Poznań University of Technology

Politechnika Poznańska

Faculty of Control, Robotics and Electrical Engineering
Institute of Robotics and Machine Intelligence

Wydział Automatyki, Robotyki i Elektrotechniki
Instytut Robotyki i Inteligencji Maszynowej

Doctoral dissertation

Point-oriented object localization and tracking in low-altitude aerial imagery

Lokalizacja i śledzenie obiektów zorientowanych punktowo na zdjęciach lotniczych zarejestrowanych na niskim pułapie

mgr inż. Bartosz Ptak

Supervisor: dr hab. inż. Paweł Drapikowski
Auxiliary supervisor: dr inż. Marek Kraft

Poznań, 2025

mgr inż. Bartosz Ptak
Point-oriented object localization and tracking in low-altitude aerial imagery
Doctoral dissertation, Poznań, 2025
Supervisor: dr hab. inż. Paweł Drapikowski
Auxiliary supervisor: dr inż. Marek Kraft

Poznań University of Technology
Institute of Robotics and Machine Intelligence
Faculty of Control, Robotics and Electrical Engineering

Abstract

Drone-based crowd monitoring is a key technology for applications in surveillance, public safety, and event management, primarily due to its dynamic, aerial perspective that surpasses the limitations of traditional ground-based systems. Recently, a new trend has emerged in tiny object localization and tracking, characterized by the use of point-oriented object sensing, which enables accurate monitoring of densely packed individuals in low-altitude aerial imagery. In this dissertation, advancements in this area are presented, including novel approaches for point-oriented object localization and a new solution for point-oriented object tracking. For localization and counting tasks, a series of enhancement mechanisms is introduced. These include the integration of motion-based features, the use of task-oriented synthetic data, and addressing the influence of varying image input resolutions in neural networks. A direct incorporation of drone altitude into the neural network architecture is also investigated, a new module that processes all pixels of high-resolution images without downscaling is proposed, and a novel loss function tailored to point-oriented localization is introduced. For object tracking and trajectory counting, an algorithm is proposed that enhances trajectory continuity and unique counting reliability in drone-based crowd monitoring, enabling the accurate tracking of individuals across video sequences. The approach extends the Simple Online and Real-time Tracking (SORT) framework by replacing the bounding-box assignment with a point-distance metric. It is further enhanced with three cost-effective techniques: camera motion compensation, altitude-aware assignment, and classification-based trajectory validation. Additionally, Deep Discriminative Correlation Filters (DDCF) are integrated, which reuse spatial feature maps from localization algorithms to improve computational efficiency and handle missed detections. To support this research, two new datasets, UP-COUNT and UP-COUNT-TRACK, are introduced, addressing challenges in modern drone imagery, including simultaneous camera and object motion, as well as changing flight altitudes. All proposed methods are quantitatively evaluated on both the publicly available DroneCrowd dataset and new datasets, demonstrating significant improvements in localization and tracking performance and achieving state-of-the-art results in drone-based people and trajectory counting. This dissertation makes substantial contributions to computer vision in aerial robotics, offering

practical tools for rapid crowd size and movement estimation. These tools have been demonstrated to be applicable in real-world scenarios.

Streszczenie

Monitorowanie tłumu z wykorzystaniem dronów stanowi kluczową technologię w obszarach nadzoru, bezpieczeństwa publicznego oraz zarządzania wydarzeniami, głównie dzięki dynamicznej, lotniczej perspektywie, która jest pozbawiona ograniczeń tradycyjnych systemów naziemnych. W ostatnim czasie pojawił się nowy trend w lokalizacji i śledzeniu bardzo małych obiektów, oparty na punktowej detekcji obiektów, umożliwiającej precyzyjne monitorowanie gęsto zgromadzonych osób w obrazach zarejestrowanych na niskim pułapu lotu. W niniejszej rozprawie przedstawiono postępy w tym obszarze, w tym nowe podejścia do punktowej lokalizacji obiektów oraz nowatorskie rozwiązanie dla punktowego śledzenia obiektów. Na potrzeby zadań lokalizacji i zliczania wprowadzono szereg usprawniających mechanizmów, obejmujących: integrację cech opartych na ruchu, wykorzystanie syntetycznych danych oraz zbadanie wpływu zmiany rozdzielczości obrazów wejściowych w sieciach neuronowych. Dodatkowo zbadano wpływ uwzględnienia wysokości lotu drona bezpośrednio w architekturze sieci neuronowej, zaproponowano nowy moduł sieci neuronowej przetwarzający wszystkie piksele obrazów wysokiej rozdzielczości bez potrzeby ich skalowania oraz opracowano nową funkcję kosztu dostosowaną do punktowej lokalizacji. W zakresie śledzenia obiektów i zliczania unikalnych trajektorii opracowano algorytm poprawiający ciągłość trajektorii oraz wiarygodność zliczania w kontekście monitorowania tłumu z dronów, umożliwiający precyzyjne śledzenie pojedynczych osób w sekwencjach wideo. Podejście to rozszerza algorytm SORT (Simple Online and Real-time Tracking), zastępując dopasowanie oparte na ramkach ograniczających metryką odległości punktowej. Ulepszenia obejmują także trzy efektywne z punktu widzenia kosztu obliczeniowego techniki: kompensację ruchu kamery, dopasowanie obiektów uwzględniające wysokość lotu oraz walidację trajektorii opartą na klasyfikacji. Ponadto, zintegrowano Deep Discriminative Correlation Filters (DDCF), które wykorzystują przestrzenne mapy cech bezpośrednio z algorytmu lokalizacji w celu zwiększenia efektywności obliczeniowej i lepszego radzenia sobie z brakującymi detekcjami. W ramach wsparcia badań wprowadzono dwa nowe zbiory danych: UP-COUNT oraz UP-COUNT-TRACK, odpowiadające na wyzwania nowoczesnego obrazowania z dronów, takie jak jednoczesny ruch kamery i obiektów oraz zmienna wysokość lotu. Wszystkie zaproponowane metody zostały poddane ilościowej ocenie zarówno na publicznie dostępnym zbiorze danych

DroneCrowd, jak i na nowo wprowadzonych, wykazując istotną poprawę dokładności lokalizacji i śledzenia, osiągając przy tym najlepsze dotychczasowe wyniki w zadaniach zliczania osób i trajektorii z perspektywy powietrznej. Praca ta wnosi istotny wkład w rozwój widzenia komputerowego w robotyce lotniczej oraz dostarcza praktycznych narzędzi do efektywnej oceny liczebności i ruchu tłumu, co zostało potwierdzone w rzeczywistych scenariuszach.

Acknowledgement

I want to express my sincere gratitude to my supervisor, Paweł Drapikowski, for his scientific guidance, constant support, and availability. I am also deeply thankful to my co-supervisor, Marek Kraft, for introducing me to the world of science and inspiring me to contribute to it with my own ideas and hard work.

I am deeply grateful to my beloved wife for her unwavering support and for filling my life with joy. Thank you for your patience, understanding, encouragement, snacks, and for listening to my endless ramblings about academic theories – many of which I'm still trying to figure out myself.

I would also like to thank my family, especially my parents, for all they've taught me and for their constant support. Their ability to cheer me on while simultaneously asking, "Are you still working on that?" has been a constant source of motivation.

Finally, to my friends and colleagues: thank you for your support and the countless brainstorming sessions we have had. Your pep talks kept me grounded, or at least as grounded as one can be during a PhD, making this journey a bit more bearable.

Contents

List of acronyms

AI *Artificial Intelligence*.

BN Batch Normalization.

CCTV Closed-Circuit Television.

CNN *Convolutional Neural Network*.

CV *Computer Vision*.

DDCF Deep Discriminative Correlation Filters.

DL *Deep Learning*.

ECO Efficient Convolution Operators.

FPS frames per second.

GNSS Global Navigation Satellite System.

GOG Globally-Optimal Greedy.

GPS Global Positioning System.

GSD Ground Sampling Distance.

GSI Gaussian-smoothed interpolation.

HOTA *Higher Order Tracking Accuracy*.

ID-F1 *Identity F1-Score*.

ID-SW *ID-Switches*.

IoU Intersection over Union.

KLD Kullback-Leibler Divergence.

L-AP *Localization Average Precision*.

MAE *Mean Absolute Error*.

MOTA *Multiple Object Tracking Accuracy*.

MPM Motion and Position Map.

MSE *Mean Square Error*.

NAS Neural Architecture Search.

nMAE *Normalized Mean Absolute Error*.

NWD Normalized Wasserstein Distance.

OC-SORT Observation-Centric SORT.

PD Pixel Distill.

ReLU Rectified Linear Unit.

RFLA Receptive Field-based Label Assignment.

SD-DETR Swin-Deformable Detection Transformer.

SORT Simple Online and Realtime Tracking.

T-AP *Tracking Average Precision*.

Tr-MAE Tracking Mean Absolute Error.

Tr-nMAE Tracking Normalized Mean Absolute Error.

UAV *Unmanned Aerial Vehicle*.

1 Introduction

UAV (*Unmanned Aerial Vehicle*) technology has recently advanced, establishing itself as a critical tool for low-altitude data acquisition across various sectors, including agriculture, surveillance, smart city infrastructure, security, and others [1]. These advancements are driven by improvements in sensor accuracy, extended battery life, and enhanced autonomous navigation capabilities, enabling UAVs to perform complex tasks with increased precision and efficiency. Drones now integrate a wide array of sensors, including GNSS (Global Navigation Satellite System), accelerometers, gyroscopes, barometers, and distance sensors, making them more reliable and easier to operate. They are particularly valuable for capturing high-resolution visual data, thus broadening their applications across industries while presenting new analytical challenges. Equipped with various visual sensors on gimbals, UAVs can host traditional RGB cameras that operate in the visible spectrum, as well as infrared [2], multispectral [3, 4], and hyperspectral [5] cameras. These sensors enable the collection of a vast range of spectral data, which is essential for specific, application-oriented tasks.

A key benefit of drones is their ability to be rapidly deployed, allowing quick ascent to desired altitudes for top-view and multi-perspective monitoring. Their accessibility and versatility make drones indispensable not only in commercial markets but also in public services such as law enforcement [6] and firefighting [7], providing enhanced situational awareness during operations. Moreover, drones' ability to record high-quality (spatial resolution) video (temporal resolution) and fuse it with precise sensor data supports advanced algorithms and data analysis. Coupled with recent developments in AI (*Artificial Intelligence*) and CV (*Computer Vision*), UAVs are increasingly capable of autonomously interpreting complex data, pushing the boundaries of their functionality and expanding their role in data-driven decision-making. In conjunction with machine learning algorithms [8], drones are now capable of detecting and tracking specific objects and processing massive data streams, transforming raw sensor data into actionable insights.

Despite all the advantages of drone video processing, bird's-eye perspective analysis presents more challenges compared to ground-based imagery sensing. Drone-based algorithms must address issues such as the changing flight altitude, camera motion, perspective distortions, and the high density of small, dynamically moving

objects against complex backgrounds. Moreover, existing datasets for this task are often insufficient due to their limited size, repetitive scenes, lack of annotations, or inherent biases. Despite these limitations, researchers have advanced remote object detection in UAV surveillance [9], primarily through deep learning methods [10]. The integration of these techniques with classical algorithms and novel drone technologies has unlocked new possibilities across various applications. This progress is particularly evident in the use of drones in smart cities, where they contribute to enhancing public safety and improving urban resource management [11].

