# Geopose-enabled Camera Imagery Interoperability with Geo-AI in Urban Digital Twins

Kalp Devangbhai Thakkar[1], Kyle Shervington[1], Soheil Sabri[1], Benjamin Lee[1]

[1] Urban Digital Twin Lab at SMST, University of Central Florida, USA
(kalpdevangbhai.thakkar, kyle.shervington, soheil.sabri, benjamin.lee)@ucf.edu

**Keywords:** GeoPose, TrainingDML-AI, Urban Digital Twin, GeoAI, Camera Imagery Interoperability, Machine Learning.

## Abstract

This paper presents a GeoPose-enabled pipeline designed to enhance camera imagery and Inertial Navigation System (INS) data interoperability within Urban Digital Twin (UDT) systems. It addresses critical challenges in data synchronization, georeferencing, and integration by leveraging low-cost tools and open standards. The proposed framework captures, processes, and aligns visual and spatial data, converting them into GeoPose and TrainingDML-AI formats to support advanced Geo-AI applications. This methodology enables seamless integration of heterogeneous datasets, facilitating machine learning tasks such as image-based object detection and geospatial analysis. Key contributions include a scalable and cost-effective solution for integrating urban data, ensuring consistency and accessibility across platforms. By advancing the capabilities of UDT systems, this work provides a standardized foundation for real-time decision-making and enhanced urban analytics, promoting smarter and more efficient management of urban spaces, infrastructure, and resources in rapidly evolving smart cities.

## 1. Introduction

In the rapidly evolving field of smart city development, Urban Digital Twins (UDT) are becoming a critical tool for data-driven urban planning, management, and decision-making (Sabri & Witte, 2023). These digital replicas of real-world urban environments enable simulations and analyses of various city infrastructures, such as traffic, environmental conditions, and public services to support urban planners and policymakers (Sabri et al., 2022). However, a persistent challenge in fully realizing the potential of UDTs lies in the interoperability of disparate geospatial datasets, such as camera imagery, and other sensor data. Ensuring seamless integration of these data sources is crucial to creating accurate, comprehensive models that reflect the real-world dynamics of urban environments.

This paper addresses the research questions centered on the capturing, integration, and interoperability of camera imagery, and geospatial metadata for performing machine learning (ML)-driven analysis within UDT frameworks. Specifically, what are the current challenges in capturing camera imagery, position, and orientation data?. How can data from various sources be synchronized and standardized to provide accurate spatial and temporal representations of urban environments? How can the integrated data be used effectively for real-time object detection and other Geo-AI tasks in smart cities?

To address these questions, we have developed a data acquisition, processing, and Geo-AI pipeline that ensures camera imagery and INS metadata are captured, synchronized, and converted into GeoPose. The pipeline extends beyond data collection, enabling ML tasks through TrainingDML-AI conversion and automatic labeling annotations. By adhering to Open Geospatial Consortium (OGC) standards like GeoPose and TrainingDML-AI, our methods lay the groundwork for better interoperability in the UDT systems.

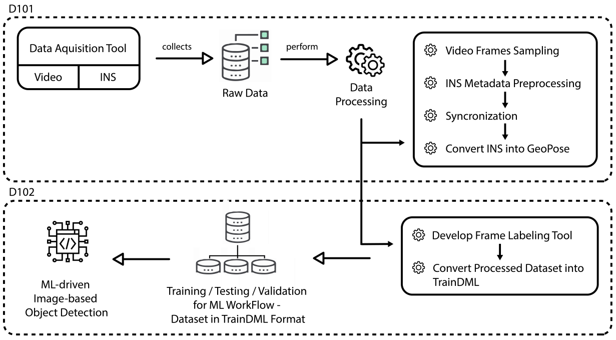

D101
Data Aquisition Tool
Video
INS
collects
Raw Data
perform
Data Processing
Video Frames Sampling
INS Metadata Preprocessing
Syncronization
Convert INS into GeoPose
D102
Develop Frame Labeling Tool
Convert Processed Dataset into TrainDML
Training / Testing / Validation for ML WorkFlow - Dataset in TrainDML Format
ML-driven Image-based Object Detection

Figure 1. Roadmap: A strategic outline detailing the integration of Camera Imagery Interoperability and Geo-AI Analysis Interoperability within UDT systems.

### 1.1 Background and Motivation

UDT have emerged as a promising solution for managing and optimizing complex urban systems by providing a spatially accurate, real-time, and dynamic view of cities (Xu et al., 2024). These digital models are enriched with data from multiple sources, such as traffic sensors, environmental sensors, and camera imagery, providing a holistic understanding of urban landscapes (Aghaabbasi & Sabri, 2025). However, as cities grow more connected, the integration of heterogeneous data sources becomes increasingly complex (Costagliola et al., 2024).

The motivation for this work stems from the need to enhance the interoperability of geospatial data in UDTs, especially data derived from camera imagery and INS systems. While these data sources can provide rich visual and spatial information, they often lack the necessary alignment in terms of time, location, and orientation to be useful for real-time decision-making. Moreover, our investigation into existing tools revealed a lack of open-source solutions capable of capturing and synchronizing camera imagery with fused GNSS/IMU data from mobile devices. This gap is particularly challenging for organizations and researchers constrained by limited budgets. The alternative investing in high-cost INS systems and professional-grade cameras could demand tens of thousands of dollars, making these solutions impractical for low-cost, scalable urban data collection.

