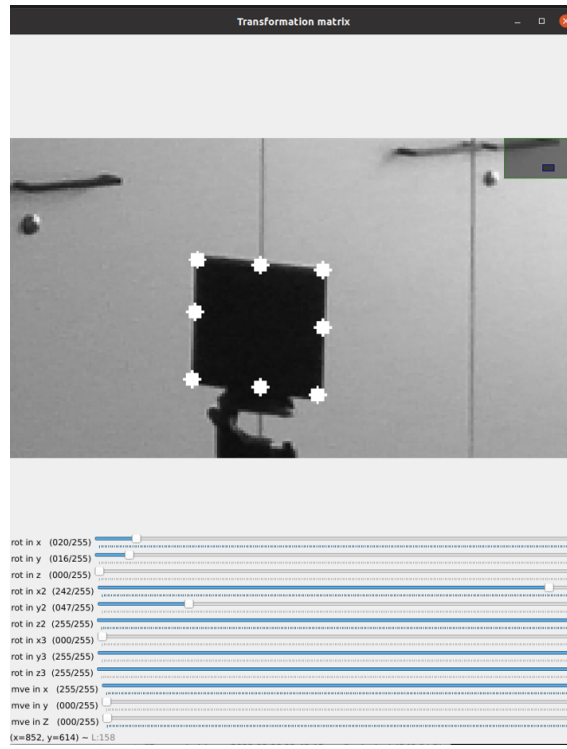



**Figure 18.** A custom GUI for LiDAR-to-camera alignment. Transformation matrix parameters can be fine-tuned for accurate matching of 3D points to 2D image locations.

## 4.7. Data annotation and segmentation

Data annotation is a key step in AI and machine learning, where data—such as images, videos, or text—is labeled to help models recognize and understand patterns. This process is especially important for supervised learning, where models rely on well-labeled datasets to learn and improve. Annotation basically has two main steps: first, the data is labeled, and then it is reviewed to ensure accuracy. This approach helps create reliable datasets that improve the performance of machine learning models.

### 4.7.1. Overview of data annotation and segmentation

There are several forms of data annotation, each tailored to different types of data and intended outcomes:

**Image annotation:** This involves the identification and labeling of images using techniques such as bounding boxes, polygonal segmentation, or pixel-wise segmentation masks. Such annotations can delineate objects within an image, classify entire images, or segment images into constituent parts, thereby aiding in object recognition, scene understanding, and more.

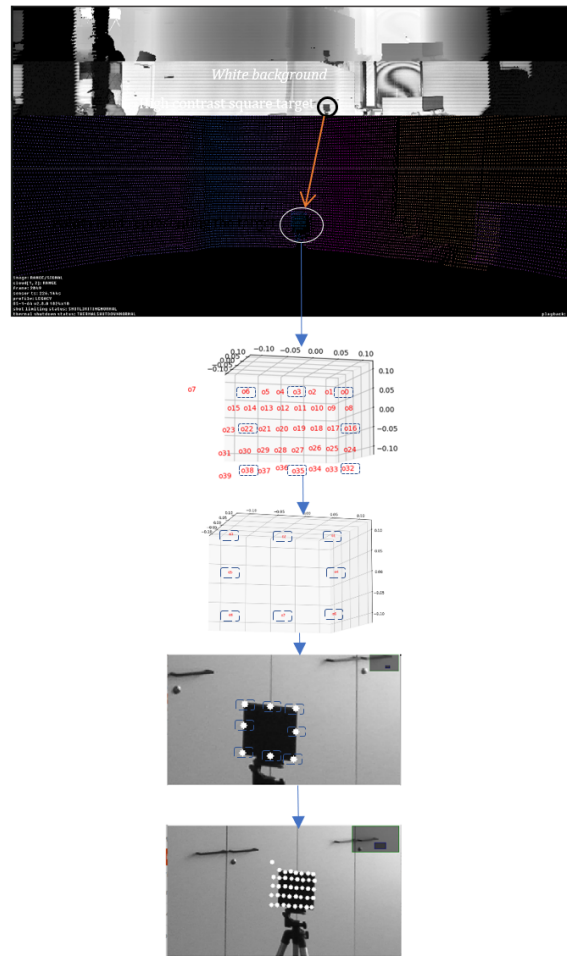



**Figure 19.** Calibration workflow, from isolating the target in the LiDAR point cloud (top image) to projecting the computed 3D points onto the camera image (bottom image) using `projectPoints`.

**Video annotation:** Similar to image annotation but applied across video frames, video annotation can track the movement of objects over time, recognize actions, and analyze behaviors. It is instrumental in applications requiring temporal and spatial analysis, such as surveillance systems and dynamic scene understanding.

**Audio annotation:** This involves labeling or tagging audio recordings to identify and classify specific sounds, speech, or acoustic features within them. This process enables machine learning models to recognize patterns, such as spoken words, music genres, environmental sounds, or emotional cues in speech, in new audio data.

**Text annotation:** This pertains to the categorization or tagging of text data with metadata, labels, or classifications. It supports Natural Language Processing (NLP) tasks like sentiment analysis, topic classification, and entity recognition, enabling machines to comprehend and generate human-like text responses.

Data annotation can be performed manually, where human annotators label the data, or automatically, through semi-supervised learning techniques that leverage algorithms to annotate data, which is then refined by human oversight. The choice between manual and automatic annotation hinges on the required accuracy, complexity of the data, and available resources.

The importance of data annotation extends beyond industries, including applications in healthcare for medical imaging analysis, and automotive for building autonomous vehicle navigation systems, among others. Data annotation not only improves the quality and reliability of machine learning models, but it also accelerates the progress of intelligent technologies across industries by making it easier to construct AI-enabled solutions. There are several key aspects of image annotation:

*Image classification.* Image classification plays a crucial role in the field of computer vision, where it involves assigning images to predefined categories using machine learning techniques. This process typically uses labeled datasets containing diverse image classes—such as "leopard" or "motor scooter" (20)—to train classifiers like convolutional neural networks (CNNs). These CNNs excel in extracting discriminative features from the training images, which enables them to recognize patterns effectively. Once trained, these classifiers can apply their learned data representations to accurately identify and categorize new, unseen images, thus automating the classification process. The effectiveness of these models is assessed based on their precision and recall, which measure their ability to correctly and consistently classify images within the categories they have learned. This capability is fundamental for applications that require reliable and automatic image understanding.

*Object recognition/detection.* Object recognition and detection refine the process of image annotation by focusing on labeling individual objects within an image, rather than assigning a single label to the whole image. For instance, in an image featuring a car, a bicycle, and a dog (21), each object would be separately identified and labeled. In addition to recognizing and labeling multiple entities within a single image, this approach involves defining the boundaries of each object using techniques like bounding boxes or pixel-wise segmentation.

*Segmentation.* Image segmentation elevates the process of annotation by partitioning an image into distinct regions, each representing a unique object or a specific area of interest. This intricate form of annotation is crucial for tasks that require an in-depth understanding of an image's content at a pixel level. Segmentation comes in three main varieties:

Semantic segmentation [LSD15] (Figure 22(b)) : This type goes beyond mere detection. It involves classifying each pixel in an image into categories, making it possible to group different parts of the image by their class. For example, in a street scene, semantic segmentation would classify pixels as belonging to roads, buildings, cars, pedestrians, etc., often without distinguishing between different instances of the same class.

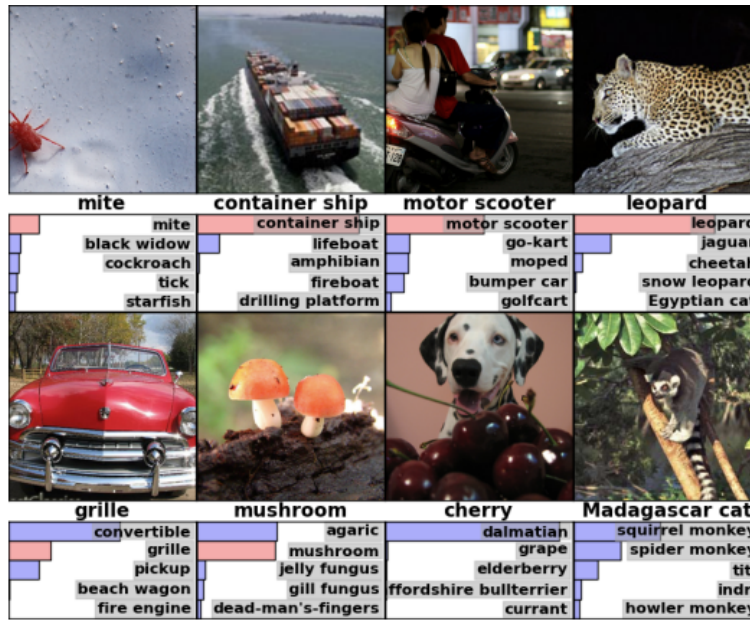



**Figure 20.** Five labels considered most probable for image classification for each image shown (source [KSH12]).

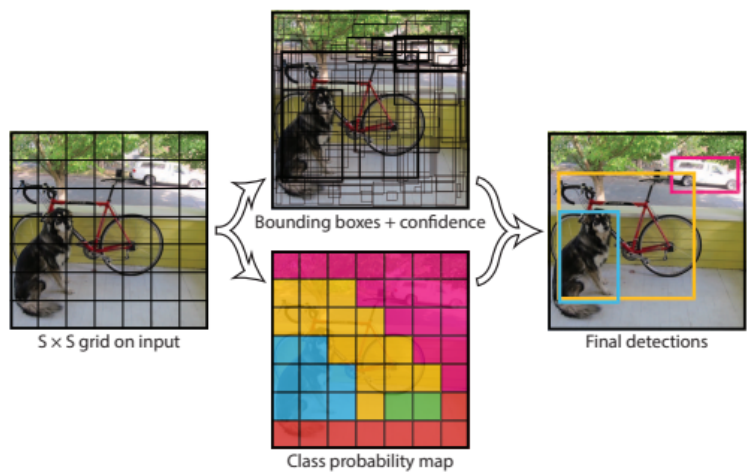



**Figure 21.** Visualization of the YOLO object detection process: (left) input image with an S × S grid, (top middle) bounding boxes with confidence scores, (bottom middle) class probability map, and (right) final detections (source [Red+16])

Instance segmentation [He+17] (Figure 22(c)): This type provides a more nuanced understanding by not only labeling every pixel of an object but also differentiating between individual instances. For instance, it would distinguish between each car on a road, assigning a unique identifier to each one.

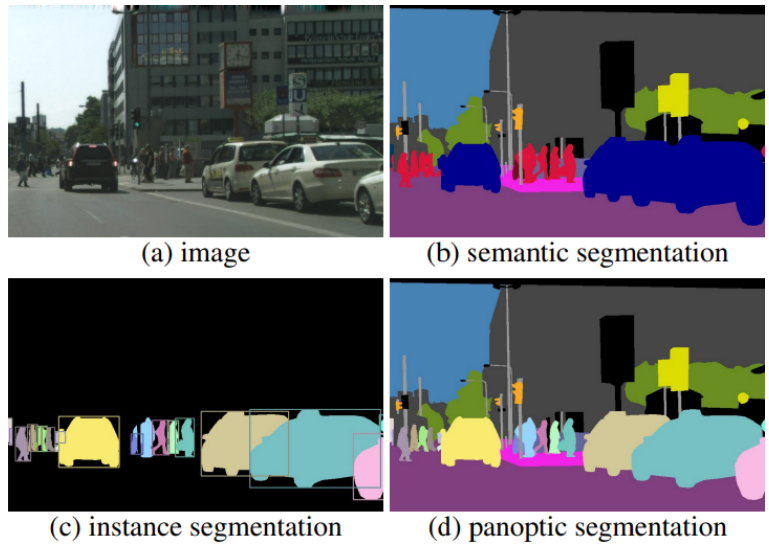



**Figure 22.** Types of segmentation, (a) street level image, (b) semantic segmentation, (c) instance segmentation and, (d) panoptic segmentation (source [Kir+19]).

Panoptic segmentation [Kir+19] (Figure 22(d)): This comprehensive approach merges the capabilities of semantic and instance segmentation. It provides a complete view by simultaneously distinguishing between the class of each pixel and identifying individual instances of objects. In fields like medical imaging, autonomous vehicle navigation, and satellite image analysis, segmentation is essential. It allows for precise delineation of objects and a clear understanding of their spatial dynamics, leading to more accurate diagnoses, safer autonomous driving, and better geographical data interpretation. The success of these applications hinges on the accurate and granular analysis that segmentation provides.

*Annotation process.* There are primarily three methods used in the annotation process: manual, automated, and hybrid annotation. Manual annotation relies on human expertise to accurately label data, while automated annotation utilizes pretrained algorithms to speed up the process. Often, a hybrid approach is employed, combining both methods to balance efficiency and accuracy. Understanding these methods and their implications is essential for effectively training and deploying machine learning models in various applications.

**Manual annotation**: This process involves human annotators carefully labeling images by identifying and marking objects, features, or areas based on the project's needs. Specialized software tools help make the task more efficient, offering features like zooming, drawing bounding boxes or polygons, and assigning labels to different elements. While manual annotation is highly accurate, especially for complex images that require understanding subtle details, it is also time-consuming and expensive, requiring significant human effort—especially for large datasets.

**Automated annotation**: It involves using pre-trained machine learning or deep learning algorithms to label elements in images. These algorithms, trained on extensive datasets, are adept at identifying and labeling standard elements in new images, making them highly efficient for straightforward tasks like creating segment masks or recognizing common objects. Automated tools, which utilize pre-trained models tailored for specific tasks, offer rapid processing capabilities, handling large volumes of data much faster than human annotators. However, they may lack the accuracy of manual annotation, particularly in complex scenarios where nuanced understanding of context is crucial. Their performance can falter with images that deviate significantly from their training data.

**Hybrid approach**: It combines the strengths of both manual and automated methods. Initially, automated tools perform a first pass of annotation, which is then reviewed and refined by human annotators. This strategy leverages the speed of automated tools while ensuring the accuracy through manual review. It significantly reduces the workload on human annotators by handling simpler tasks automatically, thus making the overall process more efficient. Furthermore, the hybrid approach allows for quality control, as human annotators can correct any errors or inaccuracies introduced by the automated tools, resulting in a high-quality, accurately annotated final dataset.

The Segment Anything Model (SAM) provided by Roboflow [Rob24] was used for annotating images within two distinct datasets: sidewalks and pedestrian route segmentation, and landmark segmentation. This process entailed uploading my datasets to the platform, enabling me to harness Roboflow's user-friendly labeling and augmentation tools to prepare my labeled data for training.

### 4.7.2. Sidewalk and pedestrian route segmentation on visual DELTA dataset

Sidewalk and pedestrian route segmentation is a crucial task in autonomous navigation, pedestrian navigation, and urban planning. This task aims to identify and extract the regions occupied by sidewalks and pedestrian paths from urban scenes. Accurate sidewalk and pedestrian route segmentation is essential for various applications, including:

- *Autonomous navigation:* Autonomous vehicles and robots need to be able to recognize and follow sidewalks and pedestrian paths in order to navigate safely and efficiently.
- *Pedestrian navigation:* Pedestrian navigation systems need to be able to provide accurate and reliable information about the location of sidewalks and pedestrian paths in order to help pedestrians find their way around.
- *Urban planning:* Urban planners need to be able to understand the use of sidewalks and pedestrian paths in order to make informed decisions about urban design and infrastructure.

In the DELTA project, a focused approach to sidewalk and pedestrian route segmentation was developed, emphasizing the integration of multimodal data and advanced deep learning techniques. At the heart of this process is the use of the SegFormer architecture, introduced in a 2021 study [Xie+21]. SegFormer is an image segmentation model that uses Transformer models in a simpler and more efficient way. Its hierarchical encoder extracts features at multiple scales without relying on positional encodings, allowing it to work well with different image resolutions. The model also uses overlapping patch embeddings and a hybrid downsampling method to improve feature extraction.

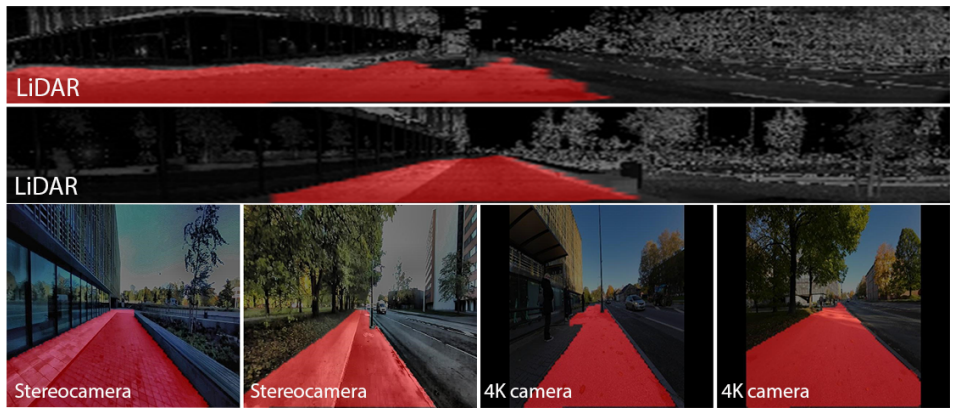



**Figure 23.** Examples of segmenting sidewalks and pedestrian routes: Six segmented images from LiDAR (reflectivity channel), stereo-camera (right-side camera), and 4K camera are displayed with color-coded labels (highlighted in red) from different geographical locations.

Contrary to the complex decoders prevalent in other Transformer-based segmentation models [Dos+20], SegFormer adopts a straightforward multilayer perceptron decoder. This lightweight decoder aggregates multi-scale information from the encoder, effectively capturing both the local and global context with a minimal parameter footprint. The integration of the encoder and decoder within a unified framework enhances the model's efficiency by eliminating the need for separate training stages. Moreover, the simplicity of the SegFormer architecture facilitates ease of understanding and implementation, promoting further exploration and development in the field. I implemented the SegFormer architecture across three distinct datasets—each comprising 950 images (for a total of 2850 images)—characterized by varying resolutions and data types: a high-resolution 4K monocular camera ($4096 \times 3072$), a stereocamera ($720 \times 405$), and LiDAR reflectivity images ($1024 \times 64$). This approach enables to address the unique challenges posed by each dataset and assess the precision of segmentation outcomes enables addressing the particular issues given by each dataset and evaluating the precision of segmentation findings. Notably, the 4K monocular camera dataset achieved the highest accuracy in segmentation, illustrating the importance of high-resolution, textured data. However, both the stereocamera and LiDAR

datasets also provided valuable insights, albeit with varying levels of precision.

Table 8 presents the performance evaluation metrics for models applied to the reflectivity channel, 4K, and ZED2i datasets, highlighting their segmentation and classification capabilities. The reflectivity channel model achieved a Dice coefficient of 0.8452, supported by a frequency-weighted accuracy of 0.9829 and a mean IoU of 0.8461, indicating reliable prediction accuracy and classification performance. The model trained on the 4K dataset outperformed others, achieving the highest Dice coefficient of 0.9676, a mean IoU of 0.9603, and a frequency-weighted accuracy of 0.9847, reflecting exceptional segmentation and predictive accuracy. Similarly, the ZED2i dataset model delivered robust results, with a Dice coefficient of 0.9384, a mean IoU of 0.9267, and a frequency-weighted accuracy of 0.9774, showcasing its strong segmentation performance and classification reliability.

**Table 8.** Results of segmentation model evaluation on three datasets (DC: Dice Coefficient, FWA: Frequency Weighted Accuracy).

| Dataset | DC | FWA | Mean IoU |
|---|---|---|---|
| Reflectivity channel | 0.8452 | 0.9829 | 0.8461 |
| 4K | 0.9676 | 0.9847 | 0.9603 |
| ZED2i | 0.9384 | 0.9774 | 0.9267 |

The diversity of sensor types requires the creation of specialized models. For instance, while models trained on the 4K monocular camera excel in processing rich texture and color details, they may not perform as well on data from stereocameras or LiDAR, where the details and resolution differ significantly. Moreover, employing a multisensor approach with models specialized for each sensor type introduces a critical fail-safe mechanism, particularly crucial in applications like autonomous vehicles where reliability is paramount. This redundancy enables the system to switch between data sources and models depending on the situation, ensuring continuous operation even if one sensor fails or its data quality is compromised by environmental factors.

### 4.7.3. Audio classification on the auditory DELTA dataset

In the evolving landscape of environmental assessments, the field of auditory analyses is gaining prominence alongside traditional visual evaluations, offering a fresh perspective on urban soundscapes. Historically, urban sounds have been relegated to the category of noise pollution, with a focus on measuring and mitigating outdoor noise—especially that originating from vehicles and industrial activities—due to its adverse effects on human health and well-being. This conventional approach often simplified the rich tapestry of urban sounds into mere decibel levels, neglecting the complex auditory experiences these soundscapes provide [YK13; Mor+18].

The recent shift towards soundscape evaluations marks a significant departure from earlier methodologies. This new paradigm promotes a comprehensive approach, taking into account not only the physical properties of sound but also the subjective human experiences of auditory settings. Soundscape [PM85] research advocates for a nuanced approach to managing urban sonic environments, recognizing the impact of sound on individuals perceptions and experiences of urban spaces. Methodologies in soundscape evaluation typically involve direct observations and in-situ surveys aimed at understanding auditory perceptions associated with specific locales [BC21; DM23]. Advances in technology, such as head-mounted displays and virtual reality [LL23], are being leveraged to simulate urban environments for research purposes, focusing on aesthetic preferences and satisfaction levels among participants. Despite these innovations, the application of such methodologies often faces limitations in urban contexts, affecting the broader applicability of findings. A notable challenge in expanding soundscape studies lies in the difficulty of automatically extracting sound sources from recordings within the dynamic and unstructured urban settings [Nog+22]. This limitation hampers the scalability of soundscape research, restricting its application across larger geographical areas. However, the diversity of sound sources in urban environments, particularly in pedestrian zones, plays a crucial role in defining the auditory landscape. These sounds, ranging from human activities to natural environmental cues, are vital for understanding and navigating urban spaces. Against this backdrop, auditory event classification emerges as a critical component of audio signal processing and machine learning, tasked with identifying and categorizing distinct sounds within an audio stream. This process is foundational to applications in environmental monitoring and urban soundscaping, beginning with the collection of high-quality audio data. By addressing the challenges in soundscape evaluation and leveraging advancements in auditory event classification, it is possible to enrich our understanding of urban environments. This extensive approach not only enhances environmental assessments but also paves the way for creating more context-aware and improved autonomous systems capable of safely interacting with the dynamic nature of urban pedestrian spaces.