1.1 Problem description and motivation

Classical object detection methods implement a regression task where the model estimates a rectangular bounding box enclosing an object, including its location and spatial dimensions within the image. While widely applied in drone-based applications [12], these methods have certain limitations compared to point-oriented object localization. Object localization, which aims to determine the image coordinates of an object's center, is especially beneficial when exact boundary definition is not essential for analysis. The approach emphasizes detecting the presence and position of objects in a scene, often achieving better performance in high-density areas and high-resolution images by leveraging pixel-level information. Furthermore, creating point-based datasets is more cost-effective and time-efficient than annotating bounding boxes, particularly in cases involving densely clustered objects or irregular shapes, where bounding box annotations are complex and error-prone. The visual interpretation of object detection and object localization tasks is presented in Figure 1.1.

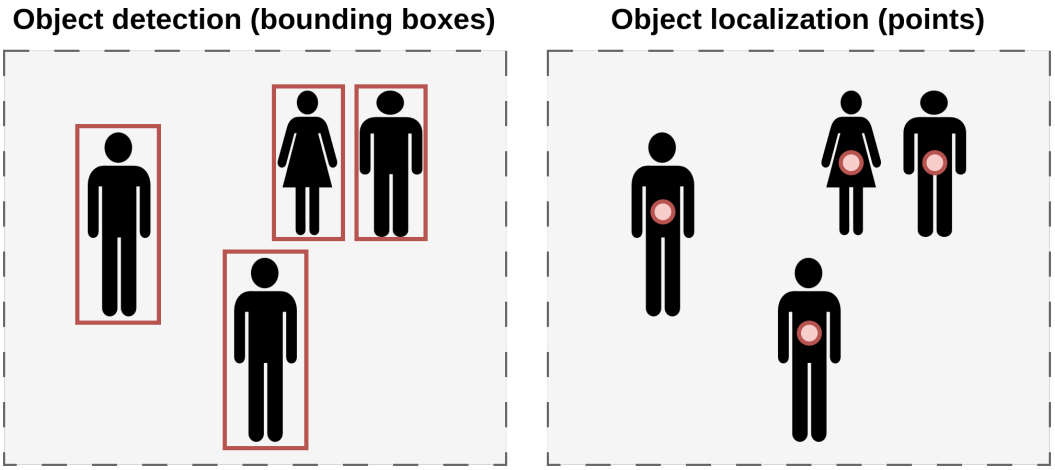


Figure 1.1: The visual interpretation of object detection and object localization tasks.

One area where significant strengths of this approach over detection have been noted is in the recognition of extremely small and densely packed objects, such as

individuals within a crowd. These objects are often classified as micro (less than 2 pixels) or very tiny (2–8 pixels), depending on the image resolution [13], where the size of the object is defined as the geometric mean of its height and width. A summary of differences between object localization and object detection tasks is shown in Table 1.1.

Table 1.1: Key differences between object detection and object localization tasks.

Feature	Object detection	Object localization
Goal	Provide object shape boundaries	Provide object center
Output	Bounding box (x, y, w, h)	Coordinates (x, y)
Scope	Focus on object presence, location, and shape	Focus on object presence and location
Applications	General recognition tasks	Specific tasks like people counting

One of the applications of object localization is crowd counting, which relies on determining the number of people visible in an image or the number of unique people trajectories present in a video recording. It can be divided into two approaches: density-based and object-based. The density-based approach employs algorithms to generate a density mask, where the sum of values corresponds to the estimated count of individuals [14]. While this method typically achieves better counting accuracy, it is limited to single-frame analysis and does not identify individual people. In contrast, object-oriented methods have gained prominence in recent years. Using the object localization paradigm, these methods identify the position of each individual within a frame, enabling accurate counting. With the known coordinates of objects, unique identification labels can also be assigned and propagated across frames, allowing people to be counted throughout an entire video sequence. This in-frame processing capability facilitates the development of more advanced and task-specific algorithms and applications.

In literature, crowd-counting algorithms are applied to both fixed-position camera systems, such as CCTV (Closed-Circuit Television) cameras [15, 16, 17, 18], and moving camera platforms, including drone-mounted cameras [19, 20]. Drones, in particular, offer flexible coverage and mobility, making them highly effective for monitoring dynamic environments. However, changes in flight altitude significantly impact both the resolution of the collected data and the extent of the observed area. Following, remote sensing imagery can thus be categorized based on the altitude at which it is acquired:

- low-altitude imagery: captured by drones operating at elevations between 1 and 120 meters above ground level, providing high ground resolution and detailed imagery suitable for localized analysis,

- medium-altitude imagery: acquired by aircraft, such as planes, flying at intermediate altitudes, offering a broader perspective and covering larger areas with moderate resolution,

- high-altitude imagery: obtained from satellites, providing global coverage and long-term monitoring capabilities and enabling the observation of large-scale phenomena.

This thesis concentrates on image processing in low-altitude aerial contexts. Although drones are capable of capturing high-resolution and richly detailed data, they introduce unique algorithmic challenges that differ significantly from those associated with static camera systems [21]. These challenges arise from scale variations due to altitude and perspective shifts, as well as distortions introduced by camera motion [22]. Furthermore, the tiny size and high density of target objects further complicate the task, as shown in Figure 1.2.

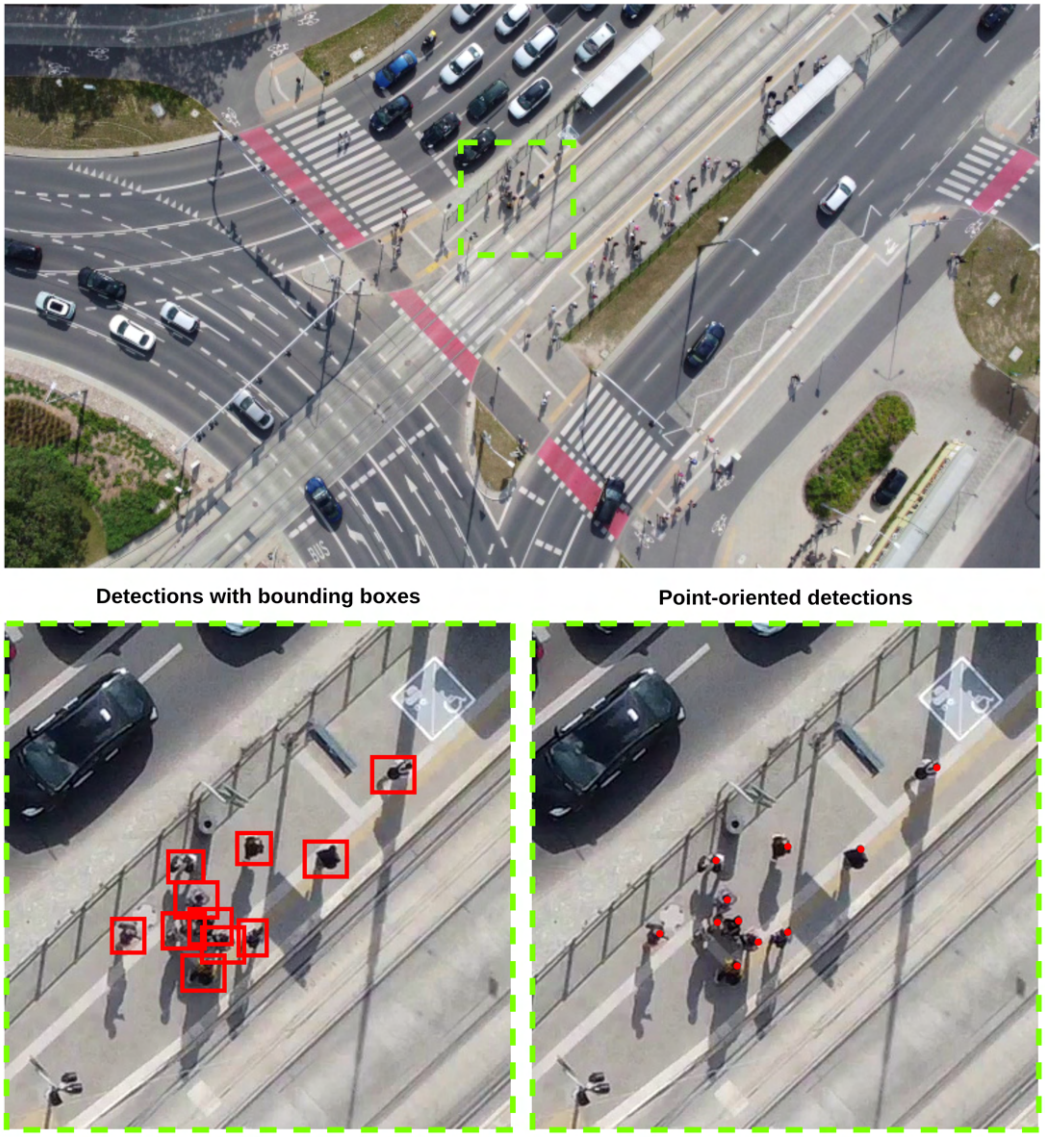


Figure 1.2: The difference in tiny, density-packed objects labeling, considering object detection and object localization tasks.

With technological advancements, drones are now equipped with high-resolution cameras that can capture detailed information about observed scenes [23]. Combined with the evolution of deep learning and computer vision techniques, these developments significantly extend the capabilities of existing algorithms. It facilitates the monitoring of moving objects in crowded areas using UAV imagery, enhancing crowd analysis during large public events [24, 25, 26] and playing a crucial role in safety and surveillance operations [27].

1.2 Research hypothesis and objectives

While object detection and tracking have been extensively studied in recent years, point-oriented object localization in low-altitude aerial imagery remains a relatively underexplored field with significant practical importance. Despite the availability of high-quality video recordings from modern aerial vehicles, insufficient attention has been given to utilizing spatial and temporal features in this data. This gap highlights the need for developing specialized methods for point-oriented localization and tracking that fully leverage the unique characteristics of low-altitude imagery. Advancements in this area would facilitate robust, high-resolution analyses across various domains, offering distinct advantages for real-world applications such as crowd counting and individual tracking.

Based on the above analysis, the following hypothesis of this dissertation is formulated:

> ***It is possible to evolve computer vision and neural network methods for point-oriented object localization and tracking tasks using spatial and temporal features extracted from a sequence of images, especially image sequences registered from low-altitude Unmanned Aerial Vehicles.***

This hypothesis led us to formulate the following principal goals of this thesis:

- to propose mechanisms that enhance point-oriented object localization in drone imagery by utilizing spatial and temporal features,
- to develop a point-oriented tracking method that improves trajectory continuity and consistency,
- to create a point-oriented object localization and tracking dataset tailored to the requirements of modern drone-based video analysis,

- to validate the proposed algorithm for individual localization and tracking in real-world use cases.

The main contributions of this research include:

- comprehensive analysis of tiny and very tiny object detection and localization methods, with a particular focus on drone-based crowd counting,
- development of object localization and tracking datasets tailored to the requirements of modern UAV-based crowd management,
- enhancement of people localization methods through the introduction of methodologies that leverage both spatial and temporal features,
- design of a point-oriented tracking algorithm that significantly improves previous approaches, particularly by enhancing trajectory continuity,
- demonstration of real-world applicability, showcasing the potential of the developed technology and highlighting its impact on research and innovation.

1.3 Influence of the research

The research presented in this dissertation has the potential to make significant contributions to both the scientific community and broader society. By developing advanced techniques for object localization, tracking, and people counting from low-altitude aerial imagery, this work provides valuable tools for event management, urban planning, public safety, and analyzing crowd behavior. Furthermore, the methodology can be adapted for tasks involving counting and tracking tiny objects in various domains, including agriculture [28], forestry [29], and livestock farming [30]. These contributions can lead to improved systems and more informed decision-making across multiple sectors.

1.3.1 Contributions to the field

Methodological advancements. This thesis introduces novel point-oriented approaches for detecting, tracking, and counting people in low-altitude aerial images, addressing limitations in traditional tiny object detection methods. By focusing on point-based localization, these methods provide a computationally efficient alternative that enhances accuracy in high-resolution and high-density scenes, where conventional bounding-box detection often struggles.