The motivation for this work lies in developing a cost-effective pipeline that can reliably capture synchronized camera imagery and INS data using widely accessible tools. By leveraging smartphone-based data acquisition and open-source applications, we aim to create a pipeline that allows for robust data collection while minimizing costs. This approach not only supports the democratization of UDT technology but also enables broader adoption in cities with limited resources.

### 1.2 Scope and Objectives

This paper focuses on developing a comprehensive GeoPose-enabled pipeline for camera imagery and INS data interoperability, a critical aspect of advancing UDT systems. The proposed scope covers the end-to-end workflow, including data acquisition, GeoPose conversion, annotation, and ML dataset preparation in the TrainingDML-AI format. The framework is designed to ensure seamless integration of camera imagery, spatial metadata, and ML models, supporting enhanced urban analytics with the potential to facilitate real-time decision-making in smart cities.

The objectives of this work are:

- To design an interoperability pipeline that seamlessly integrates camera imagery and INS data into UDT systems.
- To ensure data synchronization and processing while converting it into GeoPose format, facilitating georeferencing and spatial analysis.
- To establish an annotation framework for preparing TrainingDML-AI datasets, enabling ML tasks like object detection and scene understanding.
- To evaluate the pipeline’s effectiveness through case studies involving real-time data collection, GeoPose generation, and AI-driven object detection in urban environments.

By achieving these objectives, this work aims to provide a scalable, cost-effective solution for data interoperability and GeoAI applications in UDTs, setting the stage for future research and innovation in UDT technologies.

## 2. State of the Art

### 2.1 Related Works

Recent advancements in UDTs have emphasized the critical role of integrating geospatial metadata and camera images for improved urban modeling and analysis (Jeddoub et al., 2024). GeoPose, a standard developed by the Open Geospatial Consortium (OGC), has emerged as a crucial framework that enables a consistent representation of the position and orientation of physical objects or virtual entities in 3D space. By facilitating interoperability between diverse systems, GeoPose supports real-time applications such as infrastructure monitoring, autonomous navigation, and immersive urban simulations. However, its adoption in dynamic data streams remains limited, with synchronization challenges between high-frequency imagery and sensor data yet fully addressed.

Research in multi-sensor data fusion, leveraging techniques such as Kalman Filtering and GNSS/IMU integration, has significantly improved the accuracy of geospatial data (Wang & Li, 2018). Many datasets demonstrate the potential of synchronized sensor data for tasks including object detection and scene reconstruction. Meanwhile, advancements in GeoAI have shown promise for automating annotation and enabling predictive urban analytics, further expanding the capabilities of ML in UDT contexts.

Applications of camera imagery interoperability in UDTs span real-time traffic analysis, urban infrastructure maintenance, and disaster management. However, existing implementations often rely on costly, proprietary hardware, limiting scalability and adoption in resource-constrained environments. Addressing these challenges through cost-effective, standardized solutions remains a key research focus in the domain.

### 2.2 GeoPose and Imagery

GeoPose encapsulates the position and orientation information of a camera or sensor in a well-defined global or local coordinate system, typically by encoding latitude, longitude, and altitude for positioning, along with angle or quaternion encoding scheme for orientation. This structured approach allows for precise georeferencing of imagery data captured by mobile or stationary devices, such as drones or street-level cameras. Through this geospatial encoding, GeoPose supports alignment with high-fidelity 3D models, enhancing the integration of sensor data across multiple platforms and facilitating consistent visualization within digital twin systems (Clarke et al., 2024).

The adoption of GeoPose enables accurate geospatial representation for imagery interoperability, addressing the challenge of synchronizing heterogeneous datasets from disparate sources. Within UDTs, this is particularly relevant for real-time applications such as traffic monitoring, infrastructure management, and autonomous navigation, where precise alignment of spatial and visual data is critical. The use of GeoPose has advanced applications in photogrammetry, 3D object recognition, and augmented reality, providing a standardized way to overlay imagery onto real-world coordinates.

However, despite the capabilities of GeoPose, there are limitations in its application, especially in scenarios requiring synchronization between high-frequency imagery and positioning data. This synchronization challenge is further exacerbated in dynamic environments, where consistent timestamp alignment across devices is crucial. Additionally, the absence of a direct data standard for pairing GeoPose metadata with imagery captured at various frame rates remains a significant gap, impacting the real-time utility of GeoPose within ML workflows.

### 2.3 TrainingDML-AI

TrainingDML-AI aims to create a unified modeling language (UML) and encodings for geospatial ML training data. This is crucial because training data plays a fundamental role in Earth Observation (EO) AI/ML, especially in Deep Learning (DL), where it is used to train, validate, and test AI/ML models. The standard defines a UML model and encodings that are consistent with OGC Standards to facilitate the exchange and retrieval of training data in web environments. TrainingDML aims to improve the interoperability and reusability of training data, which is essential for developing accurate and reliable AI models (Yue & Shangguan, 2023).

### 2.4 Limitations in Current Approaches

Current approaches to camera imagery interoperability and geospatial metadata integration in UDT systems face several limitations that hinder their scalability and effectiveness. One critical challenge lies in the lack of cost-effective, open-source solutions for synchronizing high-frequency camera imagery with geospatial metadata such as GNSS/INS data. Existing systems often rely on proprietary hardware and software, which, while accurate, are prohibitively expensive and inaccessible for widespread adoption, especially in resource-constrained urban settings.