*Implementation of audio event classification.* The methodology for audio event classification in the DELTA dataset leverages deep learning models such as YAMNet [Yu+20], which is pre-trained on Google's AudioSet [Gem+17]. AudioSet provides a diverse categorization of audio clips, enhancing YAMNet's ability to recognize a wide range of audio events. Additionally, for classifying environmental sounds, we selected a subset of the ESC-50 dataset [Pic15]. From its 50 categories, 13 that are most relevant to the urban environments were chosen for focused training and validation, ensuring that the model is fine-tuned to detect audio events specific to urban environments and pedestrian spaces. Building upon the methodology proposed by [VMM22], the classification process starts by normalizing the amplitude range of the audio files, which created a consistent base for analysis. The audio signals were then converted into a visual representation,

specifically spectrograms, to facilitate the classification task. Spectrograms are visual illustrations of the spectrum of frequencies of a signal as it varies with time. In this case, they provided detailed insights into the frequency distribution and energy patterns within the audio data. The YAMNet's parameters, such as sampling rate and frame hop time, were aligned with the characteristics of the audio signals to ensure a precise temporal representation. Stereo audio files were converted into mono format by averaging the channels to meet YAMNet's input standards. Following this preprocessing, YAMNet analyzed each audio segment and produced detailed spectrograms along with class activation maps. The class activation maps provided probabilistic evaluations of various audio event categories, enhancing the model's ability to classify environmental sounds. The preprocessing stage for audio event classification using YOLOv8 [Ult24] involves a series of transformations to convert the audio signal into a suitable visual representation for the model. Here is a detailed breakdown of the steps involved:

*Fourier transformation:* The process begins with a Fourier transformation, which decomposes the time-based audio signal into its constituent frequencies. This mathematical operation is critical for revealing the frequency components of the audio signal. To mitigate spectral leakage—a phenomenon that can distort the frequency spectrum—a window function is applied to the audio signal before the Fourier transformation.

*Windowing and frame segmentation:* Post-Fourier transformation, the audio signal is segmented into overlapping frames. Each frame is then processed with a Hanning window, which smoothens the signal and minimizes discontinuities at the frame boundaries. The Hanning window function is essential for preserving the signal's continuity when analyzed frame by frame.

*Spectrogram creation and logarithmic scaling:* A Fast Fourier Transform (FFT) is performed on the windowed frames to compute their frequency content. The resulting spectrogram is a two-dimensional representation of the signal's frequency and time information. Applying logarithmic scaling to the frequency axis of the spectrogram emphasizes the lower frequencies, which aligns the representation more closely with human auditory perception.

*Decibel conversion:* To enhance the spectrogram, the magnitude of the FFT is converted into decibels (dB). This conversion to a logarithmic scale allows for a more meaningful representation of the signal's dynamic range and brings the spectrogram in line with the way humans perceive sound intensity.

*Visual representation:* The transformed spectrogram is then converted into a visual image using a colormap that encodes amplitude or energy levels into color intensities. This image serves as an intuitive visual representation of the sound, illustrating the temporal evolution of its frequency components.

After these preprocessing steps, the resulting images (spectrograms) are ready to be used as input data for the YOLOv8 model. YOLOv8, primarily known for visual object detection, can then be trained on these spectrograms to classify audio events. By treating the spectrograms as images, YOLOv8 can leverage its object

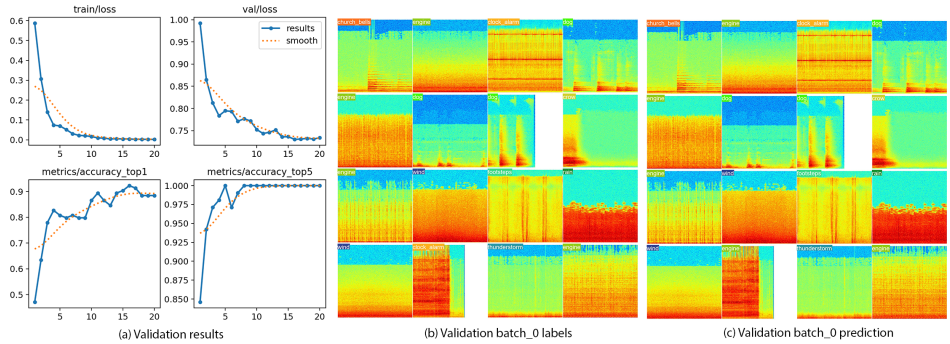



**Figure 24.** Validation results from YOLOv8 audio classification model, (**b**) spectrograms for validation audio samples with true labels versus (**c**) model predictions, depicting sound signatures for various sources.

detection capabilities to identify and categorize different sound events based on patterns within the spectrogram. Using this method, we trained a YOLOv8 model to classify various sounds based on their visual representations. The training process demonstrated a steady decrease in loss and a corresponding improvement in accuracy, achieving a top-1 accuracy of 69.2% and a top-5 accuracy of 92.3%, as shown in Figure 24(a). Additionally, Figure 24(b), (c) provides a visual comparison of actual and predicted sound classifications using spectrograms, illustrating the model's capability to distinguish sound signatures from diverse sources. The spectrograms reveal the frequency distribution and energy patterns in the audio, while the class activation maps show the probabilities of different audio events. An example of YAMNet's output for a sample audio segment is shown in Figure 25.

## 4.8. DELTA dataset

The DELTA dataset is a multimodal urban sensing resource collected using a custom-built e-scooter platform equipped with a suite of synchronized sensors. Designed to support detailed pedestrian-centric analysis, the dataset integrates spatial, visual, auditory, and motion-based data to facilitate tasks such as mapping, localization, segmentation, and environmental understanding. The dataset includes the following components:

- LiDAR Data: 15,497 LiDAR scans were captured using the Ouster OS1 sensor, providing 3D point clouds with range and reflectivity channels. These are used for spatial mapping and segmentation tasks.
- Stereo Images: 91,411 stereo image sequences were collected using the ZED2i stereocamera, including left/right RGB images and disparity maps for depth estimation and 3D reconstruction.

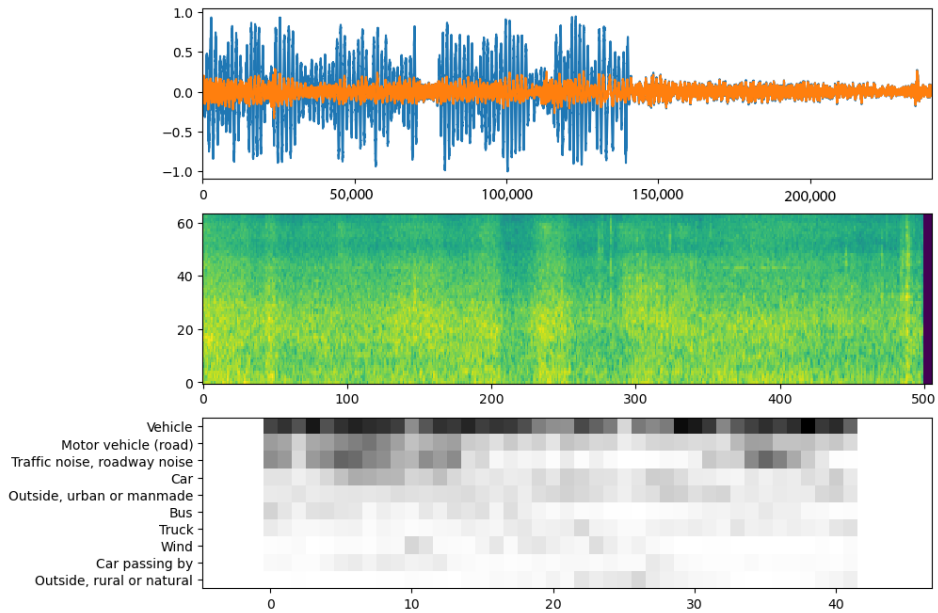



**Figure 25.** Audio analysis visualizations generated using YAMNet. The top panel shows the normalized stereo waveform of the audio chunk. The middle panel presents the corresponding spectrogram, indicating frequency content over time with color intensity representing energy levels. The bottom panel illustrates the class activation map, displaying the probabilities of various audio event classes detected across the audio timeline.

- High-Resolution 4K Images: 45,957 RGB frames were recorded using the Osmo Action 3 camera, offering high visual detail suitable for semantic analysis.
- GNSS Data: 1,598 raw GNSS points were recorded via the ZED-F9P RTK receiver, enabling accurate geo- positioning and trajectory reconstruction.
- IMU Data: Inertial measurements from the HWT901B 9-axis IMU include accelerometer, gyroscope, and magnetometer readings, supporting motion tracking and pose estimation.
- Audio Data: 25 minutes of raw ambient audio were captured using the Tascam DR-07X recorder.
- Sidewalk and Pedestrian Path Segmentation Models:
- A custom segmentation model trained on 4K action camera images, optimized for high-resolution visual inputs under diverse lighting conditions.
- A custom segmentation model trained on single-view stereo images from the ZED2i camera
- A custom segmentation model trained on the reflectivity channel of LiDAR data, effective in challenging visual conditions where traditional RGB cues may be weak or inconsistent.

- Audio Event Classification Data: The dataset includes geotagged and classified environmental audio events, generated using deep audio classification models (e.g., YAMNet and YOLOv8). Each sound event (e.g., vehicle noise, footsteps, construction) is timestamped and linked to GPS coordinates, enabling spatiotemporal analysis of the urban soundscape.

The dataset supports diverse applications in pedestrian infrastructure mapping, context-aware navigation, and multimodal AI research.

## 4.9. Discussion

The development and deployment of the DELTA platform mark a significant step toward capturing rich, multimodal data in pedestrian-centric urban environments. The platform successfully integrates a diverse range of sensors—including LiDAR, stereo and monocular cameras, GNSS, IMU, and an audio recorder—into a compact, mobile system capable of high-resolution data collection. The synchronization and calibration procedures outlined in this chapter have enabled the generation of a temporally aligned and spatially coherent dataset, suitable for tasks such as localization, segmentation, and contextual urban analysis.

Despite these contributions, several limitations in the current implementation of the DELTA platform must be acknowledged. First, while the platform offers high spatial fidelity, the collected data are inherently constrained by the hardware limitations of mobile platforms. Factors such as vibration, occlusion, and variable lighting conditions can introduce noise and inconsistency, particularly in visual and depth data. Although damping systems were integrated to mitigate vibrations, subtle sensor misalignments and drift may still affect the accuracy of downstream fusion and perception tasks.

In addition, the scope of data collection is geographically limited to the area surrounding the Delta building in Tartu, which may affect the generalizability of models trained on this dataset. Environmental diversity, such as changes in urban density, architectural style, and ambient soundscape, remains relatively constrained in the current version. The duration of audio recordings and the coverage area for pedestrian activity could also be extended in future iterations to support broader soundscape analysis.

Finally, while the dataset includes specialized models for sidewalk and pedestrian route segmentation from different sensor streams, their comparative performance under various environmental conditions is not extensively benchmarked in this chapter. Moreover, the integration of geotagged audio events—though promising—has not yet been leveraged for multimodal fusion or contextual inference.

Overall, this chapter lays the technical and methodological groundwork for multimodal urban data acquisition and sets the stage for the higher-level mapping and localization frameworks described in subsequent chapters. Addressing the current limitations through improved sensor integration, wider geographic coverage, and tighter fusion strategies will further enhance the value and applicability of the DELTA dataset in real-world scenarios.

## 4.10. Conclusion

This chapter presented the complete pipeline for handling and processing the raw multimodal data collected by the DELTA platform. It detailed the procedures for data synchronization, calibration, preprocessing, and alignment across diverse sensor modalities, including LiDAR, stereo and monocular images, audio, GNSS, and IMU data. Special attention was given to managing the challenges of cross-sensor fusion, such as temporal misalignment, noise filtering, and spatial correspondence, which are critical for ensuring consistency and reliability in downstream tasks. The chapter also introduced three distinct segmentation models developed from 4K camera images, stereo imagery, and LiDAR reflectivity data to extract pedestrian pathways and sidewalks. Additionally, a geotagged audio classification pipeline was described, providing an auditory layer for environmental context. Together, these processing steps transform raw, unstructured sensor data into structured, fused datasets suitable for semantic mapping, localization, and environmental analysis. Importantly, the dataset captured from the DELTA platform only served as a benchmark for evaluating the performance and generalizability of the street2sat and Street2GIS frameworks presented in the subsequent chapters. Furthermore, for each of the following frameworks, I have compiled a new image dataset to train the corresponding models.

# 5. URBAN MAPPING AND LOCALIZATION

Urban localization and the generation of Geographic Information System (GIS) data are foundational to modern urban planning, autonomous navigation, and mobility solutions [LT10]. GIS refers to a framework for gathering, managing, and analyzing spatial and geographic data. It integrates various data types to create detailed maps and models of urban environments. High-precision urban localization enables the accurate mapping of streets, landmarks, and infrastructure. Meanwhile, robust GIS data equips urban planners with comprehensive spatial insights, supporting the development of transportation networks, public safety initiatives, and sustainable city designs. Together, these components ensure that dynamic urban environments and emerging mobility technologies are easily integrated. The following sections presents two complementary frameworks designed to address the challenges in these domains. In the first framework, the goal was to understand whether linking street-view images with overhead perspectives could yield a richer, more context-aware environment for navigation tasks. By deploying landmark segmentation models and leveraging a generative network (street2sat), the study explored the feasibility of translating ground-level inputs into satellite views, thus creating a foundational database of spatially indexed and visually aligned tiles. This line of inquiry was not simply about achieving strong quantitative localization results; rather, it was about discovering how these different modalities interact, how features complement or overlap, and which aspects of the urban environment prove most influential in geospatial reasoning.

Building on the insights obtained from this initial exploration—such as the importance of granular pedestrian data, the complexity of accurately representing different urban features, and the challenges of maintaining reliable alignment between imagery and coordinates—the research advanced to the second framework: Street2GIS. This new framework was developed as a direct response to the lessons learned from the first approach. While the initial methodology demonstrated the potential of image synthesis and landmark-based localization, it also highlighted the need for a more robust, multimodal pipeline that could generate detailed, georeferenced shapefiles directly from imagery and positional data.

By integrating depth estimation, semantic segmentation, and advanced cross-view modeling techniques, Street2GIS aimed to produce rich, GIS-ready outputs that not only offer improved localization capabilities but also support broader urban planning and navigation applications. In essence, the first framework illuminated how generative transformations between different viewpoints could inform localization tasks, the second framework took these insights further, automating and refining the GIS data generation process to achieve a more comprehensive and precise spatial understanding thereby paving the way for more accurate, adaptive, and context-aware localization solutions.

## 5.1. street2sat: Generative AI based mapping and localization

Image-based localization plays a central role in computer vision, robotics, and navigation, enabling precise camera pose estimation for applications ranging from augmented reality to autonomous vehicles [DMR18; SLK11]. Traditional IBL methods rely on matching images to detailed 3D maps generated through techniques like LiDAR and photogrammetry, but these approaches face challenges due to high costs, storage demands, and limited scalability [LLD17; FLG14; Wan+21a]. This has led to a shift toward learning-based methods, which offer better adaptability but still struggle with scene variations and complex 2D-3D matching [Ge+20; ZYC21; Sar+19; Wan+21b; Sar+20]. Alternatives like ground-to-satellite cross-view localization reduce map size but introduce accuracy and cost concerns [Shi+20; LL19; Hu+18; Xia+22], while sensor-based approaches using LiDAR or radar provide precision at a high financial cost [Che+21; Lu+19; Ado+22; Tan+21b]. 3D maps are accurate but resource-heavy [SMT18], satellite imagery provides broader views with limited detail [ZYC21; LL19], and planimetric maps like OpenStreetMap are accessible but lack depth and environmental cues [SZC20; Seo+18].

To build a comprehensive urban map, I start by collecting georeferenced images from two perspectives: street-level and satellite views. These images, taken along paths typical of delivery robots and micromobility devices, capture the spatial details needed for mapping. Next, the collected images are processed using two specialized deep learning models. First, a custom landmark segmentation model identifies key urban features—such as buildings, statues, and bridges—in the street-level images. Simultaneously, a generative model called street2sat transforms these street-level images into corresponding satellite views, effectively bridging the gap between ground-level and aerial perspectives. After processing, all data are organized into a contextual tiled image-map. This map arranges the target area into a grid of down-sampled satellite tiles, each enriched with geocoordinates, timestamps, and the detected landmark information, forming a detailed spatial database.

For localization, when a new street-level image is received, it is first analyzed by the landmark segmentation model to detect any known landmarks. These landmarks guide a template matching process that searches the tiled map for similar landmarks. At the same time, the street2sat model generates a satellite view of the input image. By comparing this generated view with the corresponding tiles in the map, the framework identifies the best match and assigns the best latitude and longitude to the input image. Figure 26 provides a visual summary of the proposed framework.

### 5.1.1. Experimentation setup

For this specific framework, an Osmo 3 action camera in conjunction with a ZED-F9P GNSS module was utilized, as illustrated in Figure 27(a). This configuration enabled the georeferencing of images at an average interval of 3.89 meters between consecutive shots. To ensure comprehensive coverage, two thorough sweeps of the target area was conducted: one in a clockwise direction as shown in Figure 27(b) and another in a counter-clockwise direction, as shown in Figure 27(c), which covered a total distance of 3.19 kilometers. This approach was specifically designed to replicate the navigation patterns of delivery robots and micromobility devices, capturing dual perspectives from the same locations to document a wide range of visual cues encountered on sidewalks and pathways. At each capture point, three types of data were collected: RGB street-view images, geocoordinates, and timestamps

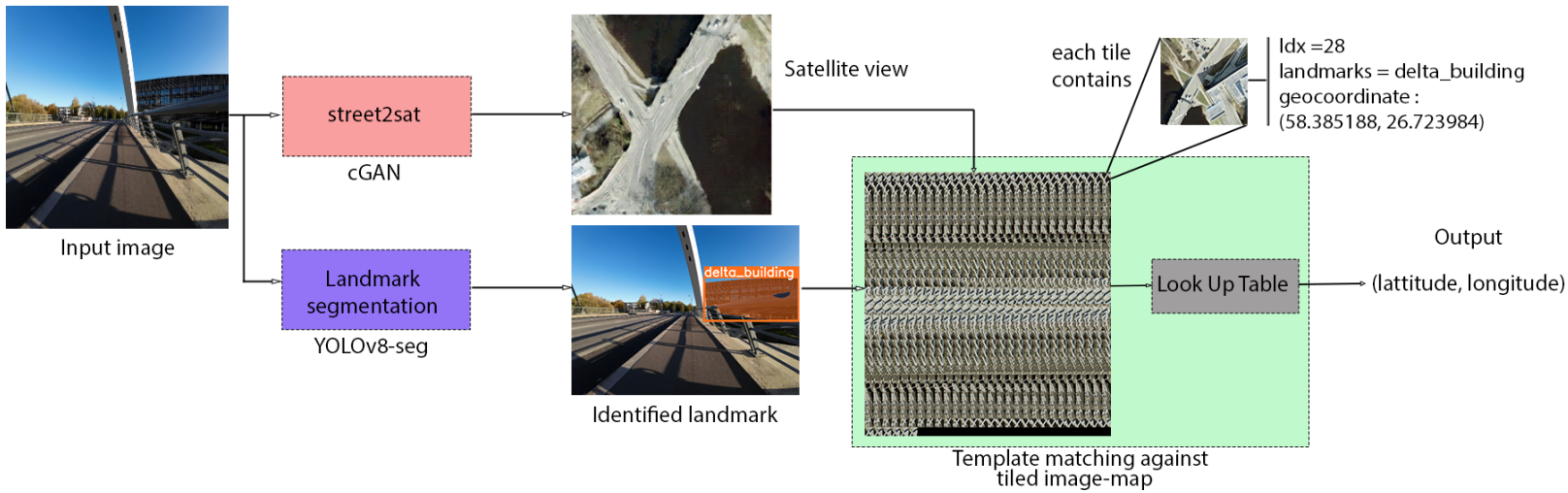



**Figure 26.** Visual representation of the proposed localization framework

### 5.1.2. Landmark segmentation

Image-based landmark segmentation is the task of identifying and delineating landmarks in images. Landmarks are typically defined as prominent and distinctive features that serve as reference points for navigation and localization. Image-based landmark segmentation is a critical step in various applications, including autonomous navigation, urban planning, and augmented reality. Traditional methods for image-based landmark segmentation have relied on hand-crafted features and rule-based algorithms. These methods are often brittle and can struggle to generalize to new datasets or environments. Deep learning-based approaches have emerged as a powerful alternative for image-based landmark segmentation. These approaches utilize CNNs to learn discriminative features from images and segment landmarks automatically. CNNs are composed of layers of convolutional filters that extract different features from images. These features are then processed through a series of pooling layers that reduce the spatial dimensions of the feature maps while preserving their most salient features.

YOLOv8 [Ult23] excels in image segmentation with its fast, accurate, and versatile design, making it ideal for use in autonomous vehicles and robotics. Its core features include the Feature Pyramid Network (FPN) for detecting and segmenting objects across various scales, enhanced by the Path Aggregation Network (PAN) for detailed image representation. The architecture is further refined with techniques like cross-stage partial convolution and bidirectional feature pyramid for efficient and comprehensive feature integration. YOLOv8 uses Focal Loss to focus training on difficult cases, improving accuracy, and employs high-resolution heads for fine detail processing. Notably, it introduces an anchor-free detection mechanism and an adaptive inference strategy, optimizing performance and computational efficiency. These advancements position YOLOv8 at the forefront of object detection technologies.

In this research, SAM [Kir+23] alongside the YOLOv8-based architecture was implemented to perform image segmentation, harnessing SAM's robust performance in zero-shot learning to reduce the typically burdensome task of manual labeling required for dataset preparation. Initially, SAM was used to annotate around 800 street-level images, focusing on 27 unique landmarks near the University of Tartu DELTA building. The use of SAM significantly accelerated the data preparation process while improving segmentation accuracy. Finally, a segmentation model based on the YOLOv8 framework was applied to the annotated images. Figure 28 showcases examples of the landmark segmentation model in operation.