New data collection and benchmarking. To support and validate these methods, the research contributes a new dataset specific to low-altitude people localization and tracking in varied urban and event settings. This dataset is curated with hand-labeled annotations suitable for point-based methods and offers a valuable benchmarking resource for future research.

1.3.2 Societal impact

Enhanced event and urban management. The ability to accurately monitor crowd density and movement from aerial perspectives enables better planning and management of public spaces and events. This research presents an algorithm that can serve as a core component of a tool for monitoring population dynamics, supporting both event organizers and urban planners in managing foot traffic, identifying bottlenecks, and maintaining safe and accessible environments.

Applications for public safety. Reliable, efficient people-counting tools can assist police and emergency services in monitoring public spaces and responding to unusual behavior patterns. During large gatherings or emergencies, this research offers a potential tool for quickly assessing crowd size and movement, enabling proactive responses and improved crowd control, and supporting law enforcement agencies in maintaining public safety.

1.4 Author activities

1.4.1 List of grants

The following is a list of research grants obtained by the author during the course of his doctoral studies.

- As part of Mobility IV within the INPUTDoc project under the NAWA STER program, the author participated in a research internship at the UAS Drone Center at the University of Southern Denmark (SDU) between April and June 2024. The mobility was conducted under the supervision of Dr. Henrik Skov Midtiby and focused on developing computer vision algorithms for analyzing marine environments using drone-collected video data. The stay fostered knowledge exchange, research collaboration, and the initiation of a joint research project, contributing significantly to advancements in environmental monitoring using aerial robotics.

- As part of the TERRINet Call 8 competition funded under the Horizon 2020 program, the author was awarded a mobility scholarship to conduct research at the GRVC Robotics Laboratory. The project focused on developing a vision- and deep learning-accelerated landing system for UAVs utilizing the PX4 autopilot and an onboard Edge AI device.

- As part of the TERRINet Call 10 competition funded under the Horizon 2020 program, the author received a mobility scholarship to carry out research at the Institut de Robòtica i Informàtica Industrial (IRI) at the Universitat Politècnica de Catalunya in Spain. The project aimed to develop a novel approach to the problem of camera network spatial topology discovery and activity control. The IRI provided a collaborative and technically advanced environment that was essential for exploring innovative methods in intelligent vision systems and distributed sensor networks.

- The author received the computation grant number 596, "Detection and localization of small objects on low-altitude aerial images" funded by the Poznań Supercomputing and Networking Center (PSNC), Poland.

1.4.2 List of projects

The following is a list of research projects in which the author actively participated during the course of his doctoral studies.

- As part of the R&D project "Increasing the quality of fiber hemp seed by robotization" (POIR.01.01.01-00-2271/20), funded by the National Center for Research and Development (NCBR), the author focused on developing a computer vision tools for intelligent agricultural robot aimed at improving the quality of fiber hemp seed. This project integrates advanced robotics and computer vision to enhance the efficiency and selectivity of agricultural practices in hemp cultivation.

- As part of the R&D project "Development of advanced autonomous drone swarm technology for digital critical infrastructure security, ad hoc inspection application and innovative creative entertainment" (POIR.01.01.01-00-0040/22), funded by the National Center for Research and Development (NCBR), the author focused on developing computer vision algorithm to increase drones intelligence. The work also includes implementing the algorithm directly on board the drone.

- As part of the COGNITION project, funded by the European Space Agency OSIP program (PO number: 4000138073), the author focused on the development of a distributed data processing system for lunar activities. The primary goal of the project was to evaluate the potential for advanced onboard data processing in spacecraft, aiming to reduce the volume of transmitted data by extracting only essential information. This approach supports increased autonomy in space vehicles such as rovers and landers, enabling more efficient and intelligent decision-making during lunar missions.

- As part of the PUT-ISS project, the author contributed to the development of a robotic vision system implemented on a space-qualified embedded computer. This system was delivered to the International Space Station as part of the LeopardISS project – one of the experiments selected for execution during Poland's first space mission, the IGNIS mission.

1.4.3 List of publications

The list of publications produced by the author during his doctoral studies strongly related to this thesis.

- CountingSim: Synthetic Way To Generate a Dataset For The UAV-view Crowd Counting Task / **Bartosz Ptak**, Dominik Pieczyński // W: Proceedings of the 3rd Polish Conference on Artificial Intelligence PP-RAI'2022, April 25-27, 2022, Gdynia, Poland - Gdynia, Polska: Uniwersytet Morski w Gdyni, 2022 - s. 20-24 **[20 pts]**

- On-Board Crowd Counting and Density Estimation Using Low Altitude Unmanned Aerial Vehicle-Looking beyond Beating the Benchmark / **Bartosz Ptak**, Dominik Pieczyński, Mateusz Piechocki, Marek Kraft // Remote Sensing - 2022, vol. 14, iss. 10, s. 2288-1-2288-18 **[100 pts, IF 5.0]**

- Elevating point-based object detection in UAVs: A deep learning method with altitude fusion / Michał Wiliński, **Bartosz Ptak**, Marek Kraft // W: Progress in Polish Artificial Intelligence Research 5: Proceedings of the 5th Polish Conference on Artificial Intelligence (PP-RAI'2024), 18-20.04.2024, Warsaw, Poland / red. Jacek Mańdziuk, Adam Żychowski, Mikołaj Małkiński - Warsaw, Poland : Warsaw University of Technology, 2024 - s. 228-234 **[20 pts]**

- Enhancing people localisation in drone imagery for better crowd management by utilising every pixel in high-resolution images / **Bartosz Ptak**, Marek Kraft / Arxiv.org - 2025, s. 1-15

- Improving trajectory continuity in drone-based crowd monitoring using a set of minimal-cost techniques and deep discriminative correlation filters / **Bartosz Ptak**, Marek Kraft (WARiE) / Arxiv.org - 2025, s. 1-17

The list of publications produced by the author during his doctoral studies not related directly to this thesis.

- Spotting advertisements from above: billboard detection and segmentation in UAV imagery / **Bartosz Ptak**, Jan Dominiak, Marek Kraft // W: Progress in Polish Artificial Intelligence Research 4 / red. Adam Wojciechowski, Piotr Lipiński - Łódź, Polska: Wydawnictwo Politechniki Łódzkiej, 2023 - s. 67-72 **[20 pts]**
- Integration of Heterogeneous Computational Platform-based, AI-capable Planetary Rover Using ROS 2 / Marek Kraft, Krzysztof Walas, **Bartosz Ptak**, Michał Bidziński, Krzysztof Stężała, Dominik Pieczyński // W: IGARSS 2023 - IEEE 2023 International Geoscience and Remote Sensing Symposium: proceedings, 16-21 July, 2023, Pasadena, California, USA: IEEE, 2023 - s. 2014-2017 **[20 pts]**
- Cognition: Distributed Data Processing System for Lunar Activities / Krzysztof Walas, Marcin Cwiek, Tomasz Strzałka, Marek Wiejak, Piotr Bosowski, Michał Kawulok, Mateusz Przeliorz, Dominik Pieczyński, **Bartosz Ptak**, Krzysztof Stężała, Michał Bidziński, Marek Kraft // W: IGARSS 2023 - IEEE 2023 International Geoscience and Remote Sensing Symposium: proceedings, 16-21 July, 2023, Pasadena, California, USA: IEEE, 2023 - s. 2010-2013 **[20 pts]**
- Deepness: Deep neural remote sensing plugin for QGIS / Przemysław Aszkowski, **Bartosz Ptak**, Marek Kraft, Dominik Pieczyński, Paweł Drapikowski // SoftwareX - 2023, vol. 23, s. 101495-1-101495-6 **[200 pts, IF: 2.4]**
- LunarSim: Lunar Rover Simulator Focused on High Visual Fidelity and ROS 2 Integration for Advanced Computer Vision Algorithm Development / Dominik Pieczyński, **Bartosz Ptak**, Marek Kraft, Paweł Drapikowski // Applied Sciences - 2023, vol. 13, iss. 22, s. 12401-1-12401-16 **[100 pts, IF: 2.5]**
- ISO-compatible personal temperature measurement using visual and thermal images with facial region of interest detection / **Bartosz Ptak**, Przemysław Aszkowski, Joanna Weissenberg, Marek Kraft, Michał Weissenberg // IEEE Access - 2024, vol. 12, s. 44262-44277 **[100 pts, IF: 3.4]**
- A fast, lightweight deep learning vision pipeline for autonomous UAV landing support with added robustness / Dominik Pieczyński, **Bartosz Ptak**, Marek Kraft, Mateusz Piechocki, Przemysław Aszkowski // Engineering Applications

of Artificial Intelligence - 2024, vol. 131, s. 107864-1-107864-13 **[140 pts, IF: 7.5]**

- Mapping urban large-area advertising structures using drone imagery and deep learning-based spatial data analysis / **Bartosz Ptak**, Marek Kraft // Transactions in GIS - 2024, vol. 28, no. 6, s. 1728-1749 **[100 pts, IF: 2.1]**
- Improved Grapes Detection and Tracking in Drones Imagery by Integrating the Coordinate Attention Mechanism / **Bartosz Ptak**, Marek Kraft // W: IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP 2024): IEEE, 2024 - s. 1-8 **[20 pts]**
- Visual Feedback System Supporting Robotic Manipulation of Hemp Plants / Marek Kraft, **Bartosz Ptak**, Mateusz Piechocki, Dominik Pieczyński, Kamil Młodzikowski, Bartłomiej Kulecki, Dominik Belter // Journal of Natural Fibers - 2025, vol. 22, no. 1, s. 2454261-1-2454261-16 **[140 pts, IF: 2.8]**
- Improving consistency of marine mammals tracking in challenging drone recordings through visual particle filter integration / **Bartosz Ptak**, Henrik Skov Midtiby, Marek Kraft // Neurocomputing - 2025, Volume 646, 14 September 2025, 130503 **[140 pts, IF: 5.5]**

1.5 Thesis structure

The next sections of the dissertation are structured as follows. **Chapter 2, Background and related work**, provides an introduction to the problem domain, an overview of existing state-of-the-art solutions, and descriptions of the available datasets and the metrics used to evaluate the performance of the algorithms. **Chapter 3, Point-oriented object localization methodology**, focuses on methods for improving object localization accuracy. **Chapter 4, Point-oriented object tracking methodology**, explores considerations specific to tracking point objects. **Chapter 5, Real-world usage scenarios**, presents examples of a real-world application that utilizes algorithms combined with a drone's sensors. Finally, **Chapter 6, Discussion**, summarizes the conducted research, discusses its limitations, and suggests potential areas for future development.

2 Background and related work

2.1 Modern computer vision

Building upon decades of research, modern computer vision has undergone a transformative evolution, driven primarily by advancements in DL (*Deep Learning*) [10]. One of the most significant networks in this field is the CNN (*Convolutional Neural Network*), which has introduced numerous improvements and novel ideas [31], significantly advancing the computer vision field. The next big leap was the introduction of Transformers, which enabled modeling long dependencies between input sequence elements, including 2D images and 3D representations [32]. Following this, generative models continued to push the boundaries of image synthesis and representation learning. Diffusion-based models have opened up new possibilities in super-resolution [33], image-to-image translation [34], and style transfer [35], thereby improving both the quality of generated images and the stability of training. In parallel, self-supervised learning methods, such as MoCO [36] and Dino [37], gained prominence by leveraging large unlabeled datasets to learn robust feature representations. Subsequent advances also include multimodal learning approaches, such as CLIP [38], which bridges vision and language tasks through contrastive pre-training. A more recent leap in visual processing is the emergence of foundation models [39], exemplified by Segment Anything Models [40], which enable general-purpose segmentation across diverse image domains. These evolutions have significantly expanded the scope of computer vision, unlocking new potential for advanced applications across various domains.