Another limitation is the inconsistency in data synchronization across diverse sensor systems. Many tools fail to align imagery and metadata with the required temporal and spatial precision, leading to inaccuracies in georeferencing and subsequent analyses. Furthermore, the absence of robust standards for integrating camera imagery and metadata at scale complicates interoperability, reducing the utility of data for machine-learning applications such as object detection and urban scene reconstruction (Weil et al., 2023).

Many existing standards for geospatial data encoding do not compete directly with TrainingDML-AI but offer complementary solutions. SpatioTemporal Asset Catalog (STAC) (Rustad et al., 2023) and the OGC O&M (Randall & Antonisse, 2012) standards are extensible and standards-based and may be useful for creating and handling geospatial data. Motion Imagery Standards Board (MISB) standards provide data and metadata related to motion imagery tracking (Randall & Antonisse, 2012). The Coalition Shared Data (CSD) is a framework for sharing military information. It is a mature approach for sharing Intelligence, Surveillance, and Reconnaissance (ISR) data (Rustad et al., 2023). None of the mentioned standards are designed to provide a markup language for training data for AI models to the same level of robustness as TrainingDML-AI.

Furthermore, existing pipelines often overlook real-time processing requirements, making them unsuitable for dynamic urban environments where immediate insights are critical. They also lack automated frameworks for annotation and standard conversion which are essential for preparing datasets for GeoAI-driven applications. These limitations underscore the need for our proposed cost-efficient, standardized pipeline for reliable data integration and enhanced ML capabilities in UDTs.

## 3. Methodology

### 3.1 Data Collection

**3.1.1 Requirements for Data Capturing Tool** The requirements for a data-capturing tool are driven by a need for accuracy, compatibility, and flexibility, ensuring seamless integration of visual and sensor data for effective ML and geospatial analysis. To achieve these goals, the selected tool must meet certain key criteria related to interoperability, cost-effectiveness, sensor support, data types, compatibility, frequency and accuracy, synchronization, export formats, usability, and processing support.

Interoperability is a priority, enabling the tool to integrate with multiple systems and standards, allowing data from various sources to be seamlessly shared or exchanged with minimal reformatting.

In terms of sensor support, the tool should capture data from several key sensors, including the camera (for video or time-stamped images), accelerometer, gyroscope, magnetometer, and GPS/GNSS for position tracking. Importantly, it should capture both raw and calibrated position and orientation data, where calibrated data are especially essential for accurate georeferencing.

Given the critical role of positioning and orientation accuracy, the tool should provide the fused data outputs in INS format. This method is built on top of other data acquisition techniques like Inertial Measurement Unit (IMU) as well as Attitude and Heading Reference Systems (AHRS), each offering increasingly refined levels of calibration and position tracking. This gradation can be conceptualized as $IMU \subset AHRS \subset INS$

Here, an IMU provides raw measurements (acceleration and rotation), an AHRS incorporates tilt and heading estimates, and an INS combines these with position data through algorithms like the Kalman filter (Sasani et al., 2016). The INS's double integration of acceleration data, for instance, is key to mitigating drift and achieving precise displacement estimation. The formula for double integration, showing how the position is derived from acceleration is given by:

$$p = \int \left( \int a\,dt \right) dt \tag{1}$$

where $p$ = position of the object
$a$ = acceleration measured by the accelerometer
$t$ = time

However, errors in accelerometer readings can accumulate due to drift, leading to quadratic error growth over time. By employing a Kalman filter to fuse external data sources (e.g., GPS) with IMU data, the INS corrects for drift and stabilizes position estimates, thus making it suitable for long-term tracking and accurate geospatial alignment.

The compatibility of the tool with both Android and iOS platforms is necessary for widespread usability. Furthermore, sampling frequency and positional accuracy are essential: data must be captured at a high frequency to allow downscaling during post-processing, as upscaling would introduce error. Synchronization capabilities are also crucial to ensure all sensor data and imagery align temporally, thus supporting integrated analysis.

The tool should facilitate data export in standard formats (e.g., JSON, CSV for sensor data, and JPG/PNG/MP4 for imagery) to ensure compatibility with downstream processing. Annotation capabilities and customization options would allow users to label data for ML and adapt settings to specific needs. Additional features, such as real-time data streaming and visualization, would enable live monitoring, while post-processing support ensures that data can be refined and calibrated, enhancing accuracy and reliability.

**3.1.2 Tools for Camera Imaginary and Metadata Acquisition** To capture high-quality visual data and precise geospatial metadata, we employed a combination of accessible yet powerful tools designed for mobile platforms. Our primary tool, Sensor Logger, supports the capture of camera imagery (in video or time-stamped image format) alongside metadata from accelerometers, gyroscopes, magnetometers, and GPS. This setup enables partially synchronized acquisition of imagery and position/orientation data necessary for georeferencing. To enhance flexibility, additional data sources, including open datasets like Hillyfields and KITTI-360 (Liao et al., 2021), provided robust supplementary imagery and geospatial metadata. Each tool and dataset was selected based on its compatibility with our requirements for data interoperability, cost-efficiency, and accuracy, ensuring comprehensive data acquisition to support ML applications in UDTs. Sample images reflecting various conditions and road environments can be referenced in Figure 2, which illustrates the diverse textures and settings encountered during collection.