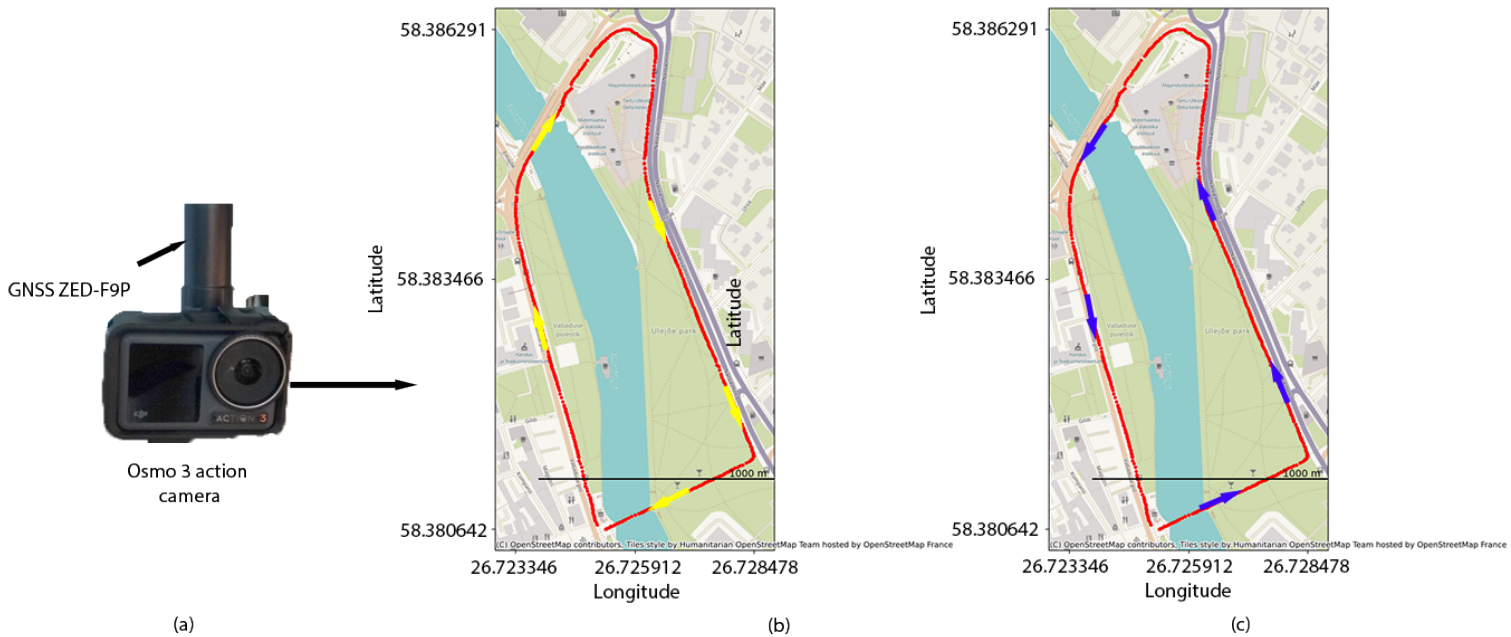



**Figure 27.** Visual depiction of the data collection process for street2sat: (a) The sensor setup used for data capture, (b) the clockwise data collection sweep, represented by yellow arrows, and (c) the counter-clockwise data collection sweep, denoted by blue arrows.

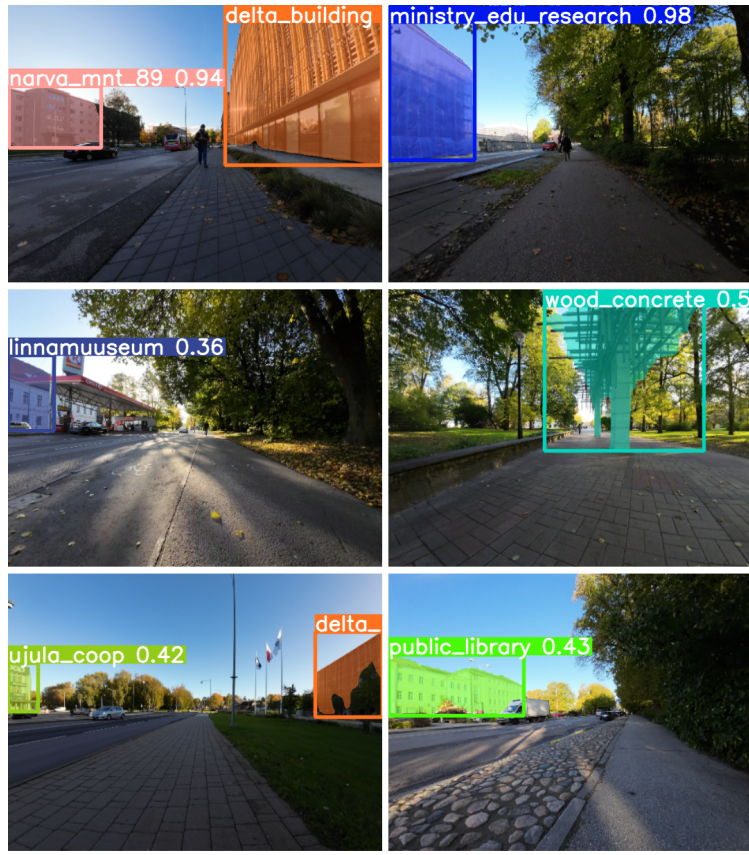



**Figure 28.** Examples of landmark segmentation on street view images in Tartu Estonia.

To assess the effectiveness of the semantic segmentation method, its precision and mean Average Precision (mAP) for both bounding boxes and masks over several epochs was evaluated. These metrics serve as essential indicators of the model's proficiency in identifying and categorizing landmarks within urban infrastructure analysis. the evaluation specifically concentrated on precision (B) and mAP50 (B) for bounding boxes, along with precision (M) and mAP50 (M) for masks. The analysis of results from the initial, median, and final epochs reveals a significant enhancement in the model's capacity for precise segmentation and landmark recognition. This improvement is detailed in Table 9, providing an overview of the segmentation quality metrics across several epochs.

### 5.1.3. The street2sat generative model

The street2sat component is a generative model specifically designed to synthesize satellite view imagery from corresponding street-level viewpoints. To achieve this, the pix2pix conditional Generative Adversarial Network (cGAN) framework [Iso+17] was implemented, which operates on paired training data comprised of

**Table 9.** Overview of the custom segmentation performance metrics across several epochs (B=Bounding boxes, M=masks).

| Metric | Initial | Median | Final |
|---|---|---|---|
| Precision (B) | 0.69 | 0.78 | 0.76 |
| mAP50 (B) | 0.43 | 0.76 | 0.79 |
| Precision (M) | 0.69 | 0.78 | 0.73 |
| mAP50 (M) | 0.42 | 0.76 | 0.75 |

**Table 10.** Training loss metrics at epoch 1000, iteration 800.

| Metric | Value |
|---|---|
| G_GAN Loss | 10.547 |
| G_L1 Loss | 26.373 |
| D_Real Loss | 0.000 |
| D_Fake Loss | 0.001 |

street-level images and their georeferenced aerial counterparts. The center of a specific square region of interest is indicated by geospatial coordinates obtained from a GNSS module, which are defined as:

$$\text{Square Area} = [(lon - d, lat - d), (lon + d, lat + d)] \tag{5.1}$$

This ensures that each street-level image is consistently matched with the corresponding satellite view of the exact location, providing precise ground truth for training. The training procedure was configured to support stable and effective model convergence. A batch size of 1 and an initial learning rate ($lr$) of 0.0002 were used. The training routine consisted of 500 initial epochs at a fixed learning rate, followed by an additional 500 epochs during which the learning rate was gradually reduced according to:

$$lr = lr_{\text{initial}} \left(1 - \frac{epoch}{n_{\text{epochs}} + n_{\text{epochs\_decay}}}\right) \tag{5.2}$$

Here, $lr_{\text{initial}}$ denotes the initial learning rate, $epoch$ is the current epoch index, $n_{\text{epochs}}$ is the number of epochs before decay, and $n_{\text{epochs\_decay}}$ specifies the duration of the decay period. All training images were uniformly resized to $256 \times 256$ pixels to ensure dimensional consistency and computational efficiency. Table 10 provides a snapshot of the training loss data for the generative model at epoch 1000 and iteration 800. The data includes metrics for generator and discriminator losses, highlighting the model's performance at this stage of training.

To address these limitations, my approach uses generative AI and custom landmark segmentation to convert street-level images into satellite-style views, improving localization by combining ground and aerial perspectives. This method enhances feature matching and spatial awareness, offering a scalable and cost-effective alternative. Furthermore, generative models can fuse diverse urban data,

enabling richer, more adaptable map representations for dynamic city environments [KGR22].

The main idea behind generating satellite images from street-level images comes down to two key benefits. First, aerial images provide a bird's-eye view that complements ground-level perspectives, making it easier to identify and align important features. This top-down perspective is especially useful in urban areas, where buildings, vehicles, and pedestrians create complex scenes with many obstacles. By combining both views, we can improve localization accuracy despite these challenges. Second, the wide field of view from an aerial image gives the model a much richer understanding of the surrounding space. This helps in matching features and making sense of the scene more effectively. By integrating both viewpoints, we create a more reliable and detailed understanding of urban environments, ultimately improving localization strategies. Figure 29 illustrates an example of the street2sat model, which generates satellite images using only street-level images. The results demonstrate how this approach preserves important environmental details and spatial relationships, leading to a more robust and context-aware localization system.

### 5.1.4. Contextual tiled image-map based mapping

The concept of the contextual tiled image-map introduces an innovative approach to urban mapping by creating an ordered sequence of generated satellite view images, each embedded with contextual data, such as its index position relative to other tiles, specific geolocation, and the identifiable landmarks within its field of view from the street-view (see Figure 30). This ordered arrangement not only offers a practical view of a vehicle's trajectory and orientation but also effectively creates a detailed mosaic that reflects the real-world environment, aiding in navigation through the urban landscape. For creating the tiled image-map a wide-angle camera paired with a ZED-F9P GNSS module (Figure 31(a)) was used to capture georeferenced images at an average interval of 3.89 meters, executing two full sweeps of the targeted region Figure 31(b) and (c). This approach, designed to mimic the navigational patterns of delivery robots or micro-mobility devices, was aimed at securing dual perspectives from identical locations to encompass the varied visual cues encountered on sidewalks and pathways. Subsequently, these images were synthesized into a large, coherent visual map, referred to as the tiled image-map Figure 31(d).

The orderly placement of these tiles serves a dual purpose: in addition to providing comprehensive spatial coverage, it encodes valuable orientation information within each tile. This encoding is made possible by leveraging data from a landmark segmentation model, which informs the system about the visible landmarks when oriented in a specific direction. Furthermore, each tile is more than just a visual representation, it is a repository of precise geographical coordinates that indicate the tile's center. The incorporation of geolocation data into the image


**Figure 29.** Examples of the street2sat model generating satellite images using only street-level input: (a) The original street-level image, (b) the corresponding satellite view from the same location sourced from OpenStreetMap, and (c) the satellite view generated by the street2sat model based solely on the street-level image.

adds an important layer of spatial context, empowering accurate localization and navigation based on the generated image-maps. Each tile in this method is generated using the street2sat image-to-image translation model. Instead of relying only on satellite or aerial images, this approach lets the model create its own representation of the map. This has two advantages: first, it prioritizes only the most relevant details for localization, rather than capturing everything in high detail. Second, it allows the model to encode scene geometry and appearance in a way that is more useful for the task, making localization more efficient. Beyond their visual representation, image-tiled maps offer significant practical advantages. They are highly scalable, enabling efficient mapping of large areas by breaking them into smaller,

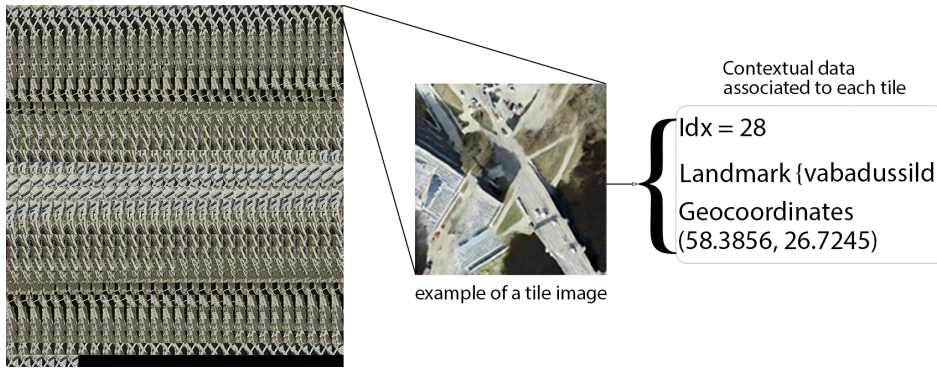



**Figure 30.** An example of the contextual data included in each tiled image-map.

manageable tiles. Their lightweight and easy-to-load structure also makes them well-suited for mobile and real-time applications, reinforcing their role in modern geospatial and navigation systems.

### 5.1.5. Template matching based localization

Image-based template matching is a widely used technique for object detection and recognition [MH21]. It involves comparing a small image fragment, known as a template, to a larger image in order to locate the location of the template in the larger image. The basic principle of image-based template matching is to compare the pixel values of the template to the corresponding pixel values in the larger image. This is typically done using a similarity metric, such as the cross-correlation [His+15a] or sum of squared differences [His+15b]. The similarity metric measures the degree to which the template and the larger image match, and a high value of the similarity metric indicates that the template is likely to be present in the larger image.

There are a variety of methods for template matching, each with its own strengths and weaknesses. Some common methods include:

***Correlation-based matching*** [ZHG06]: This is the simplest and most common method of template matching. It involves calculating the cross-correlation between the template and the reference image.

***Summed-Area Tables (SATs)*** [LSR21]: SATs are a more efficient method of template matching that can be used for large images. They store the sum of the pixels in each row and column of the reference image, which allows for faster comparisons with the template. ***Hough transform*** [Di +12]: The Hough transform is a more specialized method of template matching that is particularly well-suited for detecting objects with curved boundaries. It involves transforming the image into a parameter space where points corresponding to possible matches for the template can be identified.

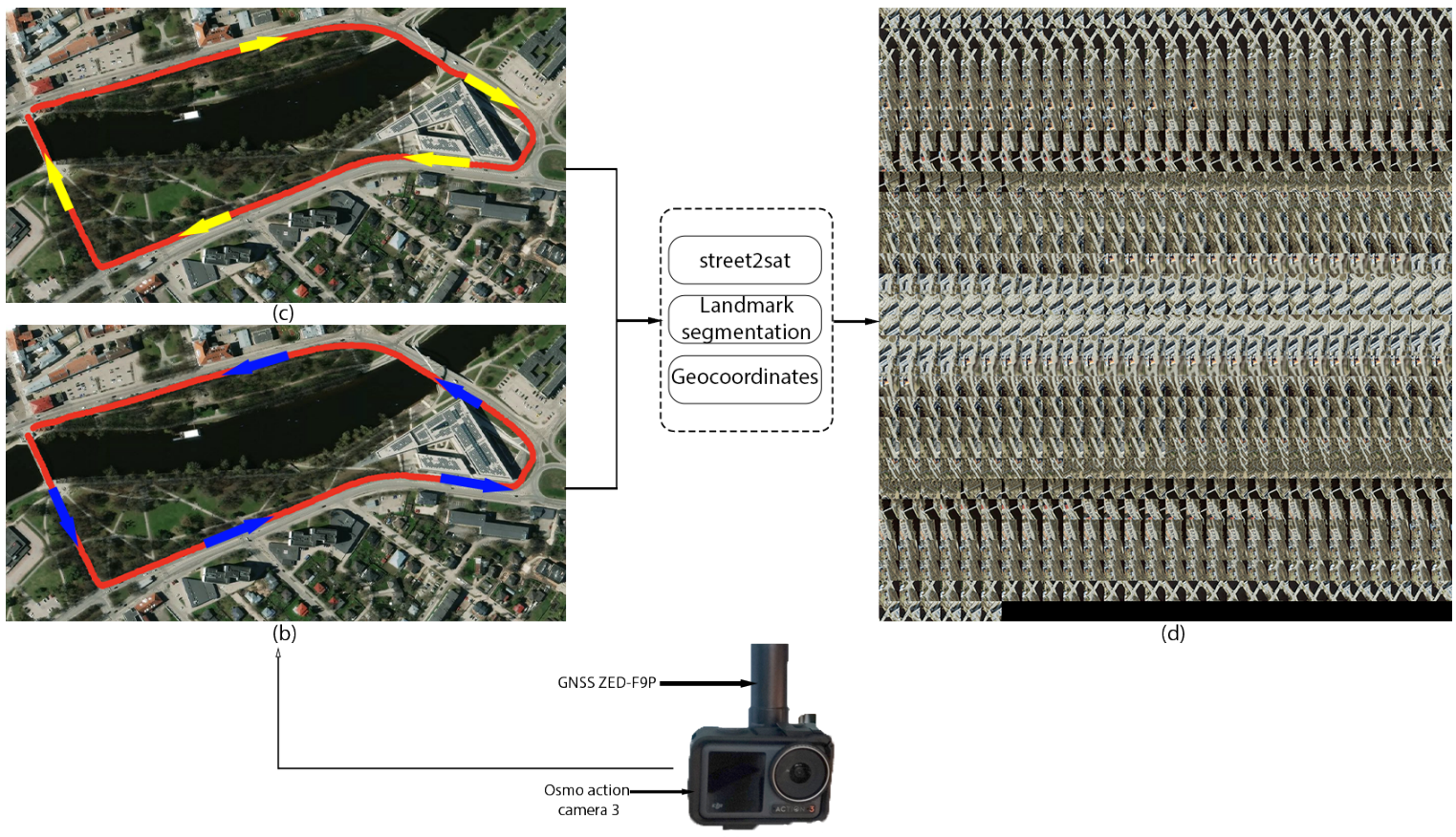



**Figure 31.** Data collection process and resulting tiled image-map: (a) sensor capturing configuration, (b) counter-clockwise data collection path indicated by blue arrows, (c) clockwise data collection path marked by yellow arrows, and (d) creation of the contextual tiled image-map database after processing through multiple pipelines.

To perform localization using a contextual tiled image-map, an image query is fed into two key processes. First, the image is processed by a landmark segmentation model that identifies recognizable landmarks. These landmarks refine the template matching algorithm by narrowing down the search area to specific tile locations within the image-map. This narrowed set of locations, represented by the set $V$, which corresponds to the valid indexes derived from the landmark data and indicates the tile locations most likely to contain the identified landmarks. Following this, the street-level input image is transformed into its corresponding satellite view using the street2sat model, $S$. The template matching algorithm employs Normalized Cross-Correlation (NCC) to systematically compare the generated satellite view $S$ against segments of a larger reference map (tiled-image map). The correlation coefficient for each segment is computed using the following equation:

$$R(x,y) = \frac{\sum_{x',y'} [S(x',y') \cdot I(x+x',y+y')]}{\sqrt{\sum_{x',y'} S(x',y')^2 \cdot \sum_{x',y'} I(x+x',y+y')^2}},$$

where $I$ represents the reference image, and $x',y'$ are coordinates relative to the segment being examined. $R(x,y)$ provides a measure of similarity between the satellite view and the reference map segment at position $R(x,y)$. The search for the best match is focused on the regions indicated by the set $V$, which comprises the tile locations linked to the recognized landmarks. Once the best match is found, its geographical location is determined by calculating the tile number, which reflects the segment's position within the reference map. This calculation is performed using the equation:

$$\text{TileNumber} = \left(\frac{y}{\text{step size}}\right) \cdot \text{row steps} + \left(\frac{x}{\text{step size}}\right) + 1.$$

The system then uses this tile number to cross-reference a lookup table that retrieves the corresponding latitude and longitude, effectively estimating the best geographical coordinates of the identified segment. By integrating both the spatial context provided by the landmark data (through the set $V$) and the visual comparison enabled by the generated satellite view, this methodology ensures a focused and efficient search within the template matching process. Figure 32 illustrates the template matching algorithm in action.

### 5.1.6. Experiment and results

To investigate the potential of the proposed localization methodology, a focused experimental setup was designed using an unseen subset (a total of 113 images) of the DELTA dataset [AVH24]. The dataset, comprising street-level images annotated with geolocations, presented a suitable testbed for examining how effectively the pipeline could bridge the gap between image-based inputs and spatial reasoning. Rather than emphasizing on specific numerical outcomes, the primary goal

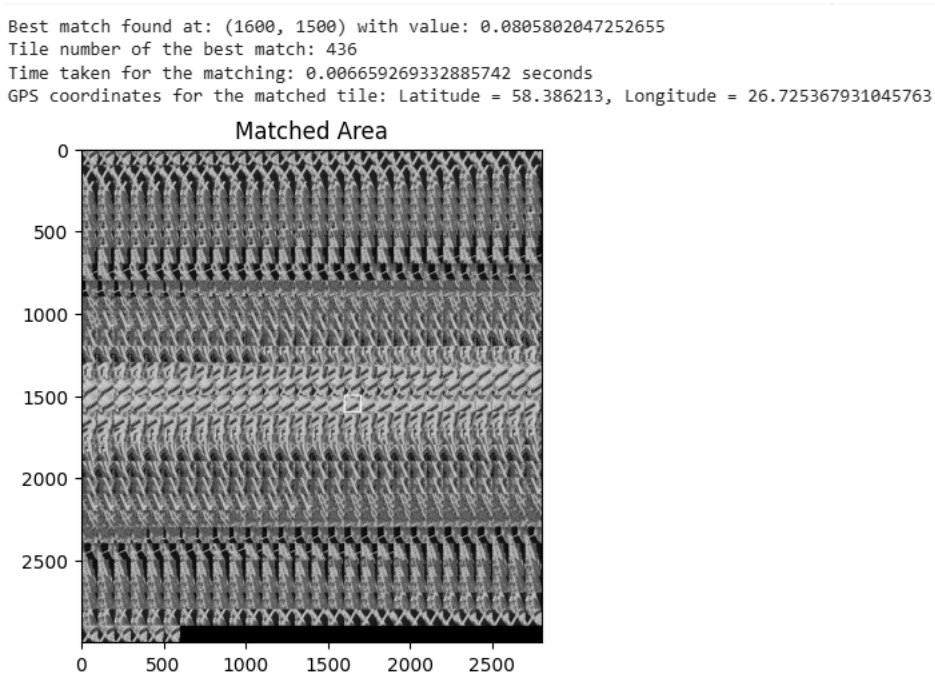



**Figure 32.** Template matching result showing the best match at image coordinates (1600, 1500) with a similarity value of 0.08. The matched area is visualized (white bounding square), along with the GPS coordinates for the matched tile.

of these experiments was to gain deeper insights into the underlying mechanisms that guide the methodology's performance. Employing the Haversine formula and supplementary statistical measures such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) allowed for a detail understanding of how closely the predictions aligned with actual geographic locations. By examining these metrics, the study aimed to uncover patterns and potential improvements in the approach, providing a richer perspective on the method's capacity to support robust geolocation tasks. Although the quantitative results are summarized in Tables 11 and 12, the emphasis remains on learning from these indicators to refine and advance the methodology.

*Haversine distance calculation.* The Haversine formula computes the shortest distance between two points on the Earth's surface, based on their latitude and longitude. This method allowed us to derive the actual distance errors for each prediction. The formula is expressed as follows:

$$\begin{aligned} d &= 2 \times 6371 \times \arcsin\left(\sqrt{a}\right), \\ a &= \sin^2\left(\frac{\Delta\text{lat}}{2}\right) + \cos(\text{lat}_1) \times \cos(\text{lat}_2) \times \sin^2\left(\frac{\Delta\text{lon}}{2}\right), \end{aligned} \tag{5.3}$$

where $\Delta$lat and $\Delta$lon are the differences in latitude and longitude, respectively, $\text{lat}_1$ and $\text{lat}_2$ are the latitudes of the true and predicted locations, and 6371 repre-

sents the Earth's radius in kilometers.

The results of the Haversine distance analysis are summarized in Table 11, indicating a mean error of 0.1175 km, a median error of 0.0739 km, and a standard deviation of 0.1219 km.