Modern advancements in core computer vision techniques have significantly impacted specialized fields such as remote sensing, enabling the efficient analysis of large-scale geospatial data collected from satellites, aerial platforms, and unmanned aerial vehicles. The ability of deep learning models to automatically extract information and features from remotely sensed imagery has further accelerated progress in this domain. In UAV-based applications, deep learning-driven object detection and segmentation have facilitated a range of operations, including autonomous navigation [41], engineering application support [42], infrastructure inspection [43, 44], urban object surveillance [45], search-and-rescue operations [46], and environmental monitoring [47]. These advancements enhance UAVs' capability to

analyze complex terrains and detect objects under diverse conditions, broadening their application in real-world scenarios.

Crucial to the success of many of these remote sensing applications is the accurate detection of objects in aerial images, as the unique perspective of aerial imagery presents significant challenges for traditional detection methods. These challenges are particularly pronounced when dealing with tiny objects and significant perspective distortions. However, deep learning-based approaches have significantly advanced aerial object detection, offering improved accuracy and robustness. Most deep learning-based methods provide axis-aligned bounding boxes [48], which define an object's location and size within an image. Recently, several approaches have explored the use of oriented bounding boxes [49], which better align with object boundaries by estimating their orientations, leading to more precise localization in this task. The rapid advancement of object detection models is also evident in drone technology [22], driven by the need to address several unique challenges in aerial imagery, including object rotation, complex backgrounds, increased difficulty in detecting small objects, scale variations affecting detection efficiency, and the sparse and uneven distribution of object categories. As a result, modern detection algorithms continue to evolve to improve robustness, adaptability, and increase the number of end-to-end solutions in UAV-based applications [50].

2.2 General people counting

People counting is a research problem that involves automatically detecting and counting the number of people present in the range of a sensor. This task can be categorized based on the environment: indoor applications, which address the needs of smart buildings [51], and outdoor scenarios, which are relevant to the broader context of smart cities [52]. Regardless of the environmental setting, people counting approaches can be further classified into visual and non-visual sensing modalities. Non-visual methods encompass a diverse range of sensors and techniques, including thermal sensing [53], acoustic measurement [54, 55], and signal analysis from cellular and WiFi networks [56, 57], among others [58]. While these approaches are often cost-effective and privacy-preserving, visual methods – typically based on RGB cameras – have gained popularity due to their ability to capture rich spatial and contextual information.

One of the earliest approaches to people counting using visual sensors was introduced in 1995 [59]. The authors employed an edge detection algorithm combined with a matching process to align detected individuals with a predefined geometric model. Subsequent early methods relied on classical computer vision techniques,

such as Gaussian modeling [60, 61], background subtraction [62, 63], Histogram of Oriented Gradients (HOG) [64, 65], optical flow [66], and the Expectation-Maximization algorithm [67].

The introduction of shallow neural networks enabled the learning of feature representations, thereby improving adaptability across various scenarios. Early neural network-based people counting methods typically used simple feedforward or convolutional architectures. In [68], an AdaBoost classifier was combined with a soft cascade mechanism to detect human heads. In [69], a basic deep CNN was used to estimate local crowd density. To address the multiscale object problem, Hydra CNN was introduced in [70]. This issue was also tackled in [71], where the authors proposed a blob-based architecture that generates scale-relevant features for both low- and high-density crowds. In [72], a mixture of CNNs specialized for various appearances was proposed to handle the diversity of multi-source data. Subsequent methods introduced more advanced architectures to address the challenges of crowd counting. A volumetric CNN was proposed in [73], enabling the processing of spatiotemporal slices from video data. CSRNet [74], a two-stage network, expanded the receptive field using dilated convolutions without increasing the number of parameters or computational cost. Finally, NAS (Neural Architecture Search)-based methods have been explored [75] to automate the search for multi-path architectures, further enhancing the robustness and adaptability of people counting networks.

In recent years, numerous methods utilizing deep learning have been introduced to directly address the challenge of localizing small objects in the people counting task. For instance, a fully point-based system that unifies counting and localization by matching a dense set of predefined candidate points was proposed. Next, this concept was advanced by introducing an End-to-End transformer-based architecture that directly regresses object positions along with confidence scores [76], offering a more streamlined solution. In contrast, the "Crowd Hat" approach was developed [77], leveraging feature maps from detection-based architectures. Their plug-and-play module processes these features to generate localization responses. The latest advancement in this domain is STEERER [78], a cutting-edge model that sets new performance standards across multiple ground-level counting and localization datasets. It effectively mitigates scale variation issues, arising from variable distances between individuals and the camera, by employing multi-scale feature selection and isolating informative components from non-informative ones within lower-resolution feature representations.

2.3 Tiny object detection and localization

Various methods for detecting tiny objects in remote sensing have been developed, particularly following the release of several aerial datasets. In 2018, the UAVDT dataset [79] was introduced, comprising one hundred drone-based video sequences annotated for vehicle detection. In 2021, multiple datasets were published, including DOTA [49], which focuses on detecting ships, bridges, vehicles, roundabouts, and soccer fields in multi-source aerial imagery; Visdrone [80], designed for urban object detection from drones at varying altitudes; and AI-TOD [81], targeting object detection in high-altitude aerial images. More recently, in 2023, the SODA-A dataset [82] was introduced, providing high-resolution images with a diverse range of labeled object classes. While these datasets have significantly contributed to the field, many suffer from inconsistent frame resolutions, and several lack annotations for extremely tiny objects [13]. Among them, AI-TOD is particularly notable, as it consists of tiled aerial images with a resolution of only 800×800 pixels but includes relatively small object annotations, establishing it as a benchmark for the tiny object detection task. This phenomenon, where not only the size of the object is important but also its percentage in the image resolution, is shown in Figure 2.1.

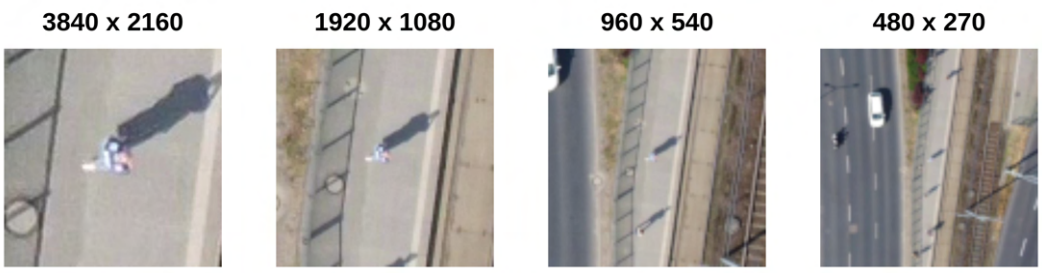


Figure 2.1: The figure presents the same object that originally occupying approximately 15×15 pixels (roughly 0.69% of the full image resolution), under progressive downscaling operations, including reductions by factors of 1x, 2x, 4x, and 8x.

Building upon AI-TOD, a range of advanced methodologies have been proposed. For instance, the authors of [83] introduced the NWD (Normalized Wasserstein Distance) as an alternative to the conventional IoU (Intersection over Union) metric, notably improving detection accuracy. Likewise, the Gaussian RFLA (Receptive Field-based Label Assignment) method [84] refines bounding box assignment using KLD (Kullback-Leibler Divergence), representing bounding boxes as two-dimensional Gaussian distributions to achieve superior scale generalization compared to NWD. In addition, the SD-DETR (Swin-Deformable Detection Transformer) [85] specifically addresses the challenge of detecting very small objects (smaller than four pixels), setting new state-of-the-art results on the AI-TOD benchmark. Although initially designed for object detection, these methodologies can be adapted for point-based

localization tasks, enabling their comparison and broadening their applicability across various domains.

With the release of the DroneCrowd dataset [86], initial approaches for detecting crowds in drone imagery began to emerge. The authors of STNNet [86] proposed a method that addresses density map estimation and object localization in densely populated scenes captured by drones. Their localization subnetwork combines classification and regression components. For object localization, object proposals are distributed across pixels. The classification branch predicts the likelihood that a proposal corresponds to an object, while the regression branch determines the precise positions of the positive proposals. In the MFA method [87], an advanced approach for crowd localization is introduced, representing the state-of-the-art in this field. The authors explore two distinct methods for generating feature maps. The first method involves heatmap generation using a UNet [88] architecture to estimate a heatmap of object positions, where each peak represents an object's location. The second method, MPM (Motion and Position Map), encapsulates both positional and directional movement information derived from sequential frames, capturing behavioral patterns within the sequence data. While these methods achieve notable results, they rely on a sliding-window approach, which can compromise the global context of the image. This limitation may reduce overall performance and increase the time required to process a single image.

2.4 Tiny object tracking

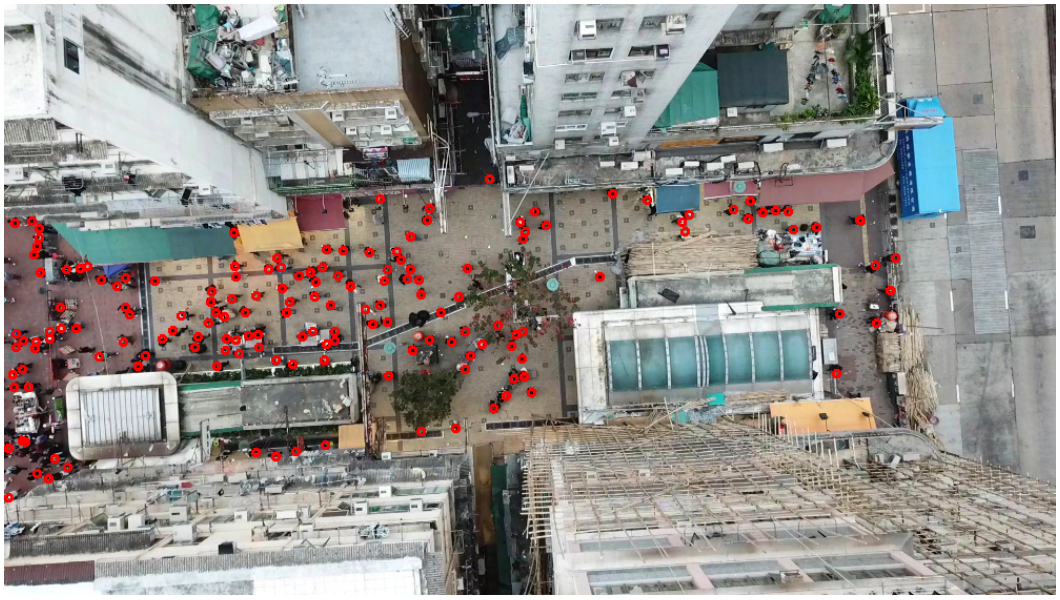
Existing efficient tracking methods primarily adopt a detect-to-track approach [89], where an object detection model first identifies object boundaries, followed by a tracking method that performs in-frame assignments, enabling the online processing of video data. This modular framework facilitates the independent development and optimization of object detection [9, 90] and tracking methods [91], allowing for the dynamic selection of components to suit specific requirements. The detect-to-track approach is particularly valuable in various real-world applications due to its modular nature. It enables the integration of specialized object detection and tracking methods, allowing each component to be tailored to address the specific challenges posed by different environments. An important aspect of tracking algorithms for real-world applications is their ability to operate online. In this paradigm, the algorithm processes frames sequentially in a forward-only manner, enabling on-board video analysis. This approach is particularly valuable in scenarios where storing and processing entire video sequences is impractical, such as in resource-constrained platforms like drones.