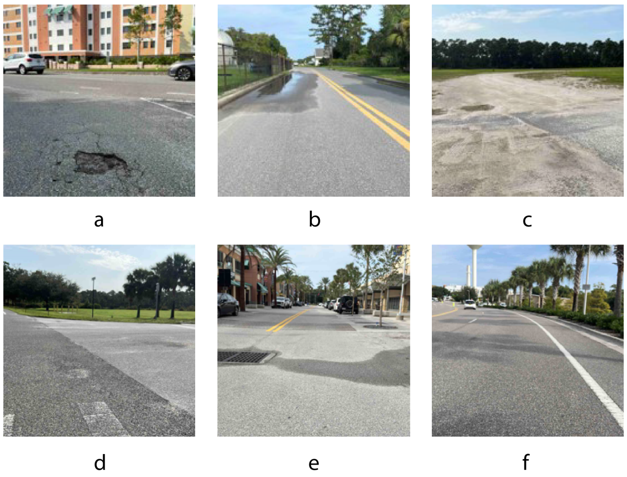


Figure 2. Sample Images of Various Road Covers and Conditions Captured from UCF Data. a: Potholes on a busy road, b: Road with water puddles, c: Gravel road with loose stones, d: Two road textures at turning region, e: Gutter with wet area, and f: Smooth, newly paved road

| Dataset | Description | Rate |
|---|---|---|
| UCF Data | Images: 1798 samples<br>INS: 1805 samples | 1 Hz<br>1 Hz |
| Hillyfields: Run 3 | Video: 05:35 mm:ss<br>INS: 33601 samples | 4.91 fps<br>100 fps |
| Kitti-360 | Images: 320k<br>INS: 4x83,000 | 10 fps<br>10 fps |

Table 1. Technical Specification and Comparison of Datasets

**3.1.3 Equipment and Setup** To ensure accurate data acquisition for our study, we used a smartphone equipped with the Sensor Logger application for capturing camera imagery along with calibrated position and orientation metadata.

**Spatial Shift Rate**

We introduced a metric called Spatial Shift Rate ($SSR$) which quantifies the rate of spatial displacement per unit of time between consecutive data points based on spatial and temporal alignment. For two consecutive frames with coordinates $P_1$ and $P_2$, and image intervals defined by time $t_1$ and $t_2$, we used the following metric:

$$SSR = \frac{\|P_1 - P_2\|}{\Delta t} \tag{2}$$

where $SSR$ = Spatial Shift Rate
$P_1, P_2$ = positional coordinates
$\|P_1 - P_2\|$ = euclidean distance
$\Delta t = t_1 - t_2$ = time interval

The Spatial Shift Rate ($SSR$) (in meters per second) provides insight into the relative positional change rate between consecutive captures. A lower $SSR$ value implies higher spatial overlap between frames, which is ideal for applications that require continuity (e.g., real-time monitoring in UDT). Conversely, a higher $SSR$ value suggests larger gaps, potentially leading to information loss or decreased continuity in the data stream, which may hinder precise analysis or model training.

**Overlap Metric Calculation**

To quantify the data overlap between GeoPose metadata (position and orientation) and imagery frames, we calculate the temporal alignment based on timestamps and compute a synchronization offset. This overlap metric ensures accurate georeferencing of imagery frames with GeoPose metadata in UDT applications.

Assuming that GeoPose metadata ($G$) and imagery frames ($I$) have different recording frequencies ($f_G$ and $f_I$, respectively), we can calculate the overlap factor, which represents the percentage of frames in $I$ that have corresponding GeoPose metadata in $G$ within a defined tolerance threshold as follows:

$$\Omega = \frac{1}{N_I} \sum_{m=1}^{N_I} \left[ \min_n \left( |t_{I_m} - t_{G_n}| \right) < \Delta t \right] \tag{3}$$

where $\Omega$ = overlap factor
$N_I$ = total number of imagery frames
$t_{I_m}$ = timestamp of the $m$-th imagery frame
$t_{G_n}$ = timestamp of the $n$-th GeoPose sample
$\Delta t$ = tolerance threshold

The indicator function returns 1 if there exists a GeoPose timestamp $t_{G_n}$ such that the absolute time difference $|t_{I_m} - t_{G_n}|$ is within the tolerance threshold $\Delta t$, indicating a match. This formula gives the proportion of imagery frames that have a matching GeoPose metadata sample within the desired temporal accuracy.

**Synchronization Error Calculation**

For successful synchronization, we want the average $\Omega$ to be less than or equal to a defined tolerance threshold $\Delta t$ that is $\Omega \leq \Delta t$ which ensures that each camera frame is matched to an INS metadata sample within a permissible time interval.

Achieving a low $\Omega$ value within the $\Delta t$ threshold confirms that camera frames are well-aligned with INS metadata. This accurate synchronization is essential for aligning visual data with spatial orientation and position data, enabling reliable geospatial analysis in UDT models.

**3.1.4 Camera Imagery Specifications** The camera imagery specifications are crucial for ensuring data quality and compatibility with subsequent processing steps, including image labeling and ML tasks. The captured imagery in the our dataset is in RGB (Red, Green, Blue) format, providing color-rich data that is valuable for diverse analysis tasks in UDT systems. Each image has a resolution of 1080x1080 pixels with a 1.2-megapixel quality, producing files in the JPEG format.