**Table 11.** Haversine distance calculation results.

| Metric | Value (km) |
|---|---|
| Mean Error | 0.1175 |
| Median Error | 0.0739 |
| Standard Deviation of Error | 0.1219 |

*MAE and RMSE.* To gain additional insight into prediction accuracy, the MAE and RMSE, which measure the average and squared average deviations, respectively, between true and predicted geocoordinates. These metrics provide complementary perspectives on prediction reliability. The formulas are as follows:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|, \tag{5.4}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}, \tag{5.5}$$

where $y_i$ and $x_i$ represent the true and predicted geocoordinate values, respectively, and $n$ is the total number of predictions. The calculated MAE and RMSE values for latitude and longitude are presented in Table 12. These results demonstrate low error margins, highlighting the practicality of the localization system.

**Table 12.** MAE and RMSE results.

| Metric | Latitude | Longitude |
|---|---|---|
| MAE | 0.00093 | 0.00070 |
| RMSE | 0.00141 | 0.00107 |

### 5.1.7. Failure analysis

In the localization prediction results, two primary sources of inaccuracies were identified. The first stems from challenges in landmark classification, influenced by several factors. In some cases, the absence of distinguishable landmarks in certain locations (Figure 33(a)) deprives the system of crucial spatial context needed for precise localization. Additionally, even when landmarks are present, they may be obscured by vegetation, urban structures, or other visual obstructions (Figures 33(b) and 33(c)), preventing effective recognition. Furthermore, system limitations such as inadequate visibility or insufficient training of the landmark detection model can result in missed identifications (Figure 33(d)).


**Figure 33.** Analysis of the system failure scenarios: (a) lack of distinguishable landmarks in the scene, (b) landmarks concealed by vegetation such as trees, (c) obstruction caused by urban structures or elements, (d) diminished visibility due to sun glare and constraints arising from inadequate model training.

The second key issue arises from inaccuracies in the satellite view generation by the street2sat model (Figure 34). When this model produces an imprecise representation of the urban environment, it disrupts the subsequent template matching process, leading to localization errors. These challenges highlight the critical need to enhance both the reliability of landmark classification and the fidelity of generated satellite views to improve the overall accuracy of the geolocation prediction system.

## 5.2. Street2GIS: An automated GIS data generation framework

The rapid evolution of autonomous navigation technologies is revolutionizing transportation systems, with a particular focus on delivery robots and micromobility services [Alv+24]. These systems depend on accurate and dynamic GIS data to navigate urban landscapes efficiently and safely [HL24]. High-resolution GIS datasets are essential for optimizing route selection, avoiding obstacles, and ensuring compliance with traffic and pedestrian regulations [Xia+24]. Despite the growing emphasis on these technologies, there remains a significant gap in the availability of detailed pedestrian infrastructure data compared to vehicular

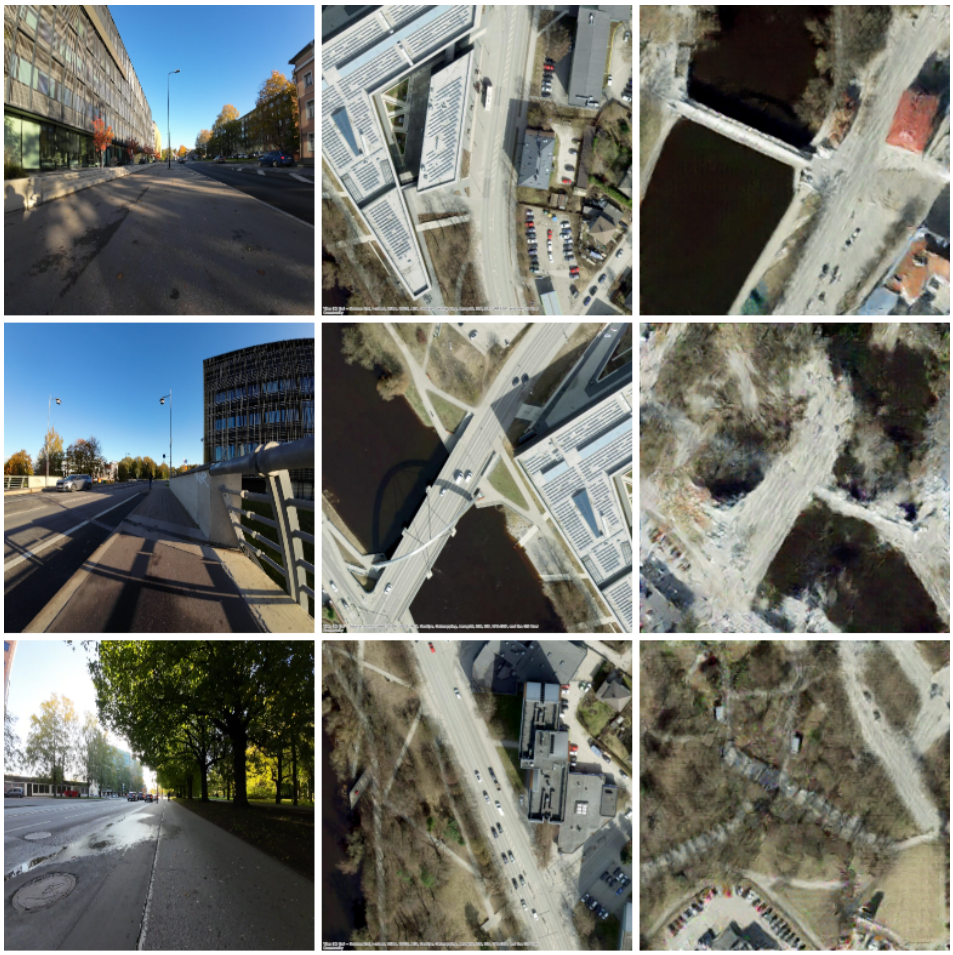



**Figure 34.** The street2sat model encountering challenges in accurately generating satellite views from street-level imagery: (a) street-level view from the DELTA dataset, (b) corresponding actual satellite view of the same location, and (c) satellite view synthesized by the street2sat model.

infrastructure as mentioned before. While road networks and vehicle pathways are well-represented in existing GIS databases, pedestrian pathways, including sidewalks and crosswalks, are often overlooked. This imbalance creates challenges for autonomous delivery robots and micromobility systems, which rely heavily on pedestrian infrastructure for operation. Research indicates that conventional methods for collecting GIS data fail to adequately capture the intricate and dynamic nature of pedestrian environments, leaving critical aspects underrepresented [LVD20].

Traditional data collection techniques, such as manual surveys and satellite imaging, are particularly ill-suited to the demands of pedestrian infrastructure mapping. These methods are labor-intensive, error-prone, and slow, resulting in

datasets that quickly become outdated. Key pedestrian-specific features, such as temporary obstructions, sidewalk conditions, and crosswalk updates, are often missed or inaccurately represented, making it difficult for autonomous systems to operate effectively. Without reliable and current GIS data, autonomous delivery robots and micromobility services face increased risks of inefficiency, compromised safety, and regulatory non-compliance [TTB21].

In urban environments with high pedestrian activity—such as city centers or university campuses—the limitations of existing GIS data are even more pronounced [Alv+24]. Frequent changes in sidewalk conditions caused by construction, events, or temporary blockages exacerbate the challenges faced by autonomous systems. These gaps in data can lead to navigation failures, such as robots encountering dead ends or violating pedestrian right-of-way rules, underscoring the urgent need for more detailed and up-to-date pedestrian infrastructure mapping [JF19].

The automation of urban feature extraction—encompassing roads, sidewalks, buildings, and vegetation—from aerial and street-level imagery has emerged as a cornerstone for advancements in urban planning, smart city development, autonomous navigation, and GIS integration. Recent advances in deep learning, particularly with CNNs and Generative Adversarial Networks (GANs), have substantially improved the precision and efficiency of extracting these features from remote sensing data [LN23]. Equally significant is the automated generation of GIS shapefiles, which capture geographic features and their attributes, enabling robust applications in urban infrastructure mapping [XK15] and autonomous systems [LN23].

Here I highlight recent innovations in deep learning and GAN-based methodologies, emphasizing their role in automating urban feature extraction and the generation of GIS shapefiles. These tools have revolutionized how high-resolution aerial and street-view imagery are utilized to create GIS-compatible datasets, directly supporting urban mobility, planning, and navigation systems. For example, deep learning models such as those by [Nin+22] demonstrate high accuracy in extracting sidewalk features, achieving precision and recall rates nearing 0.9. Similarly, [Luo+19] developed techniques to create digital sidewalk inventories from aerial images, boasting prediction accuracies of up to 92.6%, providing valuable resources for active mobility and urban accessibility initiatives. Scalable tools like Tile2Net [Hos+23] have further expanded the field by employing semantic segmentation to delineate sidewalks, crosswalks, and footpaths, producing detailed GIS shapefiles critical for pedestrian infrastructure planning and routing. In the realm of road network extraction, CNN-based methods have proven transformative. For instance, [Man19] utilized deep learning to detect and classify road segments, creating GIS-compatible shapefiles essential for urban navigation and smart city frameworks. Reviews such as [Liu+24b] underscore the effectiveness of CNNs and semi-supervised learning approaches for extracting complex road networks from remote sensing imagery.

GAN-based innovations have also pushed boundaries in urban feature mapping. The SW-GAN framework by [Che+22] integrates weakly labeled and precisely annotated data, enabling accurate road extraction in regions with limited labeled datasets. For building footprint extraction, [Li+21b] proposed a GAN-based model capable of refining segmentation outputs from aerial images into accurate polygonal shapefiles, supporting applications in urban modeling and disaster management. A notable trend is the integration of aerial and street-view data to mitigate occlusion issues and improve extraction accuracy. By combining these data sources, as demonstrated by [Nin+22], researchers have significantly enhanced the completeness and quality of urban feature maps. These advancements mark a pivotal shift toward more efficient, scalable, and detailed GIS workflows, facilitating applications across diverse domains such as urban planning, pedestrian routing, and autonomous navigation.

To enhance the efficiency of GIS data collection and updates, I have proposed a framework named Street2GIS that employs a multimodal system to automate the generation of georeferenced shapefiles from street-view imagery and GPS data. This system identifies and classifies urban features such as roads, sidewalks, buildings, and vegetation within a 70-meter radius of the site where the street-level image was provided. The framework ensures consistency by aligning its outputs with official datasets, preserving the integrity of existing urban layouts and avoiding significant alterations to network connectivity. Its efficient and scalable design offers a practical solution for generating GIS data, with the capability to operate close to near real-time. By incorporating pedestrian infrastructure data into GIS workflows, this framework aims to address challenges posed by dynamic urban environments and supports incremental improvements in urban mobility solutions. The near real-time functionality is particularly valuable in GPS-denied areas, where it can enhance the geolocalization process by relying on visual and contextual data to provide accurate spatial mapping. This capability not only improves the accuracy of GIS datasets in challenging environments but also offers a robust tool for updating critical urban data more effectively.

Street2GIS is an automated tool designed to generate accurate GIS shapefiles from street-view imagery and geospatial data, combining computer vision with geospatial processing to create detailed environmental maps for applications such as autonomous navigation, urban planning, and environmental monitoring (refer to Figure 35). It operates through a multi-step process that integrates both ground-level and aerial perspectives. The system begins with a street-view image and its geospatial coordinates, which are used to retrieve corresponding satellite imagery from Esri's World Imagery service. A monocular depth estimation model then processes the street-view image to predict depth, reconstructing a three-dimensional structure from a single two-dimensional input. This depth information enhances spatial mapping by capturing distances and scaling relationships within the scene. Using the depth map, street-view image, and satellite tile, the framework generates a semantic segmentation map from a top-down view,

identifying key urban features such as roads, sidewalks, buildings, and vegetation. From this, two essential outputs are derived: an elevation map, which represents terrain variations to ensure proper alignment of infrastructure, and an environmental feature map, which accurately delineates roads, sidewalks, and other elements. These outputs are then converted into a geo-referenced TIFF and transformed into a GIS shapefile, a widely used format for spatial analysis and integration with other geospatial datasets. The entire process is fully automated and accessible via a command-line interface, making GIS data generation faster and more efficient. Street2GIS leverages deep learning-based monocular depth estimation, aerial semantic segmentation and cross-view image synthesis to create accurate top-down maps. By transforming street-level images into rich spatial models, the framework provides a scalable, cost-effective, and accessible solution for urban mobility, infrastructure planning, and autonomous navigation.

The following sections detail the key components and evaluation methods of the Street2GIS framework. First, the data collection process is described, Next, the framework's depth estimation techniques are presented. Subsequent sections discuss the methods for semantic segmentation and environmental feature mapping achieved by integrating aerial and ground perspectives. Additional content covers the training data preparation, including the raster-to-polygon conversion process, and the techniques employed for spatial alignment and similarity assessment to validate the generated GIS data. Finally, the optimization process and evaluation metrics are presented, highlighting the framework's accuracy and reliability in modeling complex urban environments.

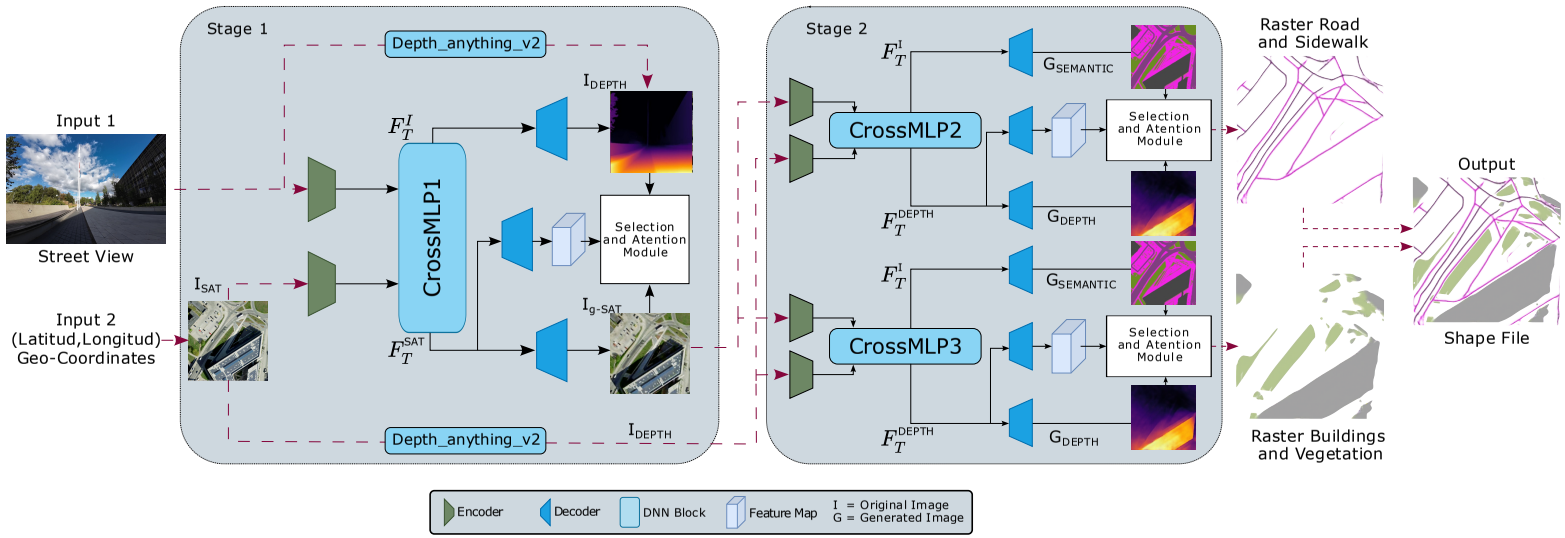



**Figure 35.** The proposed Street2GIS framework.

### 5.2.1. Experimentation setup

For my data collection process, I equipped an e-scooter with a wide-angle action camera to capture as much environmental detail and features as possible. A GNSS module was mounted on a backpack connected to a laptop for reading and recording geo-coordinates (Figure 36(a)). The camera was set to capture images at a rate of 24 frames per second, while the GNSS module recorded data at 1Hz. I was able to synchronize an image by knowing its timestamp and frequencies for both devices with their corresponding geolocation data during post-processing. The data collection spanned approximately 11.47 kilometers in Tartu, resulting in 59,185 frames and 2,466 GPS points. After filtering and processing, the dataset was refined to 2,463 images with corresponding coordinates. Each image, originally at 1920×1080 pixels, was resized to 256×256 pixels for training purposes.

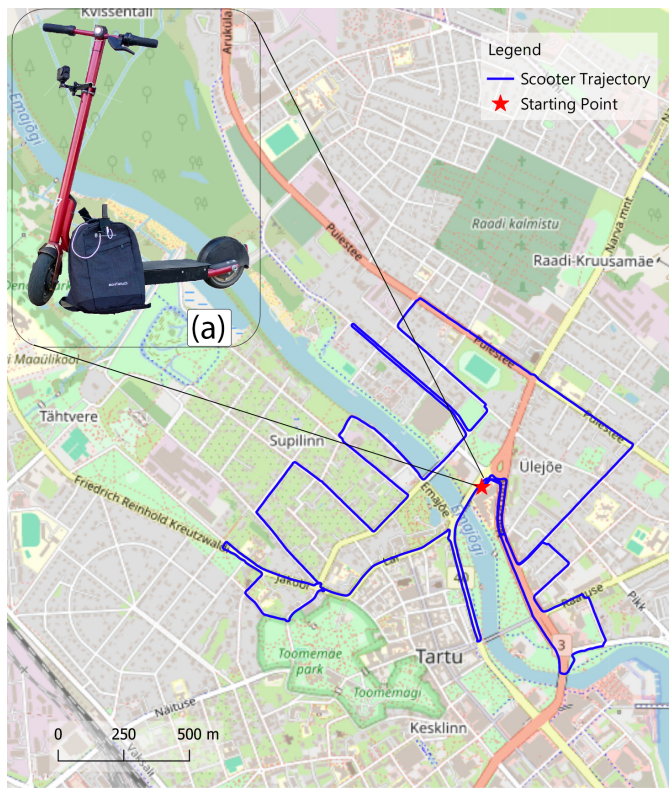



**Figure 36.** The image illustrates the geographic extent of the data collection area as represented on OpenStreetMap.

### 5.2.2. Depth estimation

The framework utilizes state-of-the-art depth estimation model to extract the spatial structure of urban environments from both ground-level and aerial perspectives, enhancing the precision of GIS outputs. For ground-level analysis, the Depth-Anything-v2 model [Yan+24] was used to generate detailed depth maps from monocular street-view images, as depicted in Figure 37(a). These depth maps provide a three-dimensional understanding of the environment, which significantly improves semantic segmentation and facilitates accurate extraction of urban features such as roads, sidewalks, buildings, and vegetation. By reconstructing the spatial relationships within the scene, the model enables more precise mapping of ground-level infrastructure.

To enhance this ground-level approach, the concept of depth estimation was extended to satellite imagery, addressing the inherent limitations of overhead images that lack explicit depth information. As shown in Figure 37(b), the Depth-Anything-v2 model is applied to extract elevation data, capturing vertical variations across both natural and man-made environments. This process produces detailed elevation maps that enhance the spatial accuracy of GIS outputs by modeling critical infrastructure features such as overpasses, bridges, terrain contours, and building heights. The integration of these depth maps into the GIS workflow ensures a comprehensive and precise representation of urban environments, bridging the gap between ground-level and aerial perspectives.

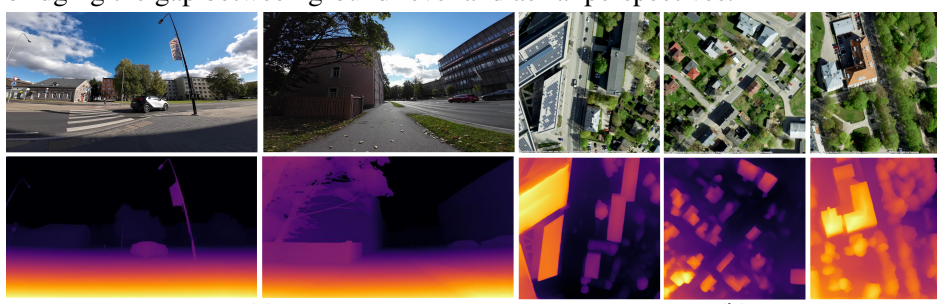



**Figure 37.** Examples of depth estimation applied to street-level and satellite images: (a) Street-level images (top row) with their corresponding depth estimations (bottom row), and (b) satellite images (top row) with their resulting elevation maps (bottom row).

### 5.2.3. Generating satellite view semantic segmentation

The semantic segmentation approach combines ground-level and aerial views to create highly accurate maps of urban environments. At the heart of this process is the CrossMLP1 architecture, an advanced model inspired by the Cascaded Cross MLPMixer framework [RTS21]. This model uses street-view RGB images alongside satellite images to develop a deeper understanding of spatial relationships. By doing so, it accurately identifies and segments urban features like streets, sidewalks, buildings, and green spaces. The integration of the detailed context

from street-level images with the extensive coverage of satellite imagery allows CrossMLP1 to capture complex spatial patterns that might otherwise be missed, ensuring that the semantic maps it produces closely mirror real-world layouts for thorough urban analysis.

To effectively train the CrossMLP1 network, I created a new dataset by pairing proprietary street-view images with their geographic coordinates, which were then used to retrieve corresponding satellite images. This dual-view dataset allowed the network to learn how ground-level and aerial perspectives relate to each other. To ensure accuracy and proper alignment, a semi-automated process was used. First, satellite images were processed using the K-means algorithm to classify key features into broad categories such as roads, sidewalks, buildings, and vegetation. These initial classifications were then refined by overlaying high-precision geospatial vectors from the Estonian Topographic Database [Lan20], improving the dataset's overall quality and detail. Attributes such as street widths, building polygons, and green space boundaries were defined based on the database's metadata. Sidewalks were defined as spaces without vegetation, roads, and buildings, ensuring their clear separation from other urban features. Each semantic layer was cross-referenced and verified for spatial accuracy to reflect real-world dimensions and arrangements. The data preparation resulted in high-quality annotations for training the CrossMLP1 network, allowing it to generate highly reliable semantic maps. By combining street-level and aerial views, the approach ensures accurate feature mapping while offering a scalable way to map complex urban areas. This framework can support applications in urban planning, GIS, and autonomous systems, providing a practical tool for analyzing and modeling urban spaces.

### 5.2.4. Generating road, sidewalk, building and vegetation

The approach for generating environmental feature rasters for urban elements such as roads, sidewalks, buildings, and vegetation relies on two enhanced CrossMLP architectures, CrossMLP2 and CrossMLP3. These models are specifically designed to capture the complex relationships between satellite semantic segmentation, RGB imagery, and elevation data, enabling detailed spatial mapping without the need for extensive manual annotation.