Recent advancements in online object tracking have focused on enhancing robustness and assignment accuracy, addressing key challenges in multi-object tracking scenarios. A pivotal contribution in this field was introduced in [92] with the SORT (Simple Online and Realtime Tracking) algorithm, which combines a Kalman filter and the Hungarian algorithm for efficient tracking. SORT achieves high tracking performance with minimal computational overhead, making it a widely used baseline for subsequent object-tracking methods, which highlights the key elements integral to contemporary tracking systems. Building on this foundation, the DeepSORT algorithm [93] extended SORT by incorporating a deep learning-based similarity metric to improve object association accuracy. Further enhancements were introduced with StrongSORT [94], which improves feature embeddings and trajectory association using GSI (Gaussian-smoothed interpolation) to address missed detections, thereby increasing tracking quality. To address limitations of the Kalman filter, such as sensitivity to state noise, temporal error magnification, and over-reliance on estimation, OC-SORT (Observation-Centric SORT) [95] was proposed. By addressing these issues, OC-SORT enhances robustness against object occlusions and non-linear movement behaviors, thereby extending the original SORT algorithm's capabilities. More recently, the BoT-SORT tracker [96] enhanced multi-object tracking by integrating camera motion compensation into the motion model, improving object association accuracy in dynamic environments. While these methods primarily focus on bounding box assignments, they provide a strong foundation for the development of point-oriented tracking methods.

Most drone-based crowd-tracking methods prioritize the localization stage, often overlooking the tracking stage, which is equally important for ensuring trajectory consistency and reducing counting errors. Recent state-of-the-art approaches, such as STNNet [86] and MFA [87], employ the GOG (Globally-Optimal Greedy) algorithm [97]. GOG is a multi-object tracking method based on a minimal-cost flow approach, enabling it to handle large input sequences and manage long-term occlusions. These features make it particularly useful for dense-object tracking and long sequences. However, as a globally optimized, offline method, it requires access to the entire sequence data, limiting its applicability in real-world tracking scenarios.

2.5 Datasets

This section focuses on datasets designed explicitly for the localization of point-oriented, tiny objects in drone-captured imagery.



(a) One of the lowest recorded images (sequence: 101).

(b) One of the highest recorded images (sequence: 042).

Figure 2.2: Example frames from the DroneCrowd dataset [86], with red markers indicating the locations of people's heads.

2.5.1 DroneCrowd

The DroneCrowd dataset [86] is the first large-scale dataset specifically designed for localizing tiny individuals in UAV-recorded videos. It consists of 112 video sequences recorded across diverse environments, including campuses, streets, parks, and plazas. Each sequence was manually annotated, yielding over four million labeled individuals. The dataset features a range of crowd densities, with the number of people per sequence varying from 25 to 455 and an average count of 144.8 individuals. To facilitate object tracking, trajectory IDs were added to the

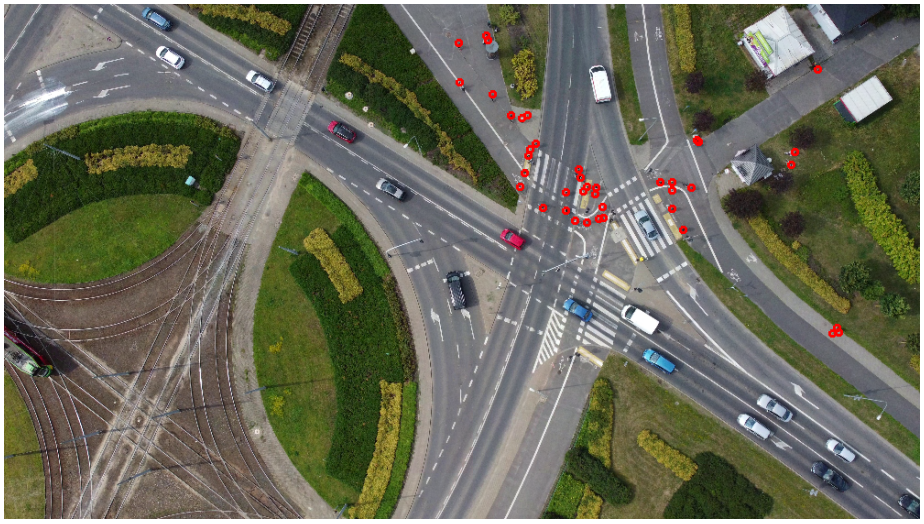
annotations, enabling the identification and continuity of each individual across frames within a sequence, resulting in over 20000 individual trajectories in total.

The dataset encompasses various illumination conditions, including cloudy, sunny, and nighttime settings, as well as different flight altitudes, providing a broad spectrum of scenarios. However, despite significant differences between sequences, intra-sequence visual variations are minimal. This can be attributed to two primary factors: first, the stationary position of the drone-mounted camera during each recording leads to a consistent background; second, the temporal span of each sequence is restricted to 300 frames captured at a rate of 25 FPS (frames per second), representing a recording duration of just 12 seconds. These characteristics result in relatively short individual trajectories and limited object displacement within sequences. Example frames from the dataset are shown in Figure 2.2, showcasing frames from sequences recorded at one of the lowest and one of the highest altitudes. The locations of individuals are highlighted with red circles.

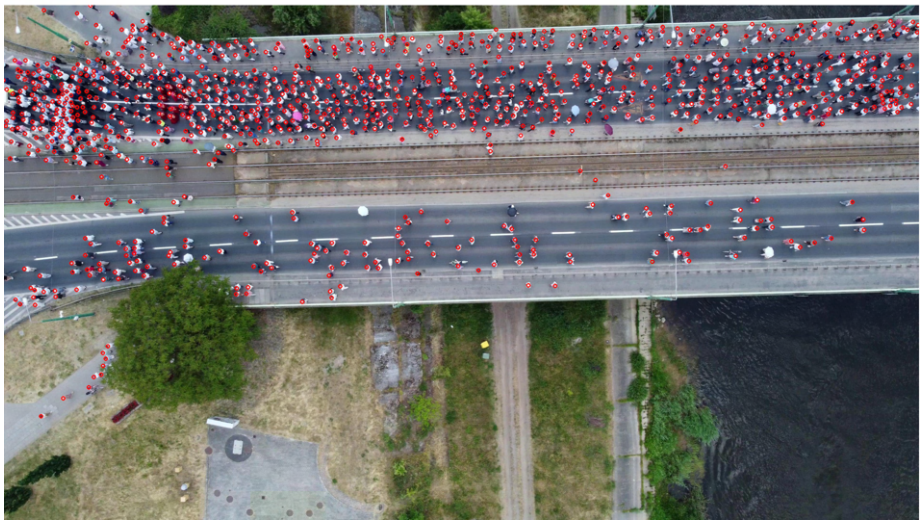
2.5.2 UP-COUNT

To address the requirements of modern people localization tasks, the UP-COUNT dataset [98] is introduced as part of the research. The introduced dataset includes drone footage captured using the DJI Mini 2 family. It encompasses diverse environments, including streets, plazas, public transport stops, parks and other green recreation places. It consists of 202 unique videos. Thus, frames were extracted with a step of one second, resulting in a diverse set of 10000 images with a resolution of 3840×2160 pixels. Acquisition conditions vary during the daytime and under different lighting conditions, creating challenging shadows. The recordings were taken at various altitudes and speeds of flight, as well as with varying densities of people. Each image is accompanied by additional information on flight altitude above ground level. Next, the labels for people's heads were manually prepared, resulting in 352487 instances. During the labeling process, each image was reviewed by two people, and the continuity of labels within each sequence was verified. Figure 2.3 contains example frames with the ground truth annotations. The top and middle images represent the lowest (26.0 meters) and the highest altitude (101.0 meters) recorded among the sequences, with an average of 60.3 meters. The bottom image presents the most crowded image, with 1039 instances, while the average object count in the dataset is 35.25. Increased variability in crowd counts and different backgrounds caused by the lack of a stationary camera position better reflect real-world scenarios. The UP-COUNT dataset is divided into three subsets for training, validation, and testing purposes, containing 141, 30, and 31 sequences, respectively. The described sequences' splits are prepared using altitude-based stratified sampling, providing a comparable altitude distribution.

(a) One of the lowest recorded images (sequence: 043).



(b) One of the highest recorded images (sequence: 142).



(c) One of the most crowded images (sequence: 16).

Figure 2.3: Example frames from UP-COUNT dataset [98] with red marks of people's heads.

To support the evaluation of point-based tracking algorithms, the UP-COUNT-TRACK dataset [99] is introduced as an extension of the original UP-COUNT dataset, incorporating tracking annotations for its test subset. This dataset comprises dense individual trajectories across 31 video sequences, encompassing a total of 33751 annotated frames. Sequence lengths range from 361 to 2541 frames, with an average of 1088.7 frames per sequence. For tracking purposes, the position of each individual was manually annotated in every frame, ensuring consistent identity labeling throughout each sequence. The number of trajectories per video ranges from 13 to 1182, with an average of 122.8 trajectories per sequence. In total, the dataset includes 3807 unique trajectories and 1360547 annotated instances, underscoring its large scale. Additionally, auxiliary metadata such as GPS coordinates and flight altitude is provided. Figure 2.4 illustrates sample frames with annotated trajectories. The dataset's diversity in recording conditions, flight altitudes, and drone motion presents a realistic and challenging benchmark, particularly suited for urban tracking scenarios.

Figure 2.4: Example frames from the UP-COUNT-TRACK dataset [99] with annotated trajectories illustrate the dataset's diverse recording environments, varying flight altitudes, and dynamic drone movements.

2.5.3 Datasets' comparison

Both datasets adopt the exact object localization and tracking paradigm, considering point-oriented annotations. To demonstrate the need to introduce a new dataset, UP-COUNT, a comparison is conducted. It is divided into two parts: the image-based people-counting task and the video-based people-trajectory counting task.

People Counting Task. This task involves estimating the positions of individuals in every frame of the dataset. It emphasizes diverse imagery and models with higher generalization capabilities. While the UP-COUNT dataset contains fewer images (10000) than DroneCrowd (33600) and has a lower label density (352487 vs. 4864280), it addresses more varied scenarios:

- Annotation Range: UP-COUNT includes a wider range of annotations per image (0–1039 vs. 25–445), enabling evaluations on both sparse and dense groups of individuals.

- Sequence Diversity: UP-COUNT features nearly double the number of unique sequences (202 vs. 112) and incorporates drone movement during recordings, making it a more realistic and challenging dataset for practical use cases.

Detailed statistical comparisons are provided in Table 2.1.

Table 2.1: Statistical comparison of the DroneCrowd and UP-COUNT datasets for the people counting task.

People counting statistics	**Dataset name**	
	DroneCrowd	**UP-COUNT**
Number of sequences	112	202
Frames resolution	1920 × 1080	3840 × 2160
Frames number	33600	10000
Min. object count	25	0
Mean object count	144.8	35.25
Max. object count	445	1039
Total count	4864280	352487
Moving camera	No	Yes
Drone altitude	No	Yes

People Trajectory Counting Task. This task focuses on estimating the number of individuals across entire video sequences. In addition to frame-level localization, a tracking algorithm is required to maintain consistent identity assignments across frames, ensuring accurate counting without duplication. The tracking component of UP-COUNT-TRACK further sets it apart from DroneCrowd. Based on Table 2.2, which compares test subsets of tracking datasets (note that UP-COUNT-TRACK provides

tracking labels only for its test subset), UP-COUNT-TRACK offers 3.7 times more labeled frames. With a comparable number of unique sequences, this results in longer and more challenging video sequences.

Table 2.2: Statistical comparison of the DroneCrowd and UP-COUNT-TRACK datasets for the people trajectory counting task (considering only test subsets).

Trajectory counting statistics	Dataset name	
	DroneCrowd	**UP-COUNT-TRACK**
Number of sequences	30	31
Min. frames number	300	361
Mean frames number	300	1088.7
Max. frames number	300	2541
Total frames	9000	33751
Min. trajectories	44	13
Mean trajectories	169.7	122.8
Max. trajectories	296	1182
Total trajectories	5092	3807

While DroneCrowd was certainly a seminal work that spurred interest in crowd counting and tracking in UAV imagery, the two introduced datasets significantly expand the baseline and provide opportunities for developing algorithms that meet the demands of modern drone-based video analysis.