The image resolution is set to 72 dpi for both X and Y axes, with a Resolution Unit of inches, providing a balance between image clarity and manageable file sizes. Encoding Process is achieved through Baseline Discrete Cosine Transform (DCT) with Huffman coding, and 8 bits per sample for efficient compression and preservation of image quality. This setup provides high fidelity in imagery, suitable for detecting fine details in urban environments, especially under varying lighting conditions and backgrounds. The specified Capture Frequency of 1 Hz (one frame per second) ensures optimal data continuity without excessive redundancy.

To achieve effective spatial coverage, the capture frequency ($f$) is determined based on the following formula:

$$f(Hz) = \frac{\text{imageCaptureDistance (meters)} \times 3600}{\text{averageSpeed (mph)} \times 1609.344} \tag{4}$$

This formula allows for the adjustment of the image capture frequency according to vehicle speed and desired frame overlap, optimizing image capture for data continuity while balancing storage and processing needs.

It’s always best to capture imagery data at high frequency, as it can be scaled down through post-processing. Additionally, the position and orientation data should be captured at a frequency that matches or exceeds that of the imagery, especially in non-video formats, to ensure proper synchronization and alignment with the collected imagery data.

## 3.2 Data Processing Pipeline

**3.2.1 Frame Extraction Using FFMPEG and Image Sampling** In the case of video imagery, we utilized FFMPEG, a popular open-source command-line toolbox, to manipulate, convert, and stream multimedia content. Specifically, FFMPEG was employed for video frame sampling, which involves extracting frames from a video at regular intervals or a predefined frame rate.

The goal of this operation is to extract video frames at a specific rate, which can be crucial for tasks like creating training datasets, analyzing video content, or synchronizing with other data sources such as INS metadata.

**3.2.2 INS Data Preprocessing** In the INS Data Preprocessing stage, the location and orientation metadata files are systematically prepared for further analysis and integration. Initially, we define the file paths, determining whether the location and orientation data are stored in separate files. If they are separate, each file is fetched from its respective URL and loaded into pandas DataFrames for processing. In cases where the metadata is consolidated into a single file, this combined file is directly read into a DataFrame.

A critical step in our preprocessing workflow is converting epoch timestamps into a human-readable UTC date-time format. This is achieved by a function that extracts the first ten digits of the epoch time and transforms them into a structured representation of date and time.

After conversion, we deduplicate the entries to eliminate any redundant timestamps in both the location and orientation datasets. This step is essential to ensure that our analysis is not skewed by repeated measurements.

Following the deduplication process, we extract relevant features from the DataFrames. For location data, we retain the necessary columns, such as latitude, longitude, and altitude. For orientation data, the extraction method varies based on the chosen encoding. If quaternion representation is used, the corresponding quaternion components are retained. Conversely, if the orientation is expressed in angles (yaw, pitch, roll), a conversion from radians to degrees is applied, as shown below:

$$\text{Degrees} = \text{Radians} \times \frac{180}{\pi} \tag{5}$$

Subsequently, the location and orientation DataFrames are merged on the common timestamp, resulting in a comprehensive dataset that encapsulates both sets of information. During this merging process, we drop any columns that might have been duplicated, such as old timestamp columns, to maintain clarity in our data structure. Any rows containing NULL values are also removed to ensure the integrity of the dataset.

A further critical transformation occurs when converting the UTC timestamps into GPS time, which is vital for applications requiring precise temporal alignment. This transformation utilizes a function that calculates the difference between UTC and GPS epochs while accounting for the current leap second difference:

$$\text{GPS Time} = (\text{UTC Time} - \text{GPS Epoch}) + \text{LEAP SECONDS} \tag{6}$$

The GPS epoch is defined as January 6, 1980, while the leap seconds, currently 18, must be subtracted to yield the accurate GPS time representation. Each combined row of date and time is processed to compute the corresponding GPS time.

Finally, to refine the dataset further, we implement a minification technique where unnecessary columns are systematically dropped, streamlining the DataFrame for subsequent analysis. This comprehensive preprocessing pipeline ensures that the INS metadata is accurately calibrated, filtered, and structured, paving the way for effective utilization in further processing and analysis stages.

**3.2.3 Synchronization of Camera Imagery and Geospatial Metadata** In this step, video frames or image instances are synchronized with their corresponding position and orientation metadata, ensuring temporal alignment between visual and geospatial data (Erdnuess, 2020). For video data, firstly the frames are extracted at a predefined rate, such as every second, to create an evenly spaced frame sequence. The number of frames is used to determine the number of corresponding INS samples, which are selected at intervals matching the frame count, ensuring a synchronized set of location and orientation data for each frame. This process is formalized by calculating an extraction interval as:

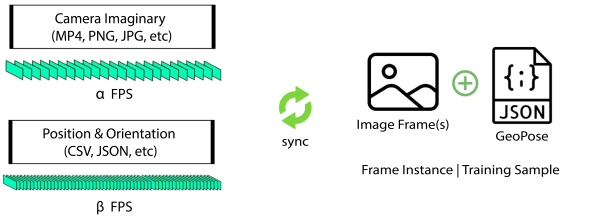


Figure 3. Synchronization of Camera Imagery and INS Metadata

$$\text{Interval} = \max\left(1, \frac{\text{Total Samples}}{\text{Desired Samples}}\right) \tag{7}$$

If the imagery is in a timestamped image format, a DataFrame of timestamps is created from the filenames and joins this data with the Geospatial metadata by matching the date and time fields. This synchronization technique relies on accurately converting epoch time to UTC and, if necessary, adjusting timestamps to GPS time to ensure precision.