To simplify the annotation process, existing GIS datasets were used as a foundation for generating large-scale labels. By utilizing the bounding box of each satellite tile, corresponding urban feature polygons from the Estonian Topographic Database were identified and retrieved. These polygons were rasterized into annotated image tiles, with each class—such as sidewalks, roads, buildings, and vegetation distinctly labeled. This automated approach ensured accurate ground truth annotations for training the network. The annotated dataset provided a complete representation of geographic features, allowing the network to learn the spatial distribution and structural characteristics of each class. This facilitated the model's ability to produce highly detailed and precise spatial representations.

To further enhance the model's spatial understanding, elevation data was integrated into the training dataset. This was achieved by applying a depth estimation model to the satellite RGB imagery, generating clear height maps that depicted variations in terrain and infrastructure elevation. The inclusion of these elevation maps allowed the model to account for vertical changes in the environment, improving its ability to place roads, buildings, and vegetation accurately in spatial representations. By incorporating this additional layer of data, the model gained a better understanding of how height influences urban layouts, leading to more accurate and realistic outputs. Figure 38 illustrates the different types training data used for each CrossMLP block.

### 5.2.5. Training data description

To train each CrossMLP block, I employed a modified Pix2Pix framework with a U-Net generator [Iso+17]. The training process utilized the Adam optimizer with a learning rate of $2 \times 10^{-4}$ and a batch size of 8. The model was trained for a total of 30 epochs, split into two phases: 15 epochs with a fixed learning rate, followed by 15 epochs with linear decay. Additionally, an early stopping condition was implemented, monitoring the loss function for improvements every five epochs.

The loss functions used during training included:

- **Adversarial Loss:** Encouraging realistic output generation.
- **L1 Reconstruction Loss:** Weighted by 100 to enforce structural similarity between predictions and targets.
- **Segmentation Loss:** Weighted by 1 to prioritize accurate classification of GIS features.

The dataset was divided into training and validation sets, with 70% of the data (1,725 images) allocated for training and 30% (738 images) for validation. The training process utilized three NVIDIA GeForce RTX 3060 GPUs, each equipped with 12 GB of memory, and required a total of 8 GB of memory for each model. On average, training each epoch took 3.5 minutes. The same hyperparameters and methodology were applied to train all CrossMLP models, ensuring consistency across experiments.

This training setup, using diverse input modalities, allowed CrossMLP models to learn robust spatial relationships. By combining depth and semantic cues from street-view and satellite imagery, the framework improved GIS feature prediction, generating raster representations of urban elements like roads, sidewalks, buildings, and vegetation. The next step is to convert these pixel-based outputs into vector polygons.

### 5.2.6. Raster to polygon conversion

Raster and vector data formats in GIS serve complementary roles: raster data, such as TIFF images, provides pixel-based representations of spatial features,

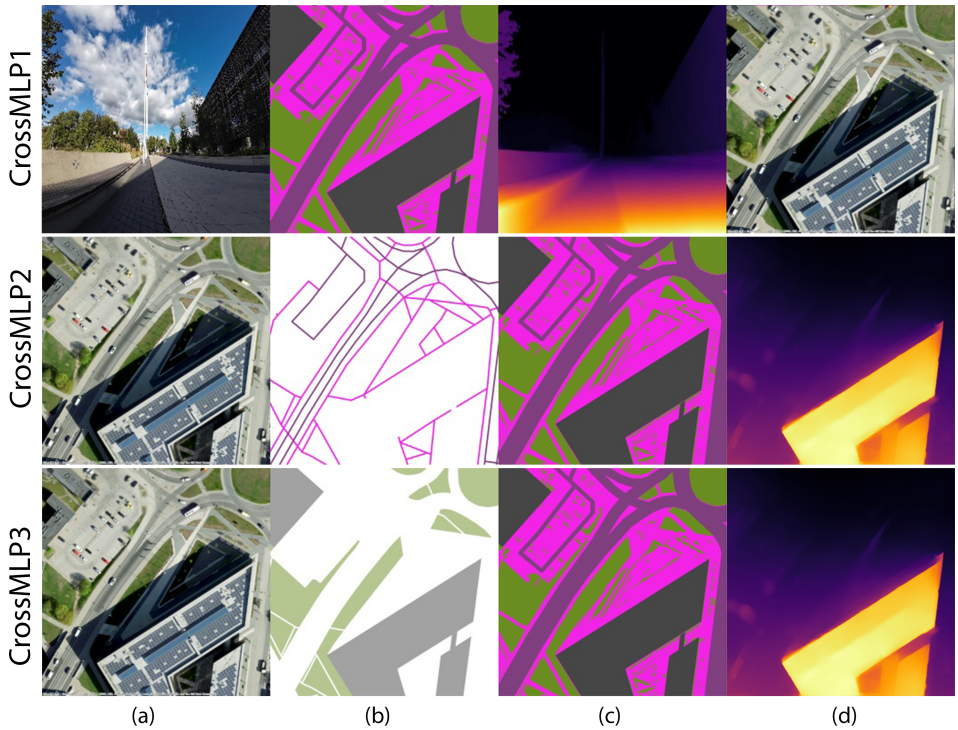



**Figure 38.** Inputs, targets, and clues used in training the CrossMLP models are illustrated across three rows: First row: (a) Input street-view RGB image, (b) target semantic segmentation map, (c) depth representation clue for the street-view, and (d) satellite RGB image clue. Second row: (a) Input satellite RGB image, (b) target road and sidewalk network, (c) semantic satellite representation clue, and (d) satellite depth representation clue. Third row: (a) Input satellite RGB image, (b) target building and vegetation placements, (c) semantic satellite representation clue, and (d) satellite depth representation clue.

while vector data, like shapefiles, uses geometric shapes—points, lines, and polygons—to define spatial elements more precisely. Converting raster images into vector polygons is a critical process that bridges these formats, enabling scalable and detailed spatial analysis.

The conversion process begins with raster image preprocessing to enhance key features while minimizing noise. Contrast adjustments are applied to emphasize essential elements, making them more distinguishable from the background. A Gaussian blur is then used to smooth high-frequency noise, creating cleaner inputs for segmentation. The preprocessed image is thresholded into a binary format, segmenting the desired features by separating foreground elements from the background based on pixel intensity. To improve the integrity of these features, morphological operations, such as gap filling and edge smoothing, are applied, ensuring continuity in the binary representation.

Once the binary raster image is refined, it is transformed into vector format. This vectorization step involves tracing the boundaries of connected pixel regions

to generate polygonal shapes that represent spatial features. These polygons capture the structure and layout of the original raster data in a geometric format suitable for GIS applications. After vectorization, additional refinement ensures the usability and accuracy of the resulting polygons. Geometric simplification reduces the complexity of the shapes by smoothing unnecessary details, making the dataset more manageable for analysis. Smaller, less significant polygons are filtered out based on their area, retaining only the most relevant features in the final output. This systematic process of raster-to-polygon conversion enables the transformation of pixel-based spatial data into detailed, scalable vector representations, facilitating more advanced spatial analysis and integration with other geospatial datasets.

### 5.2.7. Alignment and similarity measurement method

To validate the accuracy of the Street2GIS framework in generating GIS data, I developed and implemented a method for spatial alignment and similarity assessment between the generated GIS outputs and reference datasets. This approach combines feature-based alignment with structural similarity measurement to ensure precise evaluation of the framework's performance.

*Alignment method.* Accurate spatial alignment is a critical component of validating GIS data. To evaluate how effectively the Street2GIS framework captures spatial features, I implemented a multi-step spatial registration process to align generated images ($I_g$) with their corresponding reference images ($I_r$).

**Keypoint detection and matching**: I employed the Oriented FAST and Rotated BRIEF (ORB) algorithm [Rub+11] to detect and describe keypoints in both $I_g$ and $I_r$ due to its computational efficiency and robustness to rotation and scale changes. The algorithm combines the FAST keypoint detector with the BRIEF descriptor, adding orientation and multi-scale capabilities to enhance its performance in diverse scenarios. Keypoint descriptors generated by ORB were matched using a brute-force matcher with Hamming distance as the similarity metric. To ensure high-quality matches and minimize false positives, I applied Lowe's ratio test, retaining only matches that satisfy the condition:

$$\frac{\text{distance}(\boldsymbol{d}_i, \boldsymbol{d}_j)}{\text{distance}(\boldsymbol{d}_i, \boldsymbol{d}_{j'})} < \tau,$$

where $\tau = 0.75$ is the ratio threshold. This step ensured that only reliable matches were used for further processing.

**Homography estimation**: Using the filtered matches, I estimated a homography matrix $\mathbf{H}$ with the RANSAC algorithm [FB81]. The homography matrix maps points from the generated image to the reference image, accounting for projective transformations. The relationship is expressed as:

$$\begin{bmatrix} x_r \\ y_r \\ 1 \end{bmatrix} \sim \mathbf{H} \begin{bmatrix} x_g \\ y_g \\ 1 \end{bmatrix},$$

where $(x_g, y_g)$ and $(x_r, y_r)$ represent corresponding coordinates in the generated and reference images, respectively. RANSAC iteratively selects subsets of matches to compute $\mathbf{H}$, mitigating the influence of outliers and ensuring robust transformation.

**Warping and refinement**: With the estimated homography matrix $\mathbf{H}$, I warped $I_g$ to align it with $I_r$, correcting for perspective distortions and spatial misalignments. The warping process applied the transformation to each pixel in $I_g$, producing an aligned image $I'_g$:

$$I'_g = \text{warp}(I_g, \mathbf{H}).$$

To achieve sub-pixel alignment accuracy, I further refined the registration using the enhanced correlation coefficient algorithm [EP08]. This algorithm optimized warp parameters $\mathbf{W}$ by maximizing the correlation coefficient $\rho$ between $I'_g$ and $I_r$:

$$\rho = \frac{\sum_{x,y} \left(I_r(x,y) - \mu_r\right) \left(I'_g(x,y;\mathbf{W}) - \mu_g\right)}{\sqrt{\sum_{x,y} \left(I_r(x,y) - \mu_r\right)^2 \sum_{x,y} \left(I'_g(x,y;\mathbf{W}) - \mu_g\right)^2}},$$

where $\mu_r$ and $\mu_g$ are the mean intensities of $I_r$ and $I'_g$, respectively. The ECC optimization process ensured precise alignment, even at sub-pixel levels.

By integrating ORB for keypoint detection, RANSAC for homography estimation, and ECC for alignment refinement, I developed a robust methodology to align generated GIS images with ground truth data. This alignment process ensures accurate registration, enabling precise spatial feature comparisons and validating the effectiveness of the Street2GIS framework.

*Similarity measurement.* To evaluate the similarity between the aligned images produced by the Street2GIS framework and the reference images, I implemented a method using the Structural Similarity Index (SSIM) as described in [Wan+04]. SSIM provides a quantitative measure of visual similarity by analyzing brightness, contrast, and structural patterns, producing a score between 0 and 1, where 1 indicates perfect similarity.

The process begins by converting both the generated image ($I_g$) and the reference image ($I_r$) to grayscale to simplify the comparison and focus exclusively on structural content without the influence of color. Let the grayscale images be denoted as $I'_g$ and $I'_r$, respectively. To ensure a fair comparison, both grayscale images are resized to identical dimensions if necessary. Specifically, after alignment, the reference image ($I'_r$) is resized to match the dimensions of the generated image ($I'_g$).

The SSIM index is then computed between $I'_g$ and $I'_r$ using a Gaussian weighting function to account for local variations in luminance and contrast. The SSIM index is calculated as follows:

$$\text{SSIM}(I'_g, I'_r) = \frac{(2\mu_{I'_g}\mu_{I'_r} + C_1)(2\sigma_{I'_g I'_r} + C_2)}{(\mu_{I'_g}^2 + \mu_{I'_r}^2 + C_1)(\sigma_{I'_g}^2 + \sigma_{I'_r}^2 + C_2)},$$

where:

- $\mu_{I'_g}$ and $\mu_{I'_r}$ are the mean intensities of $I'_g$ and $I'_r$, computed using a Gaussian kernel.
- $\sigma_{I'_g}^2$ and $\sigma_{I'_r}^2$ are the variances of $I'_g$ and $I'_r$, representing local contrast.
- $\sigma_{I'_g I'_r}$ is the covariance between $I'_g$ and $I'_r$, reflecting their structural similarity.
- $C_1$ and $C_2$ are constants to stabilize the division when the denominators are close to zero, defined as $C_1 = (K_1 L)^2$ and $C_2 = (K_2 L)^2$, where $L$ is the dynamic range of pixel values (e.g., 255 for 8-bit images), and $K_1$ and $K_2$ are small constants (e.g., $K_1 = 0.01$, $K_2 = 0.03$).

In addition to computing the SSIM index, an SSIM map is generated, which provides a pixel-wise similarity measure across the entire image. This map visually highlights regions of dissimilarity, offering valuable insights into areas where the generated data deviates from the reference. Such visual analysis is important for identifying specific aspects of the model's performance that may require improvement, thereby aiding in iterative refinement of the framework. By employing SSIM and generating detailed similarity maps, I established a robust method for quantitatively and visually assessing the alignment and similarity of GIS outputs, ensuring the validity and reliability of the Street2GIS framework.

*Optimization.* To optimize the alignment and similarity between the generated images and the reference images, a systematic parameter optimization process was implemented. This focused on fine-tuning critical parameters during feature detection, matching, and homography estimation stages to achieve the highest possible SSIM score. The primary parameters optimized included the number of ORB features ($N_{\text{ORB}}$), the match ratio threshold ($\tau$) in Lowe's ratio test, and the RANSAC reprojection threshold ($\varepsilon$).

*Number of ORB features ($N_{ORB}$).* The ORB algorithm was utilized to detect and describe keypoints in both the generated image ($I_g$) and the reference image ($I_r$). The parameter $N_{\text{ORB}}$ determines the maximum number of keypoints extracted from each image. Adjusting $N_{\text{ORB}}$ impacts the richness of feature representation:

- **Increasing $N_{\text{ORB}}$** improves the likelihood of finding accurate correspondences between images, enhancing homography estimation and alignment accuracy. However, it increases computational complexity due to the greater number of matches evaluated.
- **Decreasing $N_{\text{ORB}}$** reduces computational load but risks insufficient keypoints, potentially degrading alignment quality.

*Match ratio threshold ($\tau$).* Lowe's ratio test was applied during the matching stage to filter ambiguous matches. For each keypoint descriptor in $I_g$, the two nearest neighbors in $I_r$ were identified based on Hamming distance. The match ratio threshold $\tau$ determined the strictness of the filtering:

- A lower $\tau$ (e.g., $\tau = 0.6$) ensures stricter matching, reducing false positives but potentially discarding valid matches.
- A higher $\tau$ (e.g., $\tau = 0.8$) allows more matches to pass, increasing potential inliers but also introducing more outliers.

*RANSAC reprojection threshold ($\varepsilon$).* The RANSAC algorithm was employed to estimate the homography matrix $\mathbf{H}$ by selecting inlier matches. The reprojection threshold $\varepsilon$ defined the maximum allowable distance between observed and projected points for a match to qualify as an inlier:

- A smaller $\varepsilon$ (e.g., $\varepsilon = 1.0$) enforces stricter criteria, improving alignment precision but risking insufficient inliers due to noise.
- A larger $\varepsilon$ (e.g., $\varepsilon = 5.0$) relaxes the criteria, increasing inlier count but potentially introducing more outliers.

*Iterative optimization process.* To determine the optimal parameters, I conducted an iterative grid search over combinations of $\Theta = \{N_{\text{ORB}}, \tau, \varepsilon\}$. The optimization process included:

1. **Defining parameter ranges:** Reasonable ranges for each parameter were established based on prior knowledge and testing:
   - $N_{\text{ORB}}$: Values from 500 to 2000.
   - $\tau$: Values between 0.6 and 0.8.
   - $\varepsilon$: Values from 1.0 to 5.0 pixels.
2. **Executing grid search:** All parameter combinations were systematically evaluated by performing the alignment process for each set.
3. **Calculating SSIM:** For each combination, the SSIM index between the aligned image ($I'_g$) and the reference image ($I'_r$) was computed.
4. **Selecting optimal parameters:** The parameter set $\Theta^*$ yielding the highest SSIM score was identified:

$$\Theta^* = \arg\max_{\Theta} \text{SSIM}(I'_g, I'_r; \Theta).$$

5. **Validating robustness:** The optimal parameters were validated on a separate dataset to ensure generalizability and robustness.

*Enhancements achieved.* The optimization process significantly improved the quality of alignment and similarity. Adjusting $N_{\text{ORB}}$ balanced keypoint richness and computational efficiency. Fine-tuning $\tau$ improved the reliability of matches, while calibrating $\varepsilon$ enhanced the RANSAC algorithm's robustness against outliers. These adjustments led to higher SSIM scores, reflecting improved structural

similarity and visual fidelity between the aligned images. Ultimately, this optimized pipeline ensured accurate and reliable validation of the Street2GIS framework.

### 5.2.8. Evaluation results

This section evaluates the performance of the Street2GIS framework across several key dimensions: computational efficiency, semantic segmentation accuracy, alignment and similarity metrics, and robustness.

*Computational efficiency.* The computational performance of the Street2GIS framework was assessed by analyzing the number of parameters and execution times for its key components. Table 13 summarizes these results.

**Table 13.** Number of parameters and execution time per block.

| **Blocks** | **Parameters ($N_{ro}$)** | **Execution Time (s)** |
|---|---|---|
| Monocular Depth | - | 0.647 |
| CrossMLP1 | 215,951,000 | 0.063 |
| CrossMLP2 & 3 | 403,823,000 | 0.063 |
| Others | - | 0.983 |

The monocular depth estimation module processed images in 0.647 seconds. CrossMLP1, the first Cross MLPMixer network with approximately 216 million parameters, completed processing in 0.063 seconds. Remarkably, the combined CrossMLP2 and CrossMLP3 networks, despite having almost double the parameters, executed within the same timeframe of 0.063 seconds. This efficiency is attributed to the architectural design of the combined CrossMLP2 and CrossMLP3 modules, which effectively integrates the two networks. CrossMLP2 specializes in roads and sidewalks, while CrossMLP3 focuses on buildings and vegetation. By treating these features as distinct classes, the framework ensures specialized processing without significant computational overhead. Parallel processing and optimized resource allocation further enhance efficiency.

The total pipeline execution time is approximately 1.75 seconds, making the framework capable of near real-time processing. For instance, the framework is optimized for pedestrian environments, where movement is relatively slow, allowing one frame to be processed every two seconds while incorporating GIS features within a 70-meter radius. This performance highlights the framework's practicality for dynamic GIS data updates, ensuring timely and efficient processing for real-world applications.

*Semantic segmentation performance.* The semantic segmentation performance of the Street2GIS framework was evaluated using precision, recall, F1-score, and support percentage for each class. Table 14 summarizes the results for the four semantic classes: Sidewalk, Roads, Buildings, and Vegetation.

The results demonstrate high precision and recall across all classes, with an overall F1-score exceeding 90%. The **Sidewalk** class achieved the highest F1-

**Table 14.** Performance comparison of semantic classes in CrossMLP1 during training.

| Class | Precision | Recall | F1-Score | Support (%) |
|---|---|---|---|---|
| Sidewalk | 90.4% | 92.1% | 93.7% | 41.2% |
| Roads | 89.3% | 90.0% | 89.6% | 17.6% |
| Buildings | 92.1% | 94.5% | 93.3% | 14.7% |
| Vegetation | 85.7% | 88.2% | 86.9% | 26.5% |
| **Overall Average** | **90.2%** | **89.7%** | **90.2%** | **100%** |

score of 93.7%, reflecting the model's strong capability in accurately identifying sidewalks. Similarly, **Buildings** performed robustly, with precision and recall exceeding 92%. The **Vegetation** class exhibited slightly lower performance, with an F1-score of 86.9%. This slight reduction may be attributed to the complexity of urban green spaces, potential occlusions in the imagery, or limited representation in street-view compared to satellite imagery. Nevertheless, the performance remains reliable for practical applications. These metrics highlight the efficacy of the semantic segmentation component within the Street2GIS framework, ensuring accurate and detailed spatial representations suitable for dynamic urban GIS applications.

**Table 15.** Performance comparison of semantic classes in CrossMLP2 and CrossMLP3. (R-S) represents roads and sidewalks, while (B-V) represents buildings and vegetation.

| Class | Alignment | Similarity | Best | Worst |
|---|---|---|---|---|
| CrossMLP2 (R-S) | 89% | 79.3% | 84.6% | 53.5% |
| CrossMLP3 (B-V) | 93% | 81.5% | 88.6% | 55.6% |
| **Overall Avg** | 91% | 80.4% | – | – |

***Alignment and similarity evaluation:*** In evaluating the outputs of the Street2GIS framework, pixel-wise comparison alone proved insufficient due to inherent zoom effects and spatial distortions in the generated data. As the framework focuses on predicting GIS elements, such simplistic evaluation methods fail to capture its full capabilities. To address this, I implemented alignment and similarity metrics to assess the spatial accuracy and visual consistency of the outputs. Table 15 presents the alignment and similarity metrics for CrossMLP2 and CrossMLP3. Alignment measures how well the spatial positioning of predicted features matches the ground truth, while similarity evaluates visual resemblance. High alignment scores indicate that the model effectively captures the spatial configuration of urban features, whereas slightly lower similarity scores highlight challenges related to lighting, occlusions, and perspective distortions.

Figure 39 illustrates example outputs generated by the proposed framework for extracting and mapping urban GIS components, specifically roads, sidewalks, buildings, and vegetation. The top row demonstrates the precision of the model in capturing road and sidewalk connections, closely matching the ground truth data. The bottom row highlights the framework's capability to accurately segment and

classify buildings and vegetation areas. The rightmost column provides a clear visualization by overlaying predicted outputs onto satellite imagery, showing that the framework maintains high geometric fidelity even in complex urban settings.

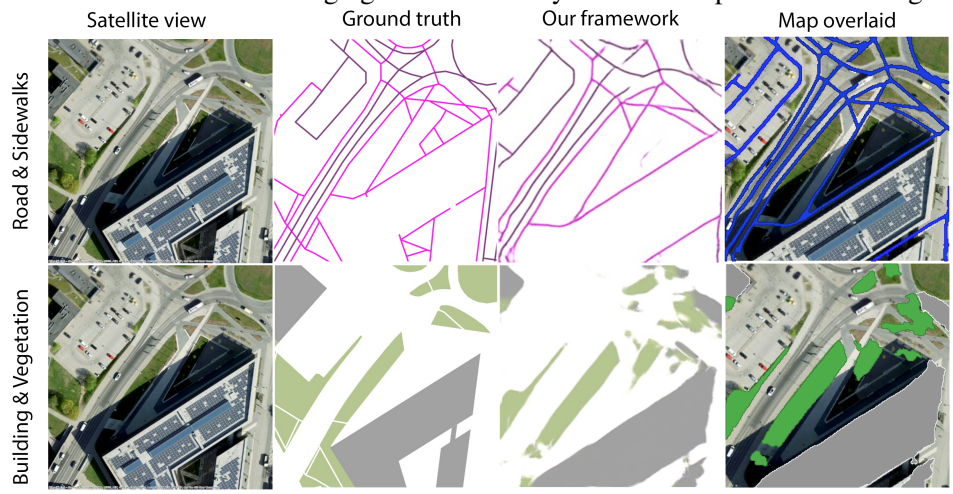



**Figure 39.** Examples of roads, sidewalks, buildings, and vegetation extracted by the framework, shown alongside their ground truth data. The last column displays overlaid shapefiles on satellite imagery for clarity.