2.6 Metrics

This section focuses on the metrics applied for tasks such as crowd object counting, point-oriented object localization, and tracking.

2.6.1 Universal metrics

Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. It measures the model's accuracy in identifying relevant objects while minimizing false positives. High precision indicates that most of the detected objects are correct. It is defined by:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.1}$$

where TP (True Positives) and FP (False Positives) are correctly and incorrectly detected objects sequentially.

Recall is the ratio of correctly predicted positive observations to all relevant observations in the dataset. It assesses the model's ability to find all relevant objects, ensuring a low rate of false negatives. It is defined by:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.2}$$

where TP (True Positives) are correctly detected objects, and FN (False Negatives), objects that the model did not detect.

F1-Score is the harmonic mean of Precision and Recall, providing a balanced metric that considers both false positives and false negatives. It is advantageous when the class distribution is uneven or when both precision and recall are equally important. It is defined by:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.3}$$

2.6.2 Localization metrics

L-AP (*Localization Average Precision*) is employed for evaluation of point-oriented methods. It is determined by the distance threshold calculated using the greedy method, as described in the procedure introduced in [86]. The results are reported for three distance thresholds: 10 pixels (L-AP@10), 15 pixels (L-AP@15), and 20 pixels (L-AP@20), along with L-AP, which provides cumulative results in thresholds ranging from 1 to 25 pixels. The metric measures the localization accuracy of the detected points, considering a match between a detected point and a ground truth point if their Euclidean distance is within the specified threshold.

2.6.3 Tracking metrics

HOTA (*Higher Order Tracking Accuracy*) [100] metric assesses multi-object tracking performance by simultaneously evaluating detection and association accuracy, offering a more nuanced assessment of tracking algorithms. HOTA addresses the limitations of traditional metrics, such as MOTA (*Multiple Object Tracking Accuracy*) [101] and ID-F1 (*Identity F1-Score*) (metric of the consistency of identity tracking across a sequence), by integrating them into a unified one that considers detection, association, and localization accuracy collectively.

ID-SW (*ID-Switches*) [102] metric measures how often a tracker incorrectly reassigns an object's ID, resulting in a break in the object's trajectory. This typically occurs due to misassignment or a loss of detection continuity. The ID-SW metric is

crucial for evaluating the consistency of a tracking algorithm in maintaining accurate object identities over time. The score is calculated by counting the total number of identity switches across all sequences.

T-AP (*Tracking Average Precision*) is applied to evaluate the tracking precision of point-oriented methods. It is determined by the distance threshold calculated using the greedy method, as described in the procedure introduced in [86]. Sets of tracks, grouped by identity and ranked by average detection confidence, are considered correct if they match ground-truth trajectories above a specified threshold. The results are also reported for the selected distance thresholds of 10 pixels (T-AP@10), along with T-AP, which provides cumulative results in thresholds ranging from 1 to 25 pixels.

2.6.4 Counting metrics

MAE (*Mean Absolute Error*) is a metric used to measure counting errors for objects in each image of a dataset. It evaluates the accuracy of an algorithm's estimate of the number of people, disregarding the locations of the objects. The metric is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.4}$$

where n is number of images, y represents the ground truth count, and $\hat{y}$ denotes the estimated count.

nMAE (*Normalized Mean Absolute Error*) is a metric similar to MAE, but it normalizes the error by dividing it by the ground truth count. This normalization enables a more accurate comparison of performance across datasets with varying object counts in scenes. It offers an intuitive interpretation by directly measuring the average percentage error in the estimated counts. It is defined as:

$$nMAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2.5}$$

where n is number of images, y and $\hat{y}$ represent the ground truth count and estimated counts.

Tr-MAE (Tracking Mean Absolute Error) is a metric used to evaluate counting errors for object trajectories within each sequence in a dataset. It is specifically designed to validate tracking methods by assessing an algorithm's ability to count

objects effectively while considering object identifications across frames. The metric is defined as:

$$Tr\text{-}MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.6}$$

where n is number of video sequences, y represents the ground truth number of trajectories, and $\hat{y}$ denotes the estimated unique object's count.

Tr-nMAE (Tracking Normalized Mean Absolute Error) calculates the relative counting error, similar to nMAE, but focuses on unique trajectories within a sequence. It can be interpreted as the percentage counting error for the entire sequence (video), providing a measure of an algorithm's performance in tracking and counting objects. It is defined as:

$$Tr\text{-}nMAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{2.7}$$

where n is number of video sequences, y represents the ground truth trajectory count, and $\hat{y}$ denotes the estimated trajectory count for a sequence.

2.7 Summary

Despite significant advancements in object localization and tracking, notable gaps remain, particularly in the context of point-oriented methods. Existing localization techniques often struggle with high-density crowds, occlusions, and the requirement for fine-grained individual identification, especially in dynamic, outdoor environments. Similarly, current point-oriented tracking algorithms face challenges related to temporal consistency. These limitations hinder their applicability to real-world scenarios, such as large-scale public events, where accurate and persistent tracking of individuals is crucial for safety and analysis. Addressing these limitations, the UP-COUNT and UP-COUNT-TRACK datasets are introduced as part of this research. Unlike prior benchmarks, they reflect the complexity of real-world drone deployments by incorporating non-stationary camera motion, a wide range of flight altitudes, and highly varied environments. UP-COUNT supports frame-level localization across both sparse and dense scenes, while UP-COUNT-TRACK extends this by offering long, temporally consistent trajectory annotations. These datasets not only provide significantly more realistic and challenging conditions than existing ones, such as DroneCrowd, but also enable the robust evaluation of both localization and tracking algorithms under diverse, real-world constraints. Their design fills a gap in the field and supports the development of methods better suited to operational applications such as public safety, crowd management, and urban analytics.

3 Point-oriented object localization methodology

The following chapter introduces a general methodology for crowd object localization and counting. It also describes a series of proposed mechanisms and approaches for these tasks, addressing tiny objects in high-resolution image sequences captured by drones. Each of the improvements is evaluated individually, with the final method validation against current state-of-the-art methods in Section 3.8.

3.1 Baseline methodology

Most methods dedicated to both CCTV-based crowd counting (for instance, [103, 14, 78]) and drone-based crowd counting (for instance, [86, 87, 98]) generate density masks as output, despite using different network architectures and various improvements. The high-level abstract implementing this approach is presented in Figure 3.1. Based on an input image, these methods employ deep neural networks

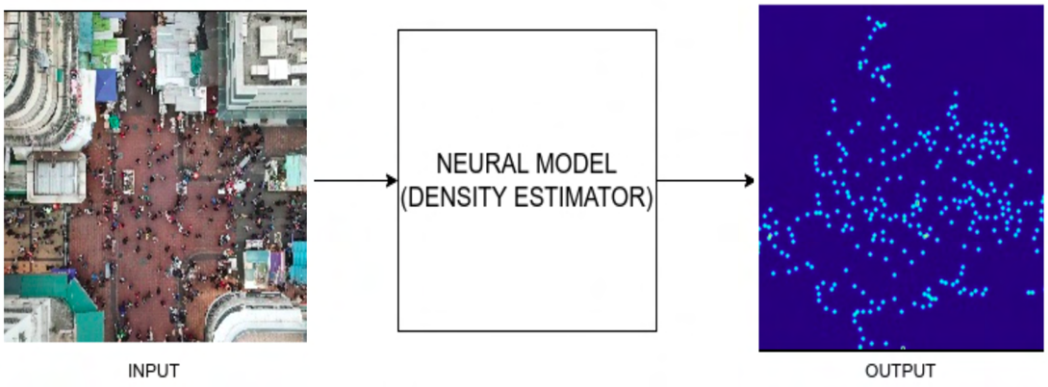


Figure 3.1: The general overview of the approach applied in most crowd counting methods: based on an input image, the neural network generates an output mask whose peaks indicate objects' locations.

to produce a density mask. This mask typically matches the image resolution, with the modeling output values ranging from zero to one, representing the probability of object presence at each pixel. The network's objective is to minimize the values for background pixels while maximizing those corresponding to objects, resulting in peaks that indicate object locations, which can be refined through post-processing.

Figure 3.2: Example images from a large crowd counting dataset with generated density masks [104].

This approach enables pixel-level accuracy, typically yielding better results in the localization of densely packed, tiny objects. Example frames with generated masks for the large crowd counting dataset, ShanghaiTech [104], are presented in Figure 3.2.

This approach is commonly referred to as a density estimation task and, similar to an image segmentation task, is frequently implemented using an encoder-decoder architecture, such as UNet [88]. This architecture implements a skip connection mechanism between successive stages, enabling the better encoding of spatial correlation features [105]. Its encoder functions to distill the input into a compact latent representation, effectively capturing its intrinsic structures and patterns. Subsequently, the decoder reconstructs the density map from this latent space, projecting it back to the output domain. This makes the neural network architecture appropriate for modeling accurate and pixel-level responses. Moreover, its modularity enables the replacement of the encoder block by others, including ResNet [106], MobileNet [107], and EfficientNet [108] architecture families. Recently, the most prominent feature extraction model is the Mix Vision Transformer (MiT), proposed as the encoder of SegFormer [109]. MiT employs a lightweight, transformer-based hierarchical representation that can produce CNN-like multi-scale features, offering enhanced representational capacity while maintaining compatibility with conventional encoder-decoder architectures.

3.2 Motion-enhanced image analysis

The amount of information that can be extracted from a single image is limited. Therefore, it is common to enrich them with additional information, especially those that extend the context of the processing. Unlike individual images, image sequences offer additional temporal features, which can be extracted through techniques such as motion analysis. This process models pixel changes between frames, calculating movement direction and amplitude, thereby providing a richer set of information compared to static images, which highlights both static and dynamic objects. Motion between successive frames can be estimated using dense optical flow, which captures pixel-wise displacement. One efficient approach for computing optical flow is the DIS (Dense Inverse Search) method [110]. While numerous deep learning-based methods for optical flow estimation have been proposed [111], they often require substantial computational resources, particularly for high-resolution data. In contrast, DIS provides a practical balance between speed and accuracy, allowing computational capacity to be preserved for additional functionalities, including integration with neural networks. Example optical flow-based motion estimations are placed in Figure 3.3. It presents arrows that represent dense motion estimations calculated between successive frames.