This combined dataset of synchronized frames and geospatial metadata enables accurate geo-referencing of imagery, associating each visual instance with precise spatial orientation and location, providing a reliable foundation for applications requiring accurate positional alignment.

### 3.3 GeoPose Conversion

**3.3.1 GeoPose Standard Overview** The GeoPose standard enables precise geospatial positioning by encoding the orientation and position of objects within a global reference frame (Smyth, 2022). Developed to support interoperability in geospatial applications, GeoPose facilitates seamless geological referencing of various data types, including camera imagery, and sensor readings into UDT systems.

The **GeoPose.Composite.Sequence.Series.Regular**, one of the eight standardization targets defined by GeoPose, is utilized to provide a structured encoding framework for positional (latitude, longitude, altitude) and orientation (angle or quaternion) data. This specification allows for the sequential, time-regular encoding of geospatial metadata, ensuring synchronized and temporally consistent alignment with imagery data. Data accuracy and interoperability across various devices and applications are enhanced by this standard, making it a reliable framework for tasks such as real-time object tracking and spatial analysis.

**3.3.2 INS to GeoPose Conversion** The INS-to-GeoPose conversion process translates raw Inertial Navigation System (INS) data into a standardized GeoPose format, accurately representing the spatiotemporal positioning of objects. The process begins by defining a reference point based on the first entry in the dataset, establishing the latitude, longitude, and altitude.

The algorithm calculates the inter-pose duration $T_d$ by determining the median time difference between consecutive GPS timestamps. This is expressed as:

$$T_d = \text{median}\left(t_{n+1} - t_n\right) \quad \text{for} \quad n = 0, 1, \ldots, N-1 \tag{8}$$

where $t_n$ = GPS timestamps
$N$ = total number of samples

This ensures the system can consistently interpret pose intervals for accurate temporal alignment.

To ensure data integrity, a SHA256 checksum is generated based on the concatenated pose data, which includes either angular or quaternion orientation encodings, depending on the specified parameter. This checksum serves as a verification measure for the accuracy and consistency of the pose data throughout the conversion process. The GeoPose structure encapsulates crucial temporal information, such as the start and stop instants $t_{\text{start}}$ and $t_{\text{stop}}$ derived from the GPS timestamps, along with the total pose count $P$, facilitating the tracking of data over time.

The conversion defines an outer frame using the reference latitude, longitude, and altitude, while inner frame series entries are created for each pose, detailing the translation and rotation parameters. The translation is defined as $[E, N, U]$, where $E$, $N$, and $U$ represent East, North, and Up coordinates, respectively. For rotation, either Euler angles $[\psi, \theta, \phi]$ (yaw, pitch, roll) or quaternion values $[q_x, q_y, q_z, q_w]$ are utilized based on the orientation encoding.

The final GeoPose JSON structure encompasses a comprehensive header with metadata, including pose count $P$ and integrity checks, along with an inter-pose duration $T_d$, outer frame data, and a series trailer, collectively ensuring a robust and interoperable representation of the positional data within the GeoPose framework. The GeoPose file creation process was discussed in the previous section.

### 3.4 Image Annotation and TrainingDML-AI

This research utilized a two-step process to prepare geospatial image data for ML. First, 42 images representing diverse road surface types were manually annotated using Computer Vision Annotation Tool (CVAT). Polygons were drawn around target features, and annotations were exported in COCO 1.0 format. Second, these annotations, along with GeoPose-encoded data and provenance metadata, were aggregated into a single TrainingDML-AI file. Each image with its corresponding geospatial metadata via python automation, ensuring all necessary information for model training was included. The resulting TrainingDML-AI file, validated for compliance, provides a structured and interoperable dataset for UDT applications.

## 4. Results and Validation

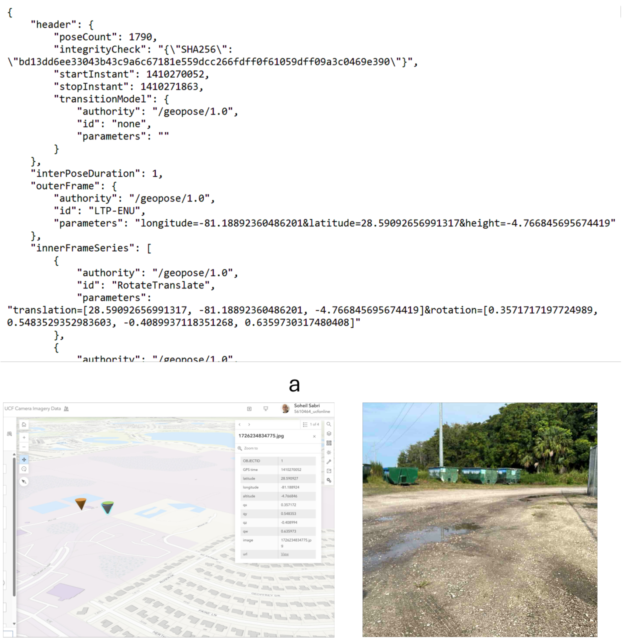
### 4.1 Results from GeoPose

The resulting GeoPose JSON file accurately encodes positional and orientation data, ensuring seamless interoperability for geospatial analysis. Figure 4 showcases a sample of the generated GeoPose JSON structure from the UCF dataset, which is then visualized and retrieved on ArcGIS online. The output contains the precise geospatial representation and timestamp alignment.