To assess the impact of incorporating street-view imagery, I compared two scenarios: one using only satellite images and another integrating both satellite and street-view data. Figure 40 visualizes these scenarios, highlighting the improved spatial accuracy and feature distinction achieved with street-view integration. The inclusion of street-view data enhanced the model's ability to capture intricate urban features and address ambiguities present in satellite imagery.

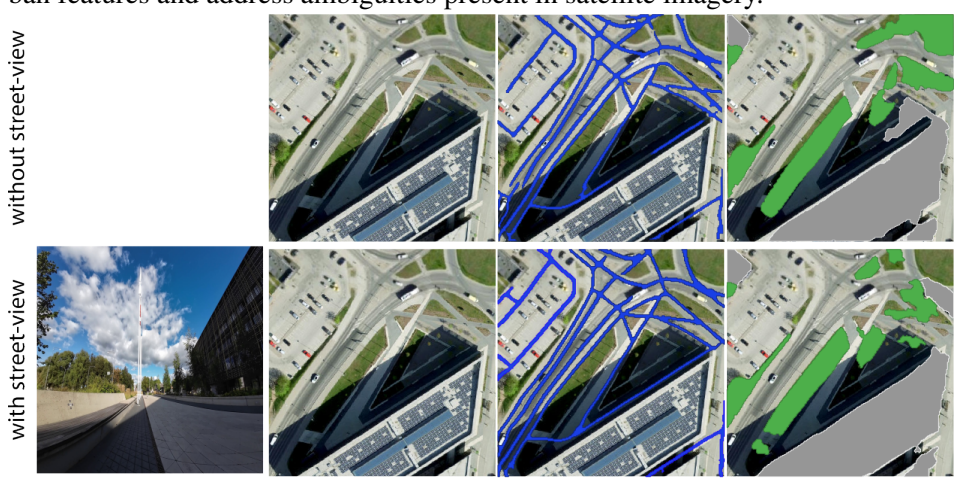



**Figure 40.** Effectiveness of the framework with and without street-view data.

To further validate the robustness of the framework, I evaluated its performance on the DELTA dataset [AVH24], which includes urban environments cap-

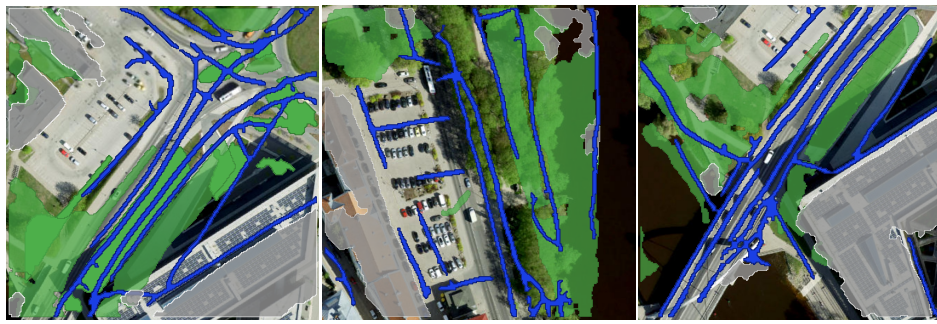


**Figure 41.** Examples of applying the framework to the DELTA dataset.

tured during different season. As shown in Figure 41, the framework effectively generated road and sidewalk networks but encountered challenges with building and vegetation placement due to seasonal variations. Table 16 compares the alignment and similarity metrics across both scenarios (with and without street-view data), demonstrating the advantage of street-view integration in minimizing temporal and seasonal inconsistencies.

**Table 16.** Performance comparison of shapefile classes: Roads and Sidewalks (R-S), and Buildings and Vegetation (B-V), with and without street-view data.

| Class | Alignment | Similarity | Best | Worst |
|---|---|---|---|---|
| With SV (R-S) | 85% | 78.9% | 83.1% | 54.7% |
| Without SV (R-S) | 81% | 72.3% | 78.2% | 46.3% |
| With SV (B-V) | 79% | 77.2% | 80.4% | 51.5% |
| Without SV (B-V) | 76% | 71.6% | 75.2% | 42.1% |

The evaluation demonstrates that integrating street-view imagery substantially enhances alignment and similarity metrics, especially for roads and sidewalks, while mitigating challenges posed by seasonal variations, particularly in the representation of buildings and vegetation. The framework effectively captures spatial relationships, enabling reliable and dynamic GIS data updates, showcasing its practicality for real-world urban applications.

## 5.3. Discussion

While the proposed street2sat and Street2GIS frameworks demonstrate promising capabilities in bridging ground-level perception with geospatial intelligence, several methodological limitations should be acknowledged to contextualize their performance and guide future improvements.

The core limitation of street2sat lies in the generalizability of the generative model across diverse urban morphologies. The framework relies on the accuracy of landmark segmentation and the ability of the generative network to translate dynamic street-level inputs into coherent satellite-like outputs. However, variations in lighting, occlusions from vegetation or urban elements, and architectural

diversity across cities pose challenges that the current model struggles to consistently overcome. In addition, the generated satellite views are synthesized rather than grounded in physical sensor data, inaccuracies in perspective alignment and scene geometry can arise, particularly in areas with limited landmark visibility or ambiguous visual cues. Although the template matching approach for localization works well in controlled situations, its usefulness in dynamic urban settings is limited due to its potential sensitivity to scale, and temporal variations.

For Street2GIS, a significant limitation is the dependency on the quality and consistency of depth estimation and semantic segmentation. While the pipeline integrates depth cues and semantic features to generate GIS-compatible shapefiles, errors in segmentation boundaries or depth artifacts can propagate through the raster-to-polygon conversion stage, leading to misaligned or distorted geospatial outputs. Furthermore, the framework currently assumes a relatively static urban scene and does not explicitly handle temporal variability, such as occlusions caused by moving objects or seasonal landscape changes. Another challenge is the lack of a large-scale comparative evaluation with official municipal GIS datasets. Although initial results are encouraging, a more systematic validation across different cities, terrains, and satellite sources is required to fully assess the robustness and long-term scalability of the approach.

Both frameworks also share broader limitations common to vision-based systems. These include sensitivity to sensor noise, potential biases in training datasets, and the computational cost of inference when deployed on edge devices. In addition, integrating these methods into existing urban data infrastructures remains a non-trivial challenge due to format compatibility, lack of real-time updating mechanisms, and limited standardization in city-level GIS workflows.

## 5.4. Conclusion

This chapter introduced two complementary frameworks that address the challenges of pedestrian-centric urban mapping and localization using multimodal data. street2sat leveraged generative AI to transform ground-level imagery into satellite-style views, enabling cross-view localization without the need for extensive 3D maps. Meanwhile, Street2GIS focused on generating GIS shape files through a modular pipeline combining depth estimation, segmentation, and raster-to-polygon conversion. Both frameworks demonstrated promising results in extracting meaningful geospatial information from ground-based sensor data and highlighted the value of integrating visual semantics, depth cues, and spatial reasoning for urban understanding.

The chapter also emphasized the versatility of these approaches in addressing different real-world constraints, such as limited sensor coverage, changing environmental conditions, and the need for scalable solutions. The DELTA dataset served as a benchmark to evaluate model performance under realistic urban scenarios, validating the generalizability and robustness of both frameworks. While

the results are encouraging, the work also revealed limitations in segmentation accuracy, cross-view matching under structural variation, and challenges in aligning outputs with official GIS standards. These insights form the basis for future research into more tightly coupled fusion architectures, improved scene understanding, and deeper integration with municipal data systems. Overall, the methodologies proposed in this chapter move toward scalable, adaptive, and context-aware solutions for urban localization and geospatial intelligence.

# 6. CONCLUSION

## 6.1. Summary of contributions and conclusion

This doctoral thesis explores urban localization and mapping through an integrated framework of innovative datasets, multimodal sensing strategies, image-based localization techniques, and automated GIS generation methods. By assembling and analyzing the DELTA dataset, a high-resolution multimodal repository of urban pedestrian pathways, the research introduces foundational insights into complex urban environments, where sensor fusion and spatiotemporal variability play pivotal roles. Leveraging generative AI models for IBL and landmark recognition, the work bridges ground-level and overhead imagery, enhancing localization accuracy and contextual understanding in rapidly changing urban settings. Finally, the Street2GIS framework demonstrates the capacity for automated shapefile generation from minimal inputs, successfully integrating monocular depth estimation, semantic segmentation, and cross-view image synthesis to produce reliable, detailed geographic information. Collectively, these advances expand the frontiers of urban mapping and navigation. They highlight the value of collecting diverse sensory data, adapting to varied environmental conditions, and employing sophisticated computational models that can generalize across different locations and time periods. Taken together, these insights form a robust and scalable foundation for context-aware mapping, setting the stage for next-generation autonomous navigation systems, city planning tools, and smart mobility solutions.

## 6.2. Real world deployment considerations

The deployment of the proposed frameworks—DELTA data acquisition, street2sat localization, and Street2GIS mapping—into real-world urban environments introduces several practical constraints that warrant critical consideration.

One key challenge is the computational cost associated with running deep learning models and multimodal fusion pipelines on embedded or mobile platforms. While training for both street2sat and Street2GIS was performed offline using high-performance computing resources, real-time inference on resource-constrained devices (e.g., e-scooters or delivery robots) remains non-trivial. Models involving image synthesis, segmentation, and cross-view matching can be computationally intensive and may not be suitable for direct deployment without optimization. Future work will explore model compression, quantization, and edge-focused architectures such as MobileNets or transformer distillation to ensure that these methods can operate efficiently in on-device settings.

In addition, the collection and processing of visual and audio data in public spaces raise important data privacy concerns. Although the DELTA platform was used exclusively for research in publicly accessible areas, scaling this system to real-world urban deployments must adhere to privacy regulations such as the GDPR. This includes considerations like automatic face and license plate anonymization, consent-based data collection in semi-private areas, and restrictions on audio recording in sensitive zones. Implementing on-device data filtering and anonymization prior to storage or transmission will be essential for ethical and legal compliance.

Lastly, integration with municipal GIS workflows presents practical hurdles. While Street2GIS produces GIS-compatible shapefiles, each municipality may have its own standards, schemas, and update protocols that must be respected. Real-world deployment would require robust tools for aligning generated data with existing infrastructure records, handling data versioning, and validating the semantic consistency of detected features. Furthermore, the lack of standardized APIs or shared formats between deep learning-based pipelines and legacy GIS systems poses an interoperability barrier. Future work will aim to establish collaborative pilot projects with local governments and urban planning agencies to validate integration procedures and build tools that bridge this gap effectively.

By addressing these deployment challenges, the methodologies proposed in this thesis can be translated into scalable, compliant, and usable systems that not only advance academic research but also contribute meaningfully to real-world urban sensing and planning.

## 6.3. Future works

Looking ahead, several key avenues for future research and development emerge: Data Expansion and Diversity: A critical next step is broadening the geographic, temporal, and environmental scope of collected datasets. By expanding data acquisition beyond initial focus regions and incorporating multiple seasons, times of day, and varied weather conditions, these datasets will become more representative of real-world complexity. Integrating additional mobility platforms—such as e-scooters and other lightweight vehicles—will facilitate more frequent and wide-ranging data updates, ensuring that models remain current and accurately reflect evolving urban landscapes.

Sensor fusion and localization enhancement: A key avenue for future work is the development of tightly coupled sensor fusion strategies that combine GNSS, IMU, and vision-based odometry to improve pose estimation in signal-degraded environments. Integrating these modalities within a unified probabilistic or learning-based framework would allow for drift correction and more robust trajectory estimation, especially in dense urban areas or areas with limited satellite visibility. Such fusion architectures will enable more consistent localization over longer distances and support real-time deployment on resource-constrained platforms.

Acoustic localization and soundscape-informed mapping: Another promising direction lies in the use of environmental sound patterns for spatial reasoning. Investigating the potential of acoustic SLAM—leveraging ambient soundscapes for location estimation—could complement existing visual and LiDAR-based techniques, especially in low-light or visually ambiguous settings. Coupling acoustic cues with generative mapping techniques may offer a low-cost alternative to full 3D reconstruction in certain applications, while also enabling a deeper understanding of urban ambiance and pedestrian experiences.

Enhanced multimodal fusion and contextual understanding: Future research should investigate improved sensor fusion strategies, integrating audio, visual, LiDAR, GNSS, and other modalities for richer contextual insights. Advanced scene understanding models, including transformer-based architectures and attention mechanisms, can further enhance object recognition and spatial reasoning. By refining the coordination between ground-level and overhead imagery through generative AI and cross-view synthesis, it will become possible to produce consistent, high-fidelity representations of urban environments with greater adaptability to environmental changes and seasonal shifts.

Algorithmic robustness and generalization: Efforts should focus on algorithmic improvements that address scaling, computational efficiency, and domain adaptation. Enhancing models to handle novel or rapidly changing scenarios, reducing their susceptibility to seasonal or geographic biases, and ensuring efficient resource usage are all critical for deploying these technologies in real-time, on-device applications.

Dynamic, real-time updating and global applicability: Implementing mechanisms for near real-time database updates will allow continuous adaptation to infrastructural modifications, construction projects, or evolving pedestrian and vehicle behaviors. Strategies to generalize beyond a single city or region—through extensive data augmentation, larger training corpora, and advanced domain adaptation—will help ensure global applicability, supporting diverse urban landscapes worldwide.

Interdisciplinary collaboration and standardization: Engaging with urban planners, local governments, and industry stakeholders will facilitate the integration of these tools into city planning and intelligent transportation frameworks. Establishing common standards and open data initiatives will encourage broader collaboration, accelerating innovation and the adoption of these methodologies on a wider scale.

# BIBLIOGRAPHY


[24] *OpenCV Documentation - Calib3d Module*. OpenCV. 2024. URL: https://docs.opencv.org/4.x/d9/d0c/group__calib3d.html.

[Ado+22] Daniel Adolfsson et al. "Lidar-level localization with radar? the cfear approach to accurate, fast, and robust large-scale radar odometry in diverse environments". In: *IEEE Transactions on robotics* 39.2 (2022), pp. 1476–1495.

[AG24] Oluwayemi-Oniya Aderibigbe and Trynos Gumbo. "Smart Cities and Their Impact on Urban Transportation Systems and Development". In: *Emerging Technologies for Smart Cities: Sustainable Transport Planning in the Global North and Global South*. Springer, 2024, pp. 105–129.

[Akb+06] Amir Akbarzadeh et al. "Towards urban 3d reconstruction from video". In: *Third International Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'06)*. IEEE. 2006, pp. 1–8.

[AKB08] Motilal Agrawal, Kurt Konolige, and Morten Rufus Blas. "Censure: Center surround extremas for realtime feature detection and matching". In: *European conference on computer vision*. Springer. 2008, pp. 102–115.

[AKP24] Deema Almaskati, Sharareh Kermanshachi, and Apurva Pamidimukkala. "Investigating the impacts of autonomous vehicles on crash severity and traffic safety". In: *Frontiers in Built Environment* 10 (2024), p. 1383144.

[Alc+11] Pablo F Alcantarilla et al. "Visibility learning in large-scale urban environment". In: *2011 IEEE International Conference on Robotics and Automation*. IEEE. 2011, pp. 6205–6212.

[ALD21] Rusul L Abduljabbar, Sohani Liyanage, and Hussein Dia. "The role of micro-mobility in shaping sustainable cities: A systematic literature review". In: *Transportation research part D: transport and environment* 92 (2021), p. 102734.

[Alf+18] Abdullah Alfarrarjeh et al. "A data-centric approach for image scene localization". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 594–603.

[Alv+24] Elin Alverhed et al. "Autonomous last-mile delivery robots: a literature review". In: *European Transport Research Review* 16.1 (2024), p. 4.

[AOV12] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. "Freak: Fast retina keypoint". In: *2012 IEEE conference on computer vision and pattern recognition*. Ieee. 2012, pp. 510–517.

[Ara+16] Relja Arandjelovic et al. “NetVLAD: CNN architecture for weakly supervised place recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5297–5307.

[ARS14] Mathieu Aubry, Bryan C Russell, and Josef Sivic. “Painting-to-3D model alignment via discriminative visual elements”. In: *ACM Transactions on Graphics (ToG)* 33.2 (2014), pp. 1–14.

[AVH24] Alireza Akhavi Zadegan, Damien Vivet, and Amnir Hadachi. “DELTA: Integrating Multimodal Sensing with Micromobility for Enhanced Sidewalk and Pedestrian Route Understanding”. In: *Sensors* 24.12 (2024), p. 3863.

[Azi+22] Dejan Azinović et al. “Neural rgb-d surface reconstruction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 6290–6301.

[Baj18] William Bajjali. “Data Acquisition and Getting Data into GIS”. In: *ArcGIS for Environmental and Water Issues*. Cham: Springer International Publishing, 2018, pp. 41–66. ISBN: 978-3-319-61158-7. DOI: 10.1007/978-3-319-61158-7_4. URL: https://doi.org/10.1007/978-3-319-61158-7_4.

[BAS19] Leonard Baum, Tom Assmann, and Henning Strubelt. “State of the art-Automated micro-vehicles for urban logistics”. In: *IFAC-Papers OnLine* 52.13 (2019), pp. 2455–2462.

[BC21] Asli Ozcevik Bilen and Zerhan Yuksel Can. “An applied soundscape approach for acoustic evaluation–compatibility with ISO 12913”. In: *Applied Acoustics* 180 (2021), p. 108112.

[Bee+20] Edward Beeching et al. “Learning to plan with uncertain topological maps”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 473–490.

[Ben+16] Archith John Bency et al. “Weakly supervised localization using deep feature maps”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer. 2016, pp. 714–731.

[Ben75] Jon Louis Bentley. “Multidimensional binary search trees used for associative searching”. In: *Communications of the ACM* 18.9 (1975), pp. 509–517.

[Ber+22] Gabriele Berton et al. “Deep visual geo-localization benchmark”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5396–5407.

[BKV23] Neha Bagwari, Sushil Kumar, and Vivek Singh Verma. “A comprehensive review on segmentation techniques for satellite images”. In: *Archives of Computational Methods in Engineering* 30.7 (2023), pp. 4325–4358.

[BLP18] Vassileios Balntas, Shenlong Li, and Victor Prisacariu. “RelocNet: Continuous metric learning relocalisation using neural nets”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 751–767.

[BMG11] Tom Botterill, Steven Mills, and Richard Green. “Bag-of-words-driven, single-camera simultaneous localization and mapping”. In: *Journal of Field Robotics* 28.2 (2011), pp. 204–226.

[Bos+18] Erkan Bostanci et al. “Sensor fusion of camera, GPS and IMU using fuzzy adaptive multiple motion models”. In: *Soft Computing* 22 (2018), pp. 2619–2632.

[BR13] Patric Beinschob and Christoph Reinke. “Strategies for 3D data acquisition and mapping in large-scale modern warehouses”. In: *2013 IEEE 9th International Conference on Intelligent Computer Communication and Processing (ICCP)*. IEEE. 2013, pp. 229–234.

[BS18] Jens Behley and Cyrill Stachniss. “Efficient Surfel-Based SLAM using 3D Laser Range Data in Urban Environments.” In: *Robotics: Science and Systems*. Vol. 2018. 2018, p. 59.

[BS22] Rounaq Basu and Andres Sevtsuk. “How do street attributes affect willingness-to-walk? City-wide pedestrian route choice analysis using big data from Boston and San Francisco”. In: *Transportation research part A: policy and practice* 163 (2022), pp. 1–19.

[BTV06] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. “Surf: Speeded up robust features”. In: *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*. Springer. 2006, pp. 404–417.

[Cai+19] Sudong Cai et al. “Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 8391–8400.

[Cal+10] Michael Calonder et al. “Brief: Binary robust independent elementary features”. In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer. 2010, pp. 778–792.

[Can81] H Cantzler. “Random sample consensus (ransac)”. In: *Institute for Perception, Action and Behaviour, Division of Informatics, University of Edinburgh* 3 (1981).

[Car+06] Francois Caron et al. “GPS/IMU data fusion using multisensor Kalman filtering: introduction of contextual aspects”. In: *Information fusion* 7.2 (2006), pp. 221–230.

[Che+19] Wentao Cheng et al. “Cascaded parallel filtering for memory-efficient image-based localization”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1032–1041.

[Che+20] Changhao Chen et al. “A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence”. In: *arXiv preprint arXiv:2006.12567* (2020).

[Che+21] Xieyuanli Chen et al. “Range image-based LiDAR localization for autonomous vehicles”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 5802–5808.

[Che+22] Hao Chen et al. “SW-GAN: Road extraction from remote sensing imagery using semi-weakly supervised adversarial learning”. In: *Remote Sensing* 14.17 (2022), p. 4145.

[Che+23] Shenglong Chen et al. “Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 195 (2023), pp. 129–152.

[Cla+17] Ronald Clark et al. “VidLoc: A Deep Spatio-Temporal Model for 6-DOF Video-Clip Relocalization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 6856–6864.

[CM21] Nicholas A Coppola and Wesley E Marshall. “Sidewalk static obstructions and their impact on clear width”. In: *Transportation research record* 2675.6 (2021), pp. 200–212.

[CN11] Mark Cummins and Paul Newman. “Appearance-only SLAM at large scale with FAB-MAP 2.0”. In: *The International Journal of Robotics Research* 30.9 (2011), pp. 1100–1123.

[Cor+15] Marius Cordts et al. “The cityscapes dataset”. In: *CVPR Workshop on the Future of Datasets in Vision*. Vol. 2. sn. 2015.

[CSR18] Ming Cai, Chunhua Shen, and Ian D. Reid. “A Hybrid Probabilistic Model for Camera Relocalization”. In: *BMVC*. Vol. 1. 2018, p. 8.

[Dav+07] Andrew J Davison et al. “MonoSLAM: Real-time single camera SLAM”. In: *IEEE transactions on pattern analysis and machine intelligence* 29.6 (2007), pp. 1052–1067.

[Dec+22] Benoit Decoux et al. “A dataset for temporal semantic segmentation dedicated to smart mobility of wheelchairs on sidewalks”. In: *Journal of imaging* 8.8 (2022), p. 216.

[Di +12] Cecilia Di Ruberto et al. “Generalized hough transform for shape matching”. In: *International Journal of Computer Applications* 975 (2012), p. 8887.

[Din+19] Ming Ding et al. “CamNet: Coarse-to-Fine Retrieval for Camera Re-Localization”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2019, pp. 2871–2880.

[Dis+01] MWM Gamini Dissanayake et al. “A solution to the simultaneous localization and map building (SLAM) problem”. In: *IEEE Transactions on robotics and automation* 17.3 (2001), pp. 229–241.