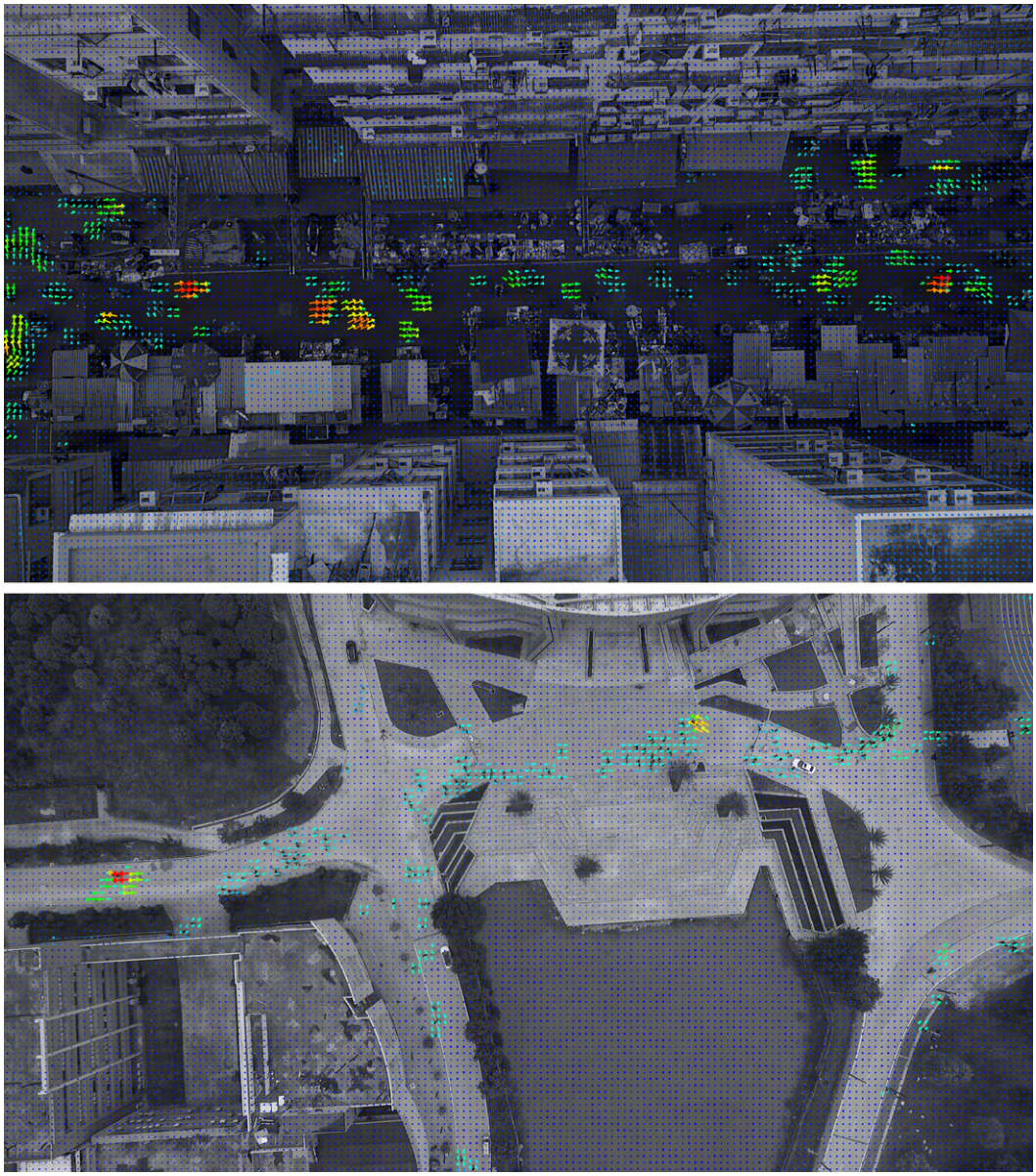


Figure 3.3: Examples of dense motion estimations in frames.

3.2.1 Motivation

In-frame motion information, computed using dense optical flow, is integrated to supplement the raw image data. The resulting displacement maps emphasize moving objects, enhancing the algorithm's focus and effectiveness in detecting dynamic elements.

3.2.2 Method

In the research [112], a pipeline incorporating dense optical flow for tiny object counting is proposed. This architecture provides an efficient approach that

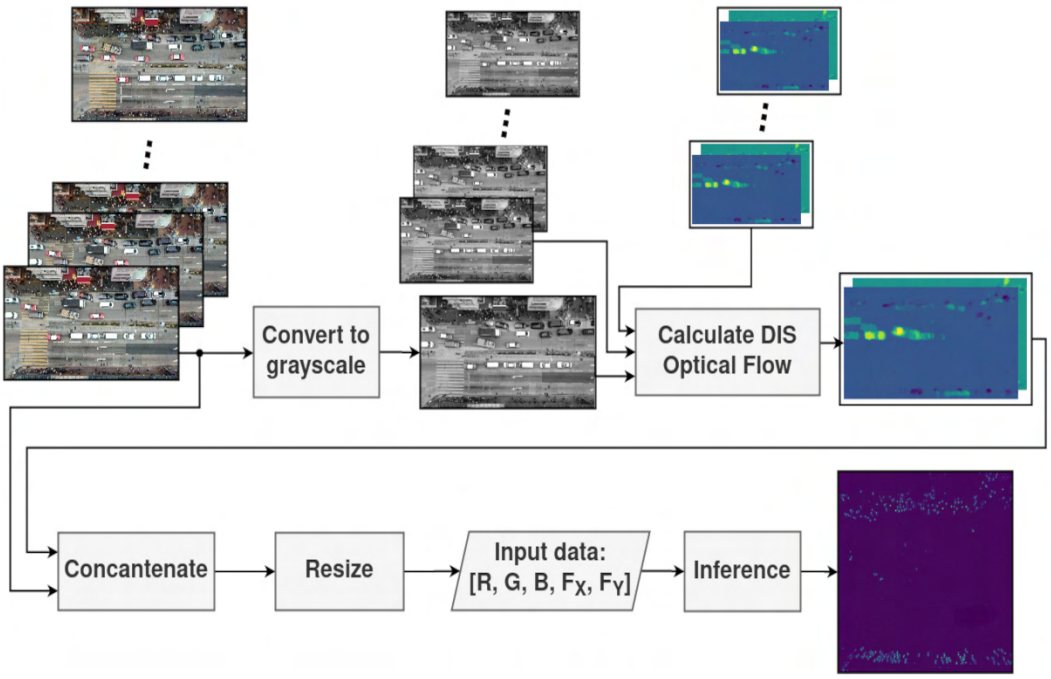


Figure 3.4: Example of computer vision pipeline involving dense optical flow along with visual data [112].

calculates motion matrices between subsequent frames and integrates them with visual features for better video processing. The DIS method is applied to compute motion information for each pixel along the X and Y axes, providing two-dimensional matrices with the same shape as the input image. These computed features are normalized and concatenated with the RGB image channels, providing additional context information by extending them with motion values. An example system's architecture, detailed in Figure 3.4, consists of three stages. First, the frame is retrieved from the video source and converted to a grayscale image. Second, the system computes the DIS optical flow using the current and previous frames. Finally, the computed optical flow matrices are concatenated with the color image channels and passed into the neural network, which outputs an estimated density mask.

3.2.3 Evaluation

To investigate the impact of using motion-enhanced input images for the crowd counting task, a neural network with and without an additional channel is evaluated. The results are presented in Table 3.1. The table reports absolute counting errors for a five-fold dataset division and their average, considering the DroneCrowd dataset. In all five dataset splits, incorporating motion matrices consistently led to lower errors, reducing the average from 22.05 (RGB only) to 21.04 (RGB + motion). While the average improvement is modest (approximately 4.6%), the consistent

error reduction across all splits indicates that motion information serves as a valuable cue for crowd counting.

Table 3.1: Comparison of the counting error (mean absolute error) in crowd counting, including and excluding motion matrices.

Dataset split	Only RGB input	RGB + motion input
1	18.56	17.81
2	20.11	19.57
3	26.36	25.15
4	18.77	17.13
5	26.45	25.56
Average	22.05	21.04

Conclusions: While the evaluation demonstrates that dense optical flow can effectively enhance input images, particularly by providing motion cues, this approach encounters significant limitations under specific conditions. When the video source, such as a drone camera, is in motion, the generated motion matrices become noisy and fail to accurately capture object movement. Therefore, the method's effectiveness in enhancing images based solely on object motion is mainly limited to applications where the camera remains stationary or exhibits only minimal movement, ensuring that the calculated flow predominantly corresponds to the actual scene dynamics.

3.3 The importance of image input resolution

Training deep-learning neural networks is a time-intensive and computationally demanding process. Limited processing resources often necessitate optimizations to address these constraints. A common approach to alleviate this issue is to reduce image resolution to fit within available memory. This optimization is especially relevant in tiny object remote sensing, where images are typically high-resolution. However, scaling down images can result in the loss of fine details, which can impact the accuracy of the analysis's results. Another optimization involves selecting a more computationally efficient network architecture by considering factors such as the number of parameters, floating point operations per second (FLOPS), and layer design choices. For example, using separable convolutions [113] instead of traditional convolutions can significantly reduce computational demands. Such architectural optimizations enable the use of higher-resolution input images. It allows for evaluating whether a smaller network with higher spatial resolution can improve evaluation metrics, emphasizing the importance of resolution over model size.

50:2691698471

3.3.1 Motivation

The use of higher input image resolution, rather than increasing model complexity through additional parameters, facilitates more effective feature extraction, leading to improved metrics and overall performance. Consequently, reducing image resolution often results in the loss of details, particularly for tiny objects.

3.3.2 Method

In the research [112], U-style neural networks for crowd counting tasks are evaluated using various input resolutions. To investigate the general trend of input resolution impact on neural networks, multiple architectures and feature extractors are utilized, ensuring wide representations. The study focused on segmentation-based encoder-decoder architectures, including the most popular ones, such as UNet [114], UNet++ [115], and DeepLabV3+ [116]. Widely adopted architectures that meet the criteria of efficient processing are examined as encoder backbones (feature extractors). The evaluation began with the ResNet [106] family due to its strong generalization capabilities across various applications [117]. Further, EfficientNet [108] is evaluated, which is recognized for its state-of-the-art performance in specific tasks. Additionally, MNASNet [118], specifically the Squeeze-and-Excitation variant (SEMNASNet), and MixNet [119] are included in the analysis, as both achieve strong performance metrics, particularly in resource-constrained environments such as mobile platforms.

3.3.3 Evaluation

The evaluation is conducted on the DroneCrowd dataset and assesses the MAE metric in drone-based crowd counting. The results are presented in Figure 3.5, which includes plots comparing different models and feature extractors across varying input image resolutions: 480×288 (R1), 640×384 (R2), 960×540 (R3), and 1280×736 (R4). Regardless of the model architecture, all results consistently demonstrate a clear trend: increasing input resolution leads to a decrease in the crowd counting error. Although the general trend remains, the magnitude of MAE reduction varies depending on the specific model and feature extractor.

Conclusions: Higher-resolution images usually preserve more details, enabling models to recognize objects better. This is an essential factor in tasks such as counting people in crowds, where fine details significantly impact accuracy. At lower resolutions, critical information is lost, leading to higher counting errors. Therefore, in tiny object localization, it is crucial to consider using higher-resolution