### 4.2 Synchronization Validation

To validate synchronization, timestamp alignment between camera imagery and INS geospatial metadata is evaluated.

As detailed in Equation 3 (Overlap Metric Calculation), the overlap factor $\Omega = 1.0$ in our dataset indicates perfect synchronization with a zero tolerance threshold $\Delta t$ = 0, ensuring



```
{
    "header": {
        "poseCount": 1790,
        "integrityCheck": "{\"SHA256\":
\"bd13dd6ee33043b43c9a6c67181e559dcc266fdff0f61059dff09a3c0469e390\"}",
        "startInstant": 1410270052,
        "stopInstant": 1410271863,
        "transitionModel": {
            "authority": "/geopose/1.0",
            "id": "none",
            "parameters": ""
        }
    },
    "interPoseDuration": 1,
    "outerFrame": {
        "authority": "/geopose/1.0",
        "id": "LTP-ENU",
        "parameters": "longitude=-81.18892360486201&latitude=28.59092656991317&height=-4.766845695674419"
    },
    "innerFrameSeries": [
        {
            "authority": "/geopose/1.0",
            "id": "RotateTranslate",
            "parameters":
"translation=[28.59092656991317, -81.18892360486201, -4.766845695674419]&rotation=[0.3571717197724989,
0.5483529352983603, -0.4089937118351268, 0.6359730317480408]"
        },
        {
            "authority": "/geopose/1.0",
```

1726234834775.jpg

| | |
|---|---|
| OBJECTID | 1 |
| GPS time | 1410270052 |
| latitude | 28.590927 |
| longitude | -81.188924 |
| altitude | -4.766846 |
| qx | 0.357172 |
| qy | 0.548353 |
| qz | -0.408994 |
| qw | 0.635973 |
| image | 1726234834775.jpg |
| url | View |

Figure 4. Output of GeoPose synchronization. a: Sample GeoPose JSON Structure from UCF Dataset, b: Visualized data on ArcGIS Online, and c: Retrieved the GeoPosed Image

that every imagery frame has an exact GeoPose timestamp since they both were captured using the same device. The $t_{G_n}$ represents $P_{estimated,i}$ which is the estimated INS position after GeoPose conversion, rather than the ground truth.

$$\Omega = \frac{1790}{1790} = 1.0 \quad \text{or} \quad 100\%$$

Here, 1790 are the number of datapoints in the our dataset. The results validate that the dataset maintains a precise temporal match between imagery frames and INS metadata, ensuring high-fidelity georeferencing crucial for accurate spatial analysis, real-time mapping, and ML applications.

### 4.3 Image Annotation Efficiency

The annotation process efficiency was benchmarked by comparing manual and semi-automated methods within CVAT. Key metrics included Annotation Time per Frame and Annotation Accuracy, with efficiency calculated as:

$$\text{Annotation Efficiency} = \frac{\text{Total Frames Annotated}}{\text{Total Annotation Time}} \tag{9}$$

| Frame | Manual (seconds) | Semi-Auto (seconds) |
|---|---|---|
| 1 | 17.12 | 10.00 |
| 2 | 15.10 | 11.01 |
| 3 | 14.54 | 9.94 |
| 4 | 15.39 | 9.45 |
| 5 | 13.08 | 10.01 |
| **Average** | **15.05** | **10.08** |

Table 2. Comparison of Time (seconds) for Manual and Semi-Auto Annotation

Table 2 shows that automated annotations significantly reduced time per frame while maintaining high accuracy, streamlining data preparation and enabling scalable ML tasks.

## 5. Discussion

### 5.1 Strengths and Contributions

The key strength of this work is establishing a cost-effective, GeoPose-enabled data acquisition and processing pipeline for UDTs. This project leverages accessible tools and open-source solutions to successfully implement a robust workflow integrating camera imagery with positional and orientation metadata, enhancing interoperability in smart city models. The conversion reliably encoded positional and orientation data, aligning with the GeoPose.Composite.Sequence.Series.Regular standardization target, proving effective interoperability across datasets. It also enables high-quality georeferenced data collection, synchronization, and conversion into TrainingDML-AI, supporting scalable ML tasks. Coupled with with real-time analytics, the GeoPose framework’s adaptability across various use cases significantly advances UDT systems’ capabilities in data sharing, cross-platform compatibility, and accurate spatial referencing for further analysis and ML workflows.

## 6. Conclusion and Future Work

By developing a robust pipeline for capturing, synchronizing, and converting spatial and visual data into the GeoPose standard, this work establishes a foundational framework for enabling seamless data integration within UDT systems. The interoperability achieved here facilitates accurate georeferencing and alignment of heterogeneous datasets, ensuring reliable inputs for downstream Geo-AI applications. These advancements underscore the importance of standardized, cost-effective solutions in advancing UDT capabilities, particularly in resource-constrained urban contexts.

### Dataset Enhancements and Validation Approach

Conversion precision can be quantified by calculating positional error. However, our dataset lacks ground truth positional data, preventing direct calculation of GeoPose conversion error. However, future validation efforts could incorporate reference datasets with known ground truth positions to further evaluate positional drift and refine transformation accuracy.