[Dju+03] Petar M Djuric et al. “Particle filtering”. In: *IEEE signal processing magazine* 20.5 (2003), pp. 19–38.

[DM23] Simon Durbridge and Damian Thomas Murphy. “Assessment of soundscapes using self-report and physiological measures”. In: *Acta Acustica* 7 (2023), p. 6.

[DMR18] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. “Superpoint: Self-supervised interest point detection and description”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2018, pp. 224–236.

[Dos+20] Alexey Dosovitskiy et al. “An image is worth 16x16 words: Transformers for image recognition at scale”. In: *arXiv preprint arXiv:2010.11929* (2020).

[EP08] Georgios D Evangelidis and Emmanouil Z Psarakis. “Parametric image alignment using enhanced correlation coefficient maximization”. In: *IEEE transactions on pattern analysis and machine intelligence* 30.10 (2008), pp. 1858–1865.

[FB81] Martin A Fischler and Robert C Bolles. “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Communications of the ACM* 24.6 (1981), pp. 381–395.

[FBC23] Carmen Fernández-Aguilar, Marta Brosed-Lázaro, and Demetrio Carmona-Derqui. “Effectiveness of Mobility and Urban Sustainability Measures in Improving Citizen Health: A Scoping Review”. In: *International Journal of Environmental Research and Public Health* 20.3 (2023), p. 2649.

[FEN07] Friedrich Fraundorfer, Christopher Engels, and David Nistér. “Topological mapping, localization and navigation using image collections”. In: *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. Ieee. 2007, pp. 3872–3877.

[FLG14] Simon Fuhrmann, Fabian Langguth, and Michael Goesele. “Mve-a multi-view reconstruction environment.” In: *GCH* 3 (2014), p. 4.

[Fro+22] Jon E Froehlich et al. “Towards Mapping and Assessing Sidewalk Accessibility Across Sociocultural and Geographic Contexts”. In: *arXiv preprint arXiv:2207.13626* (2022).

[Gao+03] Xiao-Shan Gao et al. “Complete solution classification for the perspective-three-point problem”. In: *IEEE transactions on pattern analysis and machine intelligence* 25.8 (2003), pp. 930–943.

[Ge+20] Yixiao Ge et al. “Self-supervising fine-grained region similarities for large-scale image localization”. In: *Computer Vision–ECCV 2020:*

*16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*. Springer. 2020, pp. 369–386.

[GEB15] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. "A neural algorithm of artistic style". In: *arXiv preprint arXiv:1508.06576* (2015).

[Gei+13] Andreas Geiger et al. "Vision meets robotics: The kitti dataset". In: *The International Journal of Robotics Research* 32.11 (2013), pp. 1231–1237.

[Gem+17] Jort F Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[Ghr+12] R Ghrist et al. "Topological landmark-based navigation and mapping". In: *University of Pennsylvania, Department of Mathematics, Tech. Rep* 8 (2012).

[GL21] Vivien Sainte Fare Garnot and Loic Landrieu. "Panoptic segmentation of satellite image time series with convolutional temporal attention networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 4872–4881.

[Gli+23] Szymon Glinka et al. "The use of deep learning methods for object height estimation in high resolution satellite images". In: *Sensors* 23.19 (2023), p. 8162.

[GO15] Emilio Garcia-Fidalgo and Alberto Ortiz. "Vision-based topological mapping and localization methods: A survey". In: *Robotics and Autonomous Systems* 64 (2015), pp. 1–20.

[Goo+14] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[Guo+21] Hongliang Guo et al. "Autonomous navigation in dynamic environments with multi-modal perception uncertainties". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 9255–9261.

[Gur+11] Topraj Gurung et al. "SQuad: Compact representation for triangle meshes". In: *Computer Graphics Forum*. Vol. 30. 2. Wiley Online Library. 2011, pp. 355–364.

[Har+94] Bert M Haralick et al. "Review and analysis of solutions of the three point perspective pose estimation problem". In: *International journal of computer vision* 13 (1994), pp. 331–356.

[HBK20] Thorsten Hoeser, Felix Bachofer, and Claudia Kuenzer. "Object detection and image segmentation with deep learning on Earth observation data: A review—Part II: Applications". In: *Remote Sensing* 12.18 (2020), p. 3053.

[He+17] Kaiming He et al. “Mask r-cnn”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[HE08] James Hays and Alexei A Efros. “Im2gps: estimating geographic information from a single image”. In: *2008 ieee conference on computer vision and pattern recognition*. IEEE. 2008, pp. 1–8.

[His+15a] MB Hisham et al. “Template matching using sum of squared difference and normalized cross correlation”. In: *2015 IEEE student conference on research and development (SCOReD)*. IEEE. 2015, pp. 100–104.

[His+15b] Mohd Badrul Hisham et al. “Template Matching using Sum of Squared Difference and Normalized Cross Correlation”. In: *2015 IEEE Student Conference on Research and Development (SCOReD)* (2015), pp. 100–104. URL: https://api.semanticscholar.org/CorpusID:42080729.

[HL20] Sixing Hu and Gim Hee Lee. “Image-based geo-localization using satellite imagery”. In: *International Journal of Computer Vision* 128.5 (2020), pp. 1205–1219.

[HL24] Sabir Hossain and Xianke Lin. “Enhancing mapping precision in autonomous delivery robots through tightly-coupled fusion of uncertainty-aware GPS and LiDAR odometry”. In: *International Journal of Intelligent Robotics and Applications* (2024), pp. 1–15.

[Hos+22] Maryam Hosseini et al. “CitySurfaces: City-scale semantic segmentation of sidewalk materials”. In: *Sustainable Cities and Society* 79 (2022), p. 103630.

[Hos+23] Maryam Hosseini et al. “Mapping the walk: A scalable computer vision approach for generating sidewalk network datasets from aerial imagery”. In: *Computers, Environment and Urban Systems* 101 (2023), p. 101950.

[How+17] Andrew G Howard et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications”. In: *arXiv preprint arXiv: 1704.04861* (2017).

[HQM14] Iris Heisterklaus, Ningqing Qian, and Artur Miller. “Image-based pose estimation using a compact 3d model”. In: *2014 IEEE Fourth International Conference on Consumer Electronics Berlin (ICCE-Berlin)*. IEEE. 2014, pp. 327–330.

[HR11] Joel A Hesch and Stergios I Roumeliotis. “A direct least-squares (DLS) method for PnP”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 383–390.

[Hsu+23] Li-Ta Hsu et al. “Hong Kong UrbanNav: An open-source multisensory dataset for benchmarking urban navigation algorithms”. In: *NAVIGATION: Journal of the Institute of Navigation* 70.4 (2023).

[Hu+18] Sixing Hu et al. “Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 7258–7267.

[Hua+19] Zhiqiang Huang et al. “Prior Guided Dropout for Robust Visual Localization in Dynamic Environments”. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* . 2019, pp. 2791–2800.

[Hua+21] Zhaoyang Huang et al. “Vs-net: Voting with segmentation for visual localization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 6101–6111.

[HW95] David Heckerman and Michael P Wellman. “Bayesian networks”. In: *Communications of the ACM* 38.3 (1995), pp. 27–31.

[HYS19] Jiawei Hou, Yijun Yuan, and Sören Schwertfeger. “Area graph: Generation of topological maps using the voronoi diagram”. In: *2019 19th International Conference on Advanced Robotics (ICAR)*. IEEE. 2019, pp. 509–515.

[Int+24] Marco Introvigne et al. “Real-Time Environment Condition Classification for Autonomous Vehicles”. In: *arXiv preprint arXiv:2405. 19305* (2024).

[Iso+17] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134.

[JF19] Dylan Jennings and Miguel Figliozzi. “Study of sidewalk autonomous delivery robots and their potential impacts on freight efficiency and travel”. In: *Transportation Research Record* 2673.6 (2019), pp. 317–326.

[Jon14] Peter Jones. “The evolution of urban mobility: The interplay of academic and policy perspectives”. In: *IATSS research* 38.1 (2014), pp. 7–13.

[JS11] Eagle S Jones and Stefano Soatto. “Visual-inertial navigation, mapping and localization: A scalable real-time causal approach”. In: *The International Journal of Robotics Research* 30.4 (2011), pp. 407–430.

[Jus+24] Jon Alvarez Justo et al. “Semantic Segmentation in Satellite Hyperspectral Imagery by Deep Learning”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024).

[JYM14] Hee-Seon Jin, Wonpil Yu, and Hyungpil Moon. “Merging of topological map and grid map using standardized map data representation”. In: *The Journal of Korea Robotics Society* 9.2 (2014), pp. 104–110.

[Kan+16] Vadim Kantorov et al. “Contextlocnet: Context-aware deep network models for weakly supervised localization”. In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer. 2016, pp. 350–365.

[KC17] Alex Kendall and Roberto Cipolla. “Geometric Loss Functions for Camera Pose Regression with Deep Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 5974–5983.

[KGC15] Alex Kendall, Matthew Grimes, and Roberto Cipolla. “Posenet: A convolutional network for real-time 6-dof camera relocalization”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2938–2946.

[KGR22] Yuhao Kang, Song Gao, and Robert Roth. “A review and synthesis of recent GeoAI research for cartography: Methods, applications, and ethics”. In: *Proceedings of AutoCarto*. 2022, pp. 2–4.

[Kir+19] Alexander Kirillov et al. “Panoptic segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9404–9413.

[Kir+23] Alexander Kirillov et al. “Segment anything”. In: *arXiv preprint arXiv:2304.02643* (2023).

[KMB22] Matija Kovačić, Maja Mutavdžija, and Krešimir Buntak. “New paradigm of sustainable urban mobility: Electric and autonomous vehicles—A review and bibliometric analysis”. In: *Sustainability* 14.15 (2022), p. 9525.

[Krö+01] Ben JA Kröse et al. “A probabilistic model for appearance-based robot localization”. In: *Image and Vision Computing* 19.6 (2001), pp. 381–391.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Advances in neural information processing systems* 25 (2012).

[Kuu+18] Sampo Kuutti et al. “A survey of the state-of-the-art localization techniques and their potentials for autonomous vehicle applications”. In: *IEEE Internet of Things Journal* 5.2 (2018), pp. 829–846.

[Lan20] Land Board of the Republic of Estonia. *ETAK Layer Names and Descriptions*. Accessed: 2024-11-01, Document Date: 28-10-2020. 2020. URL: `https://geoportaal.maaamet.ee/eng/Spatial-Data/Topographic-Maps/Estonian-Basic-Map-1-10-000/ETAK-layer-names-and-descriptions-p710.html`.

[Lar+16] Viktor Larsson et al. “Outlier Rejection for Absolute Pose Estimation with Known Orientation.” In: *BMVC*. 2016.

[Las+17] Zakaria Laskar et al. “Camera relocalization by computing pairwise relative poses using convolutional neural network”. In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 929–938.

[LBG20] Wenwen Li, Michael Batty, and Michael F Goodchild. *Real-time GIS for smart cities*. 2020.

[LCS11] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. “BRISK: Binary robust invariant scalable keypoints”. In: *2011 International conference on computer vision*. Ieee. 2011, pp. 2548–2555.

[Led+17] Christian Ledig et al. “Photo-realistic single image super-resolution using a generative adversarial network”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 4681–4690.

[Li+12] Yunpeng Li et al. “Worldwide pose estimation using 3d point clouds”. In: *European conference on computer vision*. Springer. 2012, pp. 15–29.

[Li+20] Qingqing Li et al. “Multi-sensor fusion for navigation and mapping in autonomous vehicles: Accurate localization in urban environments”. In: *Unmanned Systems* 8.03 (2020), pp. 229–237.

[Li+21a] Yan Li et al. “Urban vehicle localization in public LoRaWan network”. In: *IEEE Internet of Things Journal* 9.12 (2021), pp. 10283–10294.

[Li+21b] Ziming Li et al. “A deep learning-based framework for automated extraction of building footprint polygons from very high-resolution aerial imagery”. In: *Remote Sensing* 13.18 (2021), p. 3630.

[Lin+24] Qingming Lin et al. “ShapefileGPT: A Multi-Agent Large Language Model Framework for Automated Shapefile Processing”. In: *arXiv preprint arXiv:2410.12376* (2024).

[Liu+20] Wenxin Liu et al. “Tlio: Tight learned inertial odometry”. In: *IEEE Robotics and Automation Letters* 5.4 (2020), pp. 5653–5660.

[Liu+21] Hongyu Liu et al. “Pd-gan: Probabilistic diverse gan for image inpainting”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9371–9381.

[Liu+24a] Mingbo Liu et al. “A Method for Extracting High-Resolution Building Height Information in Rural Areas Using GF-7 Data”. In: *Sensors* 24.18 (2024), p. 6076.

[Liu+24b] Ruyi Liu et al. “A Review of Deep Learning-Based Methods for Road Extraction from High-Resolution Remote Sensing Images”. In: *Remote Sensing* 16.12 (2024), p. 2056.

[LL19] Liu Liu and Hongdong Li. “Lending orientation to neural networks for cross-view geo-localization”. In: *Proceedings of the IEEE/CVF*

*conference on computer vision and pattern recognition*. 2019, pp. 5624–5633.

[LL23] Yichun Lu and Siu-Kit Lau. "Soundscape evaluation method based on participatory design". In: *Proceedings of Meetings on Acoustics*. Vol. 51. 1. AIP Publishing. 2023.

[LLD17] Liu Liu, Hongdong Li, and Yuchao Dai. "Efficient global 2d-3d matching for camera localization in a large-scale 3d map". In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2372–2381.

[LN23] Zhenlong Li and Huan Ning. "Autonomous GIS: the next-generation AI-powered GIS". In: *International Journal of Digital Earth* 16.2 (2023), pp. 4668–4686.

[Low04] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60 (2004), pp. 91–110.

[LSD15] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[LSH10] Yunpeng Li, Noah Snavely, and Daniel P Huttenlocher. "Location recognition using prioritized feature matching". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11*. Springer. 2010, pp. 791–804.

[LSR21] Mansoureh Sharifzadeh Lari, Julien Straubhaar, and Philippe Renard. "Efficiency of template matching methods for Multiple-Point Statistics simulations". In: *Applied Computing and Geosciences* 11 (2021), p. 100064.

[LT10] Jesse Levinson and Sebastian Thrun. "Robust vehicle localization in urban environments using probabilistic maps". In: *2010 IEEE international conference on robotics and automation*. IEEE. 2010, pp. 4372–4378.

[Lu+19] Weixin Lu et al. "L3-net: Towards learning based lidar localization for autonomous driving". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 6389–6398.

[Luo+19] Ji Luo et al. "Developing an aerial-image-based approach for creating digital sidewalk inventories". In: *Transportation research record* 2673.8 (2019), pp. 499–507.

[LVD20] Hugh Louch, Kim Voros, and Erin David. *Availability and Use of Pedestrian Infrastructure Data to Support Active Transportation*

*Planning*. Project 20-05, Topic 50-10. Transportation Research Board, National Research Council, Washington, DC, 2020.

[LXG22] Yiyi Liao, Jun Xie, and Andreas Geiger. "Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2022), pp. 3292–3310.

[Mal+18] Evangelos Maltezos et al. "Building extraction from LiDAR data applying deep convolutional neural networks". In: *IEEE Geoscience and Remote Sensing Letters* 16.1 (2018), pp. 155–159.

[Man19] Davide Liberato Manna. "Road Feature Extraction With Deep Learning Methods". PhD thesis. Politecnico di Torino, 2019.

[Mav+23] Christoforos Mavrogiannis et al. "Core challenges of social robot navigation: A survey". In: *ACM Transactions on Human-Robot Interaction* 12.3 (2023), pp. 1–39.

[MC23] Luz E Marquez and Maria Calle. "Understanding LoRa-based Localization: Foundations and Challenges". In: *IEEE Internet of Things Journal* (2023).

[Mel+17] Iurii Melekhov et al. "Image-Based Localization Using Hourglass Networks". In: *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 2017, pp. 879–886.

[Mel+19] Iaroslav Melekhov et al. "Dgc-net: Dense geometric correspondence network". In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 1034–1042.

[MEM+20] Abdulkadir MEMDUHOGLU et al. "3D Map Experience for Youth with Virtual/Augmented Reality Applications". In: *Harran Üniversitesi Mühendislik Dergisi* 5.3 (2020), pp. 175–182.

[Mer+14] Pierre Merriaux et al. "Wheel odometry-based car localization and tracking on vectorial map". In: *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*. IEEE. 2014, pp. 1890–1891.

[MFD24] Hisham Y Makahleh, Emma Jayne Sakamoto Ferranti, and Dilum Dissanayake. "Assessing the Role of Autonomous Vehicles in Urban Areas: A Systematic Review of Literature". In: *Future Transportation* 4.2 (2024), pp. 321–348.

[MH21] Kareem Mostafa and Tarek Hegazy. "Review of image-based analysis and applications in construction". In: *Automation in Construction* 122 (2021), p. 103516.

[Mil+21] Ben Mildenhall et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". In: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[Mil+24] Tymoteusz Miller et al. “A Critical AI View on Autonomous Vehicle Navigation: The Growing Danger”. In: *Electronics* 13.18 (2024), p. 3660.

[Mir+16] Fabio Miranda et al. “Urban pulse: Capturing the rhythm of cities”. In: *IEEE transactions on visualization and computer graphics* 23.1 (2016), pp. 791–800.

[Mis+21] Márk Miskolczi et al. “Urban mobility scenarios until the 2030s”. In: *Sustainable Cities and Society* 72 (2021), p. 103029.

[Mor+18] Juan Miguel Barrigón Morillas et al. “Noise pollution and urban planning”. In: *Current Pollution Reports* 4.3 (2018), pp. 208–219.

[Mou14] Mary Luz Mouronte. “Topological analysis of the subway network of madrid”. In: *ICCGI 2014* (2014), p. 22.

[MP07] Daniel Martinec and Tomas Pajdla. “Robust rotation and translation estimation in multiview reconstruction”. In: *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2007, pp. 1–8.

[NB17] Taimoor Naseer and Wolfram Burgard. “Deep Regression for Monocular Camera-Based 6-DOF Global Localization in Outdoor Environments”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2017, pp. 1525–1530.

[Neu+17] Gerhard Neuhold et al. “The mapillary vistas dataset for semantic understanding of street scenes”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 4990–4999.

[NHB08] Andrew Nevin, Alan Scottedward Hodel, and David M. Bevly. “OBJECT REGISTRATION THROUGH STATISTICS AND HOUGH TRANSFORM”. In: 2008. URL: `https://api.semanticscholar.org/CorpusID:64262001`.

[Nin+22] Huan Ning et al. “Sidewalk extraction using aerial and street view images”. In: *Environment and Planning B: Urban Analytics and City Science* 49.1 (2022), pp. 7–22.

[NNB04] David Nistér, Oleg Naroditsky, and James Bergen. “Visual odometry”. In: *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.* Vol. 1. Ieee. 2004, pp. I–I.

[Nog+22] Ana Filipa Rodrigues Nogueira et al. “Sound Classification and Processing of Urban Environments: A Systematic Literature Review”. In: *Sensors* 22.22 (2022), p. 8608.

[Ort+22] Joseph Ortiz et al. “isdf: Real-time neural signed distance fields for robot perception”. In: *arXiv preprint arXiv:2204.02296* (2022).

[OT01] Aude Oliva and Antonio Torralba. “Modeling the shape of the scene: A holistic representation of the spatial envelope”. In: *International journal of computer vision* 42 (2001), pp. 145–175.

[OT06] Aude Oliva and Antonio Torralba. “Building the gist of a scene: The role of global image features in recognition”. In: *Progress in brain research* 155 (2006), pp. 23–36.

[OY23] Chahinez Ounoughi and Sadok Ben Yahia. “Data fusion for ITS: A systematic literature review”. In: *Information Fusion* 89 (2023), pp. 267–291.

[Par+19] Jeong Joon Park et al. “Deepsdf: Learning continuous signed distance functions for shape representation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 165–174.

[Par+20] Kibaek Park et al. “Sideguide: a large-scale sidewalk dataset for guiding impaired people”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 10022–10029.

[Pic15] Karol J Piczak. “ESC: Dataset for environmental sound classification”. In: *Proceedings of the 23rd ACM international conference on Multimedia*. 2015, pp. 1015–1018.

[PJ24] Vipul Parmar and Arnab Jana. “A review of tools and techniques for audio-visual assessment of urbanscape”. In: *Discover Cities* 1.1 (2024), p. 29.

[PKF07] Jean-Philippe Pons, Renaud Keriven, and Olivier Faugeras. “Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score”. In: *International Journal of Computer Vision* 72 (2007), pp. 179–193.

[PM85] J Douglas Porteous and Jane F Mastin. “Soundscape”. In: *Journal of Architectural and Planning Research* (1985), pp. 169–186.

[PP22] Roberto Pierdicca and Marina Paolanti. “GeoAI: a review of artificial intelligence approaches for the interpretation of complex geomatics data”. In: *Geoscientific Instrumentation, Methods and Data Systems Discussions* 2022 (2022), pp. 1–35.

[PZZ18] Prasun Purkait, Chen Zhao, and Christopher Zach. “Synthetic View Generation for Absolute Pose Regression and Image Synthesis”. In: *BMVC*. 2018, p. 69.

[Raj+22] Abdelatif Rajji et al. “Building height estimation from high resolution satellite images”. In: *International Journal of Innovation and Applied Studies* 35.2 (2022), pp. 268–281.

[RB18] Krishna Regmi and Ali Borji. “Cross-view image synthesis using conditional gans”. In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2018, pp. 3501–3510.

[RB19] Krishna Regmi and Ali Borji. “Cross-view image synthesis using geometry-guided conditional gans”. In: *Computer Vision and Image Understanding* 187 (2019), p. 102788.

[Red+16] Joseph Redmon et al. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.

[Red18] G. P. Obi Reddy. “Geographic Information System: Principles and Applications”. In: *Geospatial Technologies in Land Resources Mapping, Monitoring and Management*. Ed. by G. P. Obi Reddy and S. K. Singh. Cham: Springer International Publishing, 2018, pp. 45–62. ISBN: 978-3-319-78711-4. DOI: 10.1007/978-3-319-78711-4_3. URL: https://doi.org/10.1007/978-3-319-78711-4_3.

[Ret+20] Paulo HL Rettore et al. “Road data enrichment framework based on heterogeneous data fusion for ITS”. In: *IEEE transactions on intelligent transportation systems* 21.4 (2020), pp. 1751–1766.

[Ria23] Artem Riabko. “Methods of Satellite Images Segmentation Analysis”. In: *2023 IEEE 7th International Conference on Methods and Systems of Navigation and Motion Control (MSNMC)*. IEEE. 2023, pp. 163–167.

[RL24] Saki Rezwana and Nicholas Lownes. “Interactions and Behaviors of Pedestrians with Autonomous Vehicles: A Synthesis”. In: *Future Transportation* 4.3 (2024), pp. 722–745.

[Rob24] Roboflow. *Roboflow: Computer Vision for Everyone*. Accessed: 2024-12-06. 2024. URL: https://roboflow.com/.