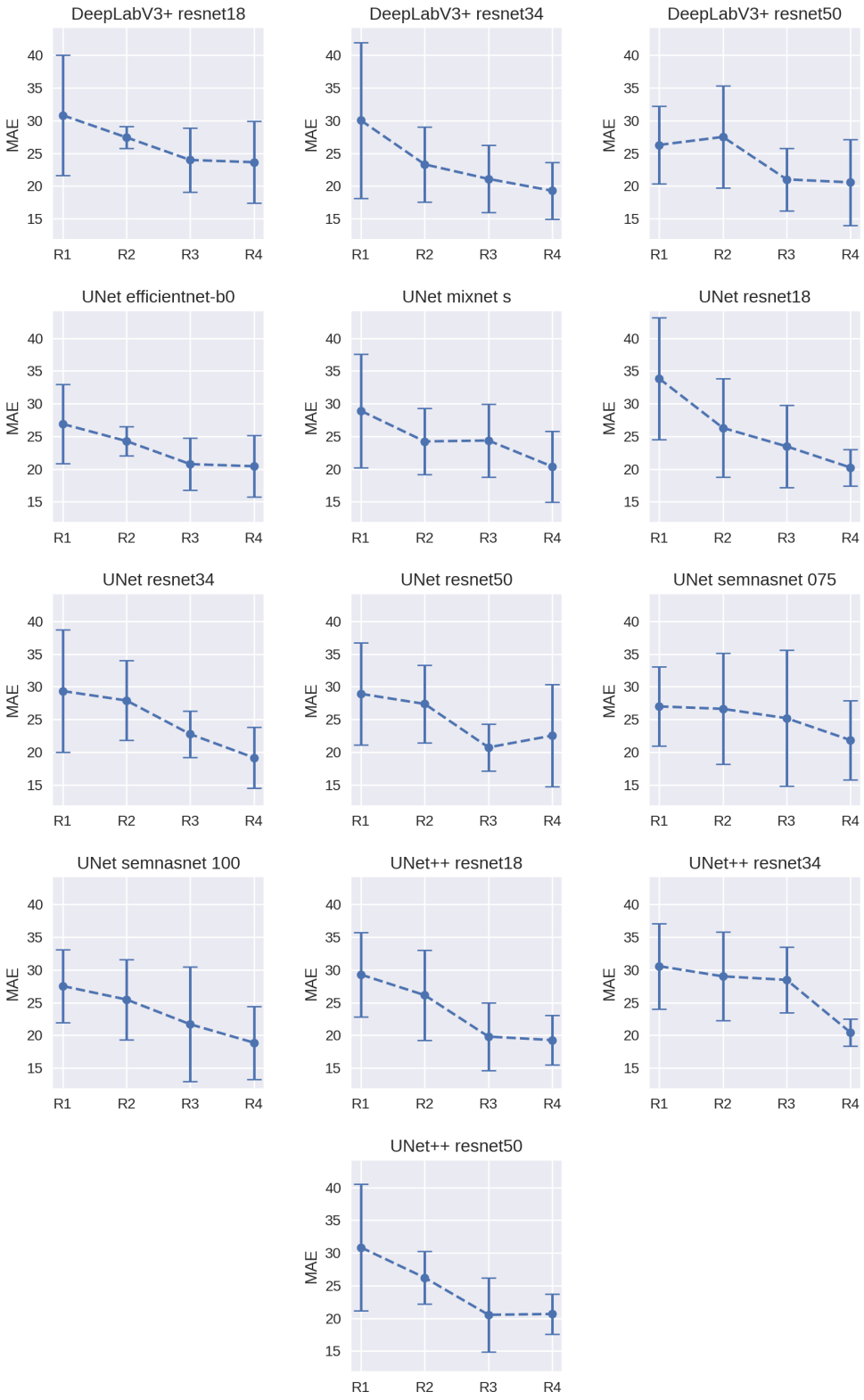


Figure 3.5: The comparison of models' architectures for various input image resolutions [112]. R1 to R4 correspond to the following input sizes: 480×288, 640×384, 960×540, and 1280×736, respectively.

inputs, even if it necessitates employing a model with fewer parameters, while also accounting for the computational requirements of such an approach.

3.4 The impact of synthetic data

Large amounts of data are essential for enabling neural networks to generalize effectively. For example, the ImageNet dataset [120], one of the largest and most comprehensive for image classification, comprises approximately 14 million images distributed across 21,841 categories. Similarly, the COCO dataset [121], developed for object detection tasks, contains around 328,000 images with over 1.5 million object instances spanning 80 categories. However, available datasets for many computer vision tasks are limited, often due to the challenges of labeling or the low prevalence of specific topics. This limitation is particularly evident in tasks like drone-based people localization, where insufficient data increases the risk of overfitting during the training process.

To address this gap, numerous studies have utilized synthetic data to augment existing datasets. For example, in [122], the authors introduced a synthesizer capable of generating high-quality object detection data for novel domains, effectively mitigating performance degradation in object detectors caused by significant domain shifts in agricultural scenes. Similarly, the authors of [123] utilized a video game to develop a data collector with labeled outputs, enabling the generation of synthetic crowd scenes that replicate those captured by video surveillance cameras.

3.4.1 Motivation

Generating a synthetic dataset using a game engine and incorporating it into the neural network training process enhances performance metrics in the drone-based people-counting task.

3.4.2 Method

In the research [124], a synthetic dataset for a drone-based crowd counting task is developed using the Unity3D game engine [125] and the Perception toolkit [126]. This process involved the development of a simulation capable of producing synthetic images paired with corresponding ground-truth annotations. The simulated environment is based on a city map primarily reflecting urban settings, such as downtown districts, public squares, and parks. Independent individuals were randomly deployed into a city environment, compelled to navigate using a mesh network, and

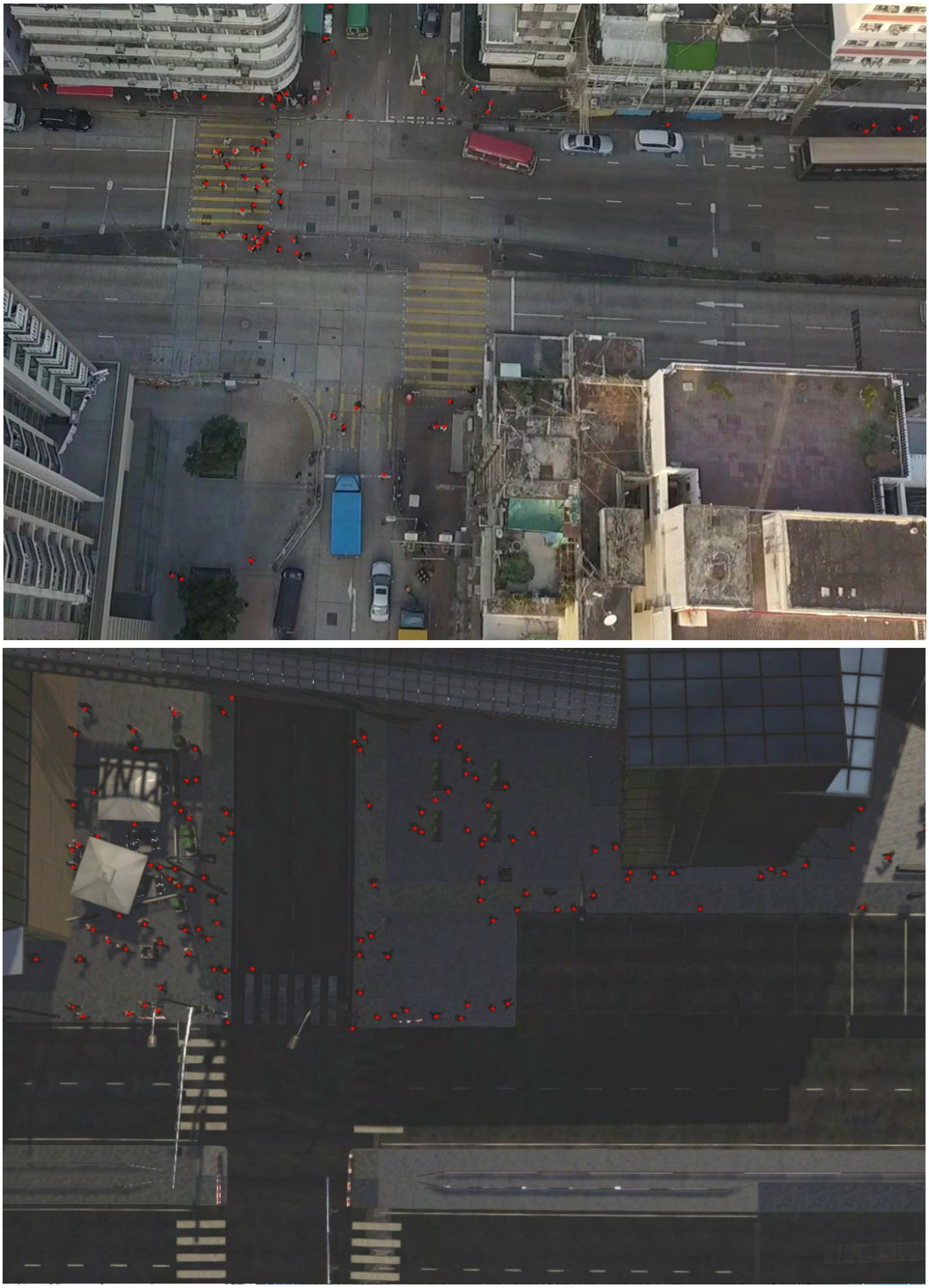


Figure 3.6: The top image shows a sample from the DroneCrowd dataset with human heads annotated with red dots. The bottom image demonstrates samples from the new synthetic dataset [124].

were tasked with generating their routes between destinations. This setup enabled the simulation of city-like crowd dynamics and the inclusion of non-typical scenarios, such as protests or events. Data acquisition is carried out using eight cameras,

mimicking aerial vehicle-mounted setups, concurrently capturing the ground-truth coordinates of individuals.

The simulator was run multiple times with different setups to generate a total of 65 unique sequences, each characterized by variations in the number of people, camera altitudes, and lighting conditions. The dataset is divided into training, validation, and test subsets, each comprising distinct sequences. These subsets contain 155,203, 16,648, and 16,018 frames, respectively. Figure 3.6 illustrates an example of a generated frame alongside a comparison with an image from the DroneCrowd dataset.

3.4.3 Evaluation

To investigate the influence of synthetic data on the crowd counting task, an experiment is conducted that covers various network architectures and strategies. Using UNet [114], different encoders, including ResNet [106], SEMNASNet [118], and EfficientNet [108] families, are evaluated, providing a wide representation of neural network architectures. To conduct a comprehensive evaluation of the interaction between synthetic data and various feature extraction paradigms, four distinct training and evaluation strategies are proposed. These strategies are designed to assess the influence of pretraining sources and architectural components on model performance in the downstream task. Specifically, the following are considered:

- Full fine-tuning with ImageNet initialization: The entire network is trained end-to-end, with weights initialized from a model pretrained on the ImageNet dataset. This strategy serves as a baseline, leveraging generic visual features acquired from large-scale natural image data.

- Full fine-tuning with synthetic pretraining: The complete network is trained end-to-end, using weights initialized from a model pretrained exclusively on synthetic data. This allows for an assessment of how well synthetic pretraining transfers when the entire network is adapted to the target task.

- Partial fine-tuning with frozen encoder: Only the decoder (task-specific head) of the network is trained, while the encoder (feature extractor) is kept fixed with weights derived from synthetic pretraining. This configuration evaluates the quality and transferability of the learned feature representations from synthetic data, independent of downstream task-specific adaptations in the encoder.

- Output head tuning with frozen encoder and decoder: The encoder and decoder components are frozen, and only the final output head is trained. This most constrained setting tests the expressiveness and adaptability of the pretrained model's final-stage features when minimal task-specific tuning is allowed.

These configurations facilitate a systematic investigation into the extent to which synthetic data can serve as a viable substitute or complement to real-world data in feature extraction and model generalization.

Table 3.2: The comparison of various models with different weights initialization and freezing methodologies. The people counting mean absolute error (MAE) is measured.

Encoder	Initial weights and training mode			
	ImageNet (full training)	Synthetic (full training)	Synthetic (freeze encoder)	Synthetic (trainable model head)
resnet18	25.049	23.504	**23.048**	28.295
resnet34	22.053	22.037	**20.214**	56.272
semnasnet-075	20.385	22.026	**17.753**	63.828
semnasnet-100	21.561	**21.307**	22.639	149.150
efficientnet-b0	24.684	28.796	**22.984**	74.165

The evaluation results are presented in Table 3.2, considering the MAE metric and DroneCrowd dataset. The results reveal a clear trend: initializing with pretrained synthetic data weights and freezing the encoder consistently yields the lowest MAE across most architectures, achieving MAE values of 23.048, 20.214, 17.753, and 22.984 for ResNet18, ResNet34, SemNASNet-075, and EfficientNet-B0 architectures, respectively.

Conclusions: Weight initialization using synthetic data that is specifically generated for the target task yields better performance metrics compared to general ImageNet weights. This empirical finding indicates that feature representations learned from task-specific synthetic data exhibit a high degree of relevance and alignment with the downstream objective of people counting. The results highlight the strong transferability of features learned in-domain, suggesting that synthetic data can serve not only as a viable substitute for real-world data in pretraining but also as a highly effective mechanism for enhancing model