### Real-Time Data Acquisition and Processing

Expanding this work toward real-time data acquisition and processing remains a critical goal. Future directions include adopting stream processing frameworks and edge computing techniques to enable real-time synchronization, georeferencing, and analysis. These advancements will enhance the utility of UDT systems in dynamic urban environments, offering actionable insights for real time city management and decision-making.

## References

Aghaabbasi, M., & Sabri, S. (2025). Potentials of digital twin system for analyzing travel behavior decisions. *Travel Behaviour and Society*, *38*, 100902. https://doi.org/10.1016/j.tbs.2024.100902

Clarke, J., Smyth, S., Smith, R., & Morley, J. (2024). Creating a 3D Multi-Dataset Bubble in Support of OGC Testbed-19 and Metaverse Standards Prototypes [Series Title: Lecture Notes in Geoinformation and Cartography]. In T. H. Kolbe, A. Donaubauer & C. Beil (Eds.), *Recent Advances in 3D Geoinformation Science* (pp. 155–167). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-43699-4_10

Costagliola, A. R., Sabbioni, A., Bujari, A., Montanari, R., & Bellavista, P. (2024). A Multi-faceted Interoperability Model for Reliable and Trustworthy Urban Digital Twins. *Proceedings of the 2024 International Conference on Information Technology for Social Good*, 373–376. https://doi.org/10.1145/3677525.3678684

Erdnuess, B. (2020, May). Proper synchronization of geospatial metadata in motion imagery and its evaluation. In K. Palaniappan, G. Seetharaman, P. J. Doucette & J. D. Harguess (Eds.), *Geospatial Informatics X* (p. 11). SPIE. https://doi.org/10.1117/12.2558696

Jeddoub, I., Nys, G.-A., Hajji, R., & Billen, R. (2024). Data integration across urban digital twin lifecycle: A comprehensive review of current initiatives. *Annals of GIS*, 1–20. https://doi.org/10.1080/19475683.2024.2416135

Liao, Y., Xie, J., & Geiger, A. (2021). KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D [Version Number: 2]. https://doi.org/10.48550/ARXIV.2109.13410

Randall, L. S., & Antonisse, H. J. (2012). OGC observations and measurements standard to support feature-based motion imagery tracking. *Full Motion Video (FMV) Workflows and Technologies for Intelligence, Surveillance, and Reconnaissance (ISR) and Situational Awareness*, *8386*, 155–163. https://doi.org/10.1117/12.919937

Rustad, S. E., Hansen, B. J., Bjørndal, M. G., & Haugen, T. (2023). Exposing Military Sensor Data using SpatioTemporal Asset Catalog (STAC). *2023 International Conference on Military Communications and Information Systems (ICMCIS)*, 1–7. https://doi.org/10.1109/ICMCIS59922.2023.10253573

Sabri, S., Chen, Y., Lim, D., Rajabifard, A., & Zhang, Y. (2022). AN INNOVATIVE TOOL FOR OPTIMISED DEVELOPMENT ENVELOPE CONTROL (DEC) ANALYSIS AND SCENARIO BUILDING IN DIGITAL TWIN [Conference Name: ISPRS TC IV¡br¿17th 3D GeoInfo Conference - 19&ndash;21 October 2022, Sydney, Australia Publisher: Copernicus GmbH]. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, *XLVIII-4-W4-2022*, 117–123. https://doi.org/10.5194/isprs-archives-XLVIII-4-W4-2022-117-2022

Sabri, S., & Witte, P. (2023). Digital technologies in urban planning and urban management. *Journal of Urban Management*, *12*(1), 1–3. https://doi.org/10.1016/j.jum.2023.02.003

Sasani, S., Asgari, J., & Amiri-Simkooei, A. R. (2016). Improving MEMS-IMU/GPS integrated systems for land vehicle navigation applications. *GPS Solutions*, *20*(1), 89–100. https://doi.org/10.1007/s10291-015-0471-3

Smyth, C. S. (2022, November). OGC GeoPose 1.0 Data Exchange Draft Standard. https://docs.ogc.org/dis/21-056r10/21-056r10.html

Wang, L., & Li, S. (2018). Enhanced Multi-sensor Data Fusion Methodology based on Multiple Model Estimation for Integrated Navigation System. *International Journal of Control, Automation and Systems*, *16*(1), 295–305. https://doi.org/10.1007/s12555-016-0200-x

Weil, C., Bibri, S. E., Longchamp, R., Golay, F., & Alahi, A. (2023). Urban Digital Twin Challenges: A Systematic Review and Perspectives for Sustainable Smart Cities. *Sustainable Cities and Society*, *99*, 104862. https://doi.org/10.1016/j.scs.2023.104862

Xu, H., Omitaomu, F., Sabri, S., Zlatanova, S., Li, X., & Song, Y. (2024). Leveraging generative AI for urban digital twins: A scoping review on the autonomous generation of urban data, scenarios, designs, and 3D city models for smart city advancement. *Urban Informatics*, *3*(1), 29. https://doi.org/10.1007/s44212-024-00060-w

Yue, P., & Shangguan, B. (2023, September). OGC Training Data Markup Language for Artificial Intelligence (TrainingDML-AI) Part 1: Conceptual Model Standard. https://docs.ogc.org/is/23-008r3/23-008r3.html