[RTS21] Bin Ren, Hao Tang, and Nicu Sebe. “Cascaded cross mlp-mixer gans for cross-view image translation”. In: *arXiv preprint arXiv:2110.10183* (2021).

[Rub+11] Ethan Rublee et al. “ORB: An efficient alternative to SIFT or SURF”. In: *2011 International conference on computer vision*. Ieee. 2011, pp. 2564–2571.

[Sar+19] Paul-Edouard Sarlin et al. “From coarse to fine: Robust hierarchical localization at large scale”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 12716–12725.

[Sar+20] Paul-Edouard Sarlin et al. “Superglue: Learning feature matching with graph neural networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 4938–4947.

[Sch+23] Brigitte Schulte-Fortkamp et al. *Soundscapes: Humans and their acoustic environment*. Springer, 2023.

[Seo+18] Paul Hongsuck Seo et al. “Cplanet: Enhancing image geolocalization by combinatorial partitioning of maps”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 536–551.

[Sev+21] Andres Sevtsuk et al. “A big data approach to understanding pedestrian route choice preferences: Evidence from San Francisco”. In: *Travel behaviour and society* 25 (2021), pp. 41–51.

[SFC21] Dahlen Silva, Dávid Földes, and Csaba Csiszár. “Autonomous vehicle use and urban space transformation: A scenario building and analysing method”. In: *Sustainability* 13.6 (2021), p. 3008.

[She+21] Akhil Shetty et al. “Safety challenges for autonomous vehicles in the absence of connectivity”. In: *Transportation research part C: emerging technologies* 128 (2021), p. 103133.

[Shi+20] Yujiao Shi et al. “Optimal feature transport for cross-view image geo-localization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 07. 2020, pp. 11990–11997.

[Sib+13] Dominik Sibbing et al. “Sift-realistic rendering”. In: *2013 International Conference on 3D Vision-3DV 2013*. IEEE. 2013, pp. 56–63.

[SLK11] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. “Fast image-based localization using direct 2d-to-3d matching”. In: *2011 International Conference on Computer Vision*. IEEE. 2011, pp. 667–674.

[SMT18] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. “Visual SLAM and structure from motion in dynamic environments: A survey”. In: *ACM Computing Surveys (CSUR)* 51.2 (2018), pp. 1–36.

[SP11] Niko Sünderhauf and Peter Protzel. “Brief-gist-closing the loop by simple means”. In: *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE. 2011, pp. 1234–1241.

[Suc+21] Edgar Sucar et al. “iMAP: Implicit mapping and positioning in real-time”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 6229–6238.

[Sun+19] Chang Sun et al. “Wide-view sidewalk dataset based pedestrian safety application”. In: *IEEE Access* 7 (2019), pp. 151399–151408.

[Sun+20] Pei Sun et al. “Scalability in perception for autonomous driving: Waymo open dataset”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 2446–2454.

[SZ03] Sivic and Zisserman. “Video Google: A text retrieval approach to object matching in videos”. In: *Proceedings ninth IEEE international conference on computer vision*. IEEE. 2003, pp. 1470–1477.

[SZC20] Noe Samano, Mengjie Zhou, and Andrew Calway. “You are here: Geolocation by embedding maps and images”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*. Springer. 2020, pp. 502–518.

[Tan+19] Hao Tang et al. “Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation”. In: *Pro-*

*ceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2417–2426.

[Tan+21a] Shitao Tang et al. "Learning camera localization via dense scene matching". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1831–1841.

[Tan+21b] Tim Y Tang et al. "Self-supervised learning for using overhead imagery as maps in outdoor range sensor localization". In: *The International Journal of Robotics Research* 40.12-14 (2021), pp. 1488–1509.

[TNA20] Linh Duy Tran, Son Minh Nguyen, and Masayuki Arai. "GAN-based noise model for denoising real images". In: *Proceedings of the Asian Conference on Computer Vision*. 2020.

[Tou+18] Sory I Toure et al. "Land cover and land use change analysis using multi-spatial resolution data and object-based image analysis". In: *Remote Sensing of Environment* 210 (2018), pp. 259–268.

[TTB21] Agnieszka Telega, Ivan Telega, and Agnieszka Bieda. "Measuring walkability with GIS—methods overview and new approach proposal". In: *Sustainability* 13.4 (2021), p. 1883.

[UHW23] Usman Ahmad Usmani, Ari Happonen, and Junzo Watada. "Revolutionizing Transportation: Advancements in Robot-Assisted Mobility Systems". In: *International Conference on ICT for Sustainable Development*. Springer. 2023, pp. 603–619.

[ULH16] Steffen Urban, Jens Leitloff, and Stefan Hinz. "Mlpnp-a real-time maximum likelihood solution to the perspective-n-point problem". In: *arXiv preprint arXiv:1607.08112* (2016).

[Ult23] Ultralytics. *Ultralytics GitHub Repository*. `https://github.com/ultralytics/ultralytics`. Accessed: 2024. 2023.

[Ult24] Ultralytics. *Ultralytics GitHub Repository*. 2024. URL: `https://github.com/ultralytics/ultralytics`.

[UN00] Iwan Ulrich and Illah Nourbakhsh. "Appearance-based place recognition for topological localization". In: *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*. Vol. 2. Ieee. 2000, pp. 1023–1029.

[Val+24] Améline Vallet et al. "Generating high-resolution land use and land cover maps for the greater Mariño watershed in 2019 with machine learning". In: *Scientific Data* 11.1 (2024), p. 915.

[Ves+18] Emanuele Vespa et al. "Efficient octree-based volumetric SLAM supporting signed-distance and occupancy mapping". In: *IEEE Robotics and Automation Letters* 3.2 (2018), pp. 1144–1151.

[VMM22] Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. "You only hear once: a YOLO-like algorithm for audio segmentation and sound event detection". In: *Applied Sciences* 12.7 (2022), p. 3293.

[VZD22] Kerman Viana, Asier Zubizarreta, and Mikel Diez. "A Reconfigurable Framework for Vehicle Localization in Urban Areas". In: *Sensors* 22.7 (2022), p. 2595.

[Wal+17] Florian Walch et al. "Image-based localization using lstms for structured feature correlation". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 627–637.

[Wan+04] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[Wan+19] Bowen Wang et al. "AtLoc: Attention Guided Camera Localization". In: *arXiv preprint arXiv:1909.03557* (2019).

[Wan+21a] Dan Wang et al. "Multi-view 3d reconstruction with transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5722–5731.

[Wan+21b] Yanhong Wang et al. "Block-based image matching for image retrieval". In: *Journal of Visual Communication and Image Representation* 74 (2021), p. 102998.

[WB+95] Greg Welch, Gary Bishop, et al. "An introduction to the Kalman filter". In: (1995).

[WBA22] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. "Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction". In: *2022 International Conference on 3D Vision (3DV)*. IEEE. 2022, pp. 433–442.

[Wel+19] Galen Weld et al. "Deep learning for automatically detecting sidewalk accessibility problems using streetscape imagery". In: *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 2019, pp. 196–209.

[Wen+20] Weisong Wen et al. "UrbanLoco: A full sensor suite dataset for mapping and localization in urban scenes". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE. 2020, pp. 2310–2316.

[WMH17] Jie Wu, Lei Ma, and Xiaolin Hu. "Delving Deeper into Convolutional Neural Networks for Camera Relocalization". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2017, pp. 5644–5651.

[Xia+22] Zimin Xia et al. "Visual cross-view metric localization with dense uncertainty estimates". In: *European Conference on Computer Vision*. Springer. 2022, pp. 90–106.

[Xia+24] Jiahao Xia et al. “Pedestrian-accessible infrastructure inventory: enabling and assessing zero-shot segmentation on multi-mode geospatial data for all pedestrian types”. In: *Journal of imaging* 10.3 (2024), p. 52.

[Xie+21] Enze Xie et al. “SegFormer: Simple and efficient design for semantic segmentation with transformers”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 12077–12090.

[XK15] Sonila Xhafa and Albana Kosovrasti. “Geographic Information Systems (GIS) in urban planning”. In: *European Journal of Interdisciplinary Studies* 1.1 (2015), pp. 74–81.

[Yan+24] Lihe Yang et al. “Depth Anything V2”. In: *arXiv preprint arXiv:2406.09414* (2024).

[YIN+24] Shen YING et al. “High Definition Map Model for Autonomous Driving and Key Technologies”. In: *Geomatics and Information Science of Wuhan University* 49.4 (2024), pp. 506–515.

[YK13] Ming Yang and Jian Kang. “Psychoacoustical evaluation of natural and urban sounds in soundscapes”. In: *The Journal of the Acoustical Society of America* 134.1 (2013), pp. 840–851.

[Yu+20] Hongkun Yu et al. “TensorFlow model garden”. In: *GitHub* (2020).

[ZH24] Alireza Akhavi Zadegan and Amnir Hadachi. “Generative-AI based Map Representation and Localization”. In: *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Advances in Urban-AI*. 2024, pp. 34–42.

[Zha+18] Yihang Zhang et al. “Spatial-temporal fraction map fusion with multi-scale remotely sensed images”. In: *Remote Sensing of Environment* 213 (2018), pp. 162–181.

[Zha+22] Chenni Zhang et al. “Building height extraction from GF-7 satellite images based on roof contour constrained stereo matching”. In: *Remote sensing* 14.7 (2022), p. 1566.

[ZHG06] Feng Zhao, Qingming Huang, and Wen Gao. “Image matching by normalized cross-correlation”. In: *2006 IEEE international conference on acoustics speech and signal processing proceedings*. Vol. 2. IEEE. 2006, pp. II–II.

[Zho+20] Qunjie Zhou et al. “To learn or not to learn: Visual localization from essential matrices”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2020, pp. 3319–3326.

[Zho+23] Xingguang Zhong et al. “Shine-mapping: Large-scale 3d mapping using sparse hierarchical implicit neural representations”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 8371–8377.

[Zhu+22] Zihan Zhu et al. “Nice-slam: Neural implicit scalable encoding for slam”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 12786–12796.

[Zhu+23] Zihan Zhu et al. “Nicer-slam: Neural implicit scene encoding for rgb slam”. In: *arXiv preprint arXiv:2302.03594* (2023).

[ZK23] Yuerong Zhang and Maria Kamargianni. “A review on the factors influencing the adoption of new mobility technologies and services: autonomous vehicle, drone, micromobility and mobility as a service”. In: *Transport reviews* 43.3 (2023), pp. 407–429.

[ZMH25] Alireza Akhavi Zadegan, Jose Medina, and Amnir Hadachi. “Street2GIS: Multimodal Generative Framework for Pedestrian Infrastructure Mapping”. In: *Proceedings of the 17th International Conference on Joint Urban Remote Sensing (JURSE)*. To be published. Higher School of Communication of Tunis (SUP’COM). IEEE, 2025.

[ZS14] Ji Zhang and Sanjiv Singh. “LOAM: Lidar odometry and mapping in real-time.” In: *Robotics: Science and systems*. Vol. 2. 9. Berkeley, CA. 2014, pp. 1–9.

[ZYC21] Sijie Zhu, Taojiannan Yang, 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*. 2021, pp. 3640–3649.

# 7. APPENDIX A

## 7.1. Reference architecture for the development of the DELTA dataset

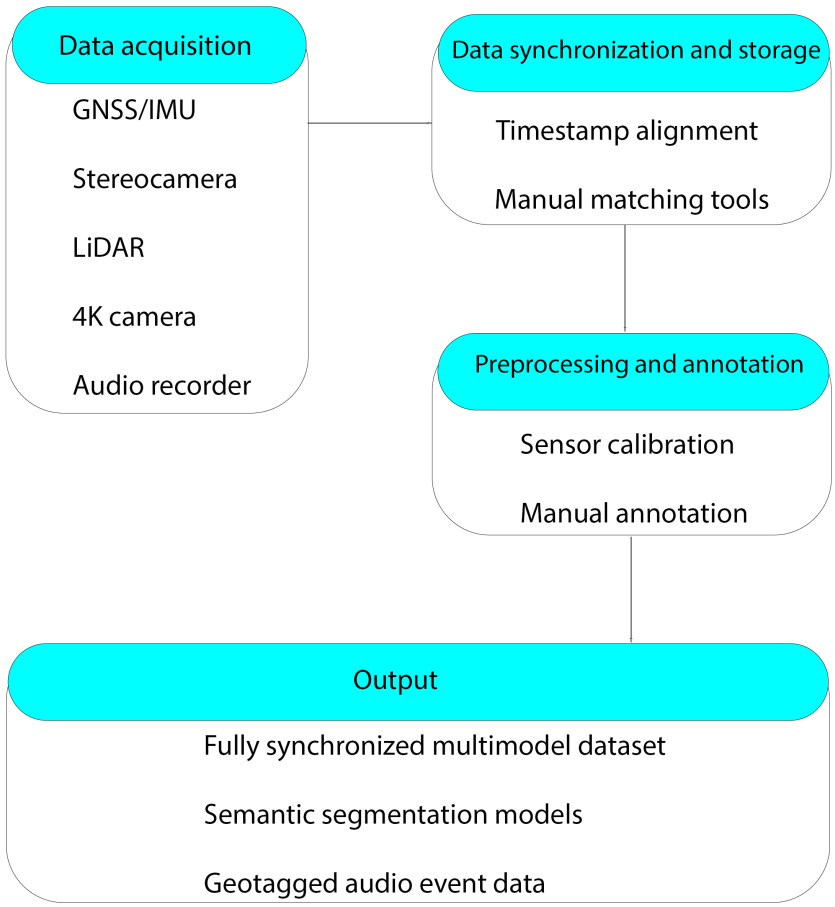



**Figure 42.** DELTA dataset reference architecture

# 8. ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor, Professor Amnir Hadachi, for his continuous guidance, support, and encouragement throughout my PhD journey. His thoughtful advice, patience, and belief in my work have been essential in helping me grow both academically and personally.

I would also like to express my sincere appreciation to Dr. Pelle Jakovits for kindly allowing me to conduct my experiments in his lab and for providing me with the space and tools I needed for my research. I am also deeply thankful to my dear and supportive friends in our lab for their encouragement and assistance.

In addition, I would like to extend my heartfelt thanks to the staff and members of the Institute of Computer Science at the University of Tartu for their unwavering support and kindness. From administrative help to academic advice and day-to-day assistance, their contributions have made a significant difference in my academic journey. Their dedication to creating a supportive and welcoming environment for students is something I will always be grateful for. Finally, I want to express my sincere gratitude to my family and friends for their constant support, patience, and encouragement throughout this endeavor. Their understanding, love, and belief in me have been a constant source of strength, without which this accomplishment would not have been possible.

# SISUKOKKUVÕTE

## Multimodaalne lähenemine kaardistamise ja lokaliseerimise täpsustamiseks, integreerides generatiivset tehisintellekti ja jalakäijatele-orienteeritud andmeid.

Tänapäeva linnad on elujõulised ja keerulised keskkonnad, kus kõnniteed, jalgrattateed, sõiduteed ja ülekäigurajad sulanduvad ühisteks avalikeks ruumideks. Kuna innovaatilised liikumislahendused, nagu pakirobotid ja elektrilised tõukerattad, jagavad jalakäijatega samu teid, on vajadus täpsete, detailsete ja pidevalt uuenevate kaartide järele suurem kui kunagi varem. Traditsioonilised kaardid on aga tavaliselt kujundatud autokesksete teede jaoks, jättes jalakäijatele mõeldud alad, nagu kõnniteed, ülekäigurajad ja jagatud ruumid, suhteliselt kaardistamata. See tähendab, et selliseid olulisi alasid ei uuendata tihti ja piisava täpsusega. Samas see info on vajalik uuenduslike liikuvustehnoloogiate ja kaasava linnaplaneerimise toetamiseks.

See doktoritöö lahendab seda probleemi, kasutades täiustatud multimodaalseid ja mitmesensorilisi andmekogumise meetodeid. Need meetodid võimaldavad oluliselt täiustada linnade kaardistamist ja lokaliseerimist. Selle asemel, et tugineda ainult satelliitpiltidele või aeganõudvale käsitsi uuendamisele, rakendatakse tehisintellekt (AI), et automatiseerida ja parandada kaardistamisprotsessi. Arendades uusi tehnikaid, loob see töö ainulaadseid kaardiperspektiive, ühendab efektiivselt erinevaid vaatenurki ja lihtsustab linnakaartide loomist.

Mitmekesiste reaalsete andmete kogumiseks linnakeskkonnas kohandas töö autor elektrilise tõukeratta, mis on varustatud täiustatud sensorite komplektiga. See hõlmab stereokaameraid, mis jäädvustavad detailset visuaalset infot, nagu värvid ja tekstuurid, LiDAR-andureid täpseks sügavus- ja kaugusmõõtmiseks, GPS-i täpseks asukohajälgimiseks ja audiosalvestit, mis jäädvustab keskkonnahelisid, nagu mööduvad sõidukid või ehitusmüra. See spetsiaalne liikuv platvorm võimaldab ulatuslikku ja tõhusat andmete kogumist just kõnniteedel ja teistele jalakäijatele mõeldud aladele. Kõik need multimodaalsed andmed on ühendatud spetsiaalsesse andmekogusse nimega DELTA, mis on pühendatud just jalakäijate radadele. Selle andmekogu põhjal töötas autor välja spetsialiseeritud AI-raamistiku nimega street2sat, mis teisendab tänavatasandi pildid satelliidivaadeteks. See lähenemine ühendab maapinna pilte õhuvaadetega, võimaldades täpsemat kaardistamist ja paremat lokaliseerimist, tuvastades ja ühendades olulisi linnamaastiku tunnuseid. Ühendades sujuvalt erinevaid vaateid, aitab street2sat luua kontekstiteadlikke kaarte, mis on ühtsemad ja kasulikumad erinevateks rakendusteks.

Eesmärgiga muuta kogutud andmed linnaplaneerimise ja navigatsioonisüsteemide jaoks praktilisemaks, töötas töö autor välja täiendava raamistiku nimega Street2GIS. See raamistik automatiseerib GIS-andmete loomise, teisendades sügavus- ja visuaalset infot standardsetesse GIS-formaatidesse, nagu *shapefile*'id, mis kaardistavad selgelt olulised linnamaastiku tunnused, nagu teed, kõnniteed, hooned ja tai-

mestik. Street2GIS ühendab sügavushinnangu, mis jäädvustab kaugusi ja ruumilisi seoseid, ning semantilise segmenteerimise, mis tuvastab ja märgistab stseeni erinevaid elemente. Integreerides need protsessid sujuvasse töövoogu, võimaldab Street2GIS tõhusalt luua täpseid GIS-andmeid, toetades paremat linnakeskkonna analüüsi ja otsuste tegemist.
See AI-põhine kaardistamislähenemine aitab kaasa ligipääsetavamate, reageerivamate ja kaasavamate linnakeskkondade loomisele. Parema ja detailsema teabega jalakäijate ruumide kohta saavad linna planeerijad ja poliitikakujundajad paremini kujundada infrastruktuuri, mis teenib kõigi linnaelanike, sealhulgas jalakäijate, jalgratturite ja liikumisraskustega inimeste mitmekesiseid vajadusi. Lisaks toetab see terviklik kaardistusstrateegia tõhusalt uuenduslikke linnatehnoloogiaid, nagu autonoomsed pakirobotid ja intelligentsed transpordisüsteemid, võimaldades neil turvalisemalt ja tõhusamalt navigeerida kaasaegsete linnade keerulises ja dünaamilises maastikus. Kuna linnakeskkonnad jätkavad arenemist ja kasvamist, on oluline võtta kasutusele paindlik, innovaatiline ja andmepõhine lähenemine jalakäijate alade kaardistamiseks. See tagab turvalisemad ja kaasavamad avalikud ruumid ning võimaldab linnadel tuleviku liikuvusuuendusi tõhusalt rakendada, edendades jätkusuutlikku linnakasvu, mis tuleb kasuks kõigile kogukonna liikmetele.

# CURRICULUM VITAE

## Personal data

| | |
|---|---|
| Name: | Alireza Akhavi Zadegan |
| Date of Birth: | 07.07.1987 |
| Citizenship: | Iranian |
| Contact: | aakhv110@gmail.com |

## Education

| | |
|---|---|
| 2022–2025 | University of Tartu, Doctoral degree, Computer Science |
| 2014–2018 | University Malaya, Master's degree, Machine learning and Computer Vision |
| 2010–2013 | Sheffield Hallam University, Bachelor's degree, Electrical and Electronics Engineering |

## Employment

| | |
|---|---|
| 2022–2025 | University of Tartu – Junior Research Fellow |

## Scientific work

Main fields of interest:

- Mechatronics
- Computer Vision
- Computer Graphics

# ELULOOKIRJELDUS

## Isikuandmed


| | |
|---|---|
| Nimi: | Alireza Akhavi Zadegan |
| Sünniaeg: | 07.07.1987 |
| Kodakondsus: | Iraan |
| E-post: | aakhv110@gmail.com |


## Haridus

| | |
|---|---|
| 2022–2025 | Tartu Ülikool, doktorikraad arvutiteaduses |
| 2014–2018 | Malaya Ülikool, magistrikraad masinõppes ja arvutinägemises |
| 2010–2013 | Sheffield Hallami Ülikool, bakalaureusekraad elektroonikas ja elektroonika inseneerias |

## Teenistuskäik

| | |
|---|---|
| 2022–2025 | Tartu Ülikool – Nooremteadur |

## Teadustegevus

Peamised uurimisvaldkonnad:

- Mehhatroonika
- Arvutinägemine
- Arvutigraafika

# DISSERTATIONES INFORMATICAE PREVIOUSLY PUBLISHED IN DISSERTATIONES MATHEMATICAE UNIVERSITATIS TARTUENSIS

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω-rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo**. Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Šor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

# DISSERTATIONES INFORMATICAE UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh**. Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas**. Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi**. Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich**. Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka**. Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinemaa**. Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto**. Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere**. Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.
58. **Kirill Milintsevich.** Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity. Tartu 2024, 130 p.
59. **Maksym Del.** Multilingual and Multi-Domain Representational Patterns Across Trpansformer-Based Models. Tartu 2024, 131 p.
60. **Kristo Raun.** Adaptive Out-of-order Handling in Streaming Conformance Checking. Tartu 2024, 118 p.
61. **Toivo Vajakas.** Towards integration of mobile network data into analyzing human mobility. Tartu 2024, 103 p.
62. **Katsiaryna Lashkevich.** Data-Driven Analysis and Optimization of Waiting Times in Business Processes. Tartu 2024, 169 p.
63. **Alejandra Duque-Torres.** Classifying, Constraining and Ranking Metamorphic Relations. Tartu 2025, 159 p.

64. **Mariia Bakhtina.** A Method for Information Security and Privacy Management in Smart Solutions. Tartu 2025, 199 p.
65. **Andre Tättar.** Multilingual Machine Translation for Under-Resourced Languages. Tartu 2025, 170 p.
66. **Mahmoud Shoush.** Prescriptive Process Monitoring Under Uncertainty and Resource Constraints. Tartu 2025, 178 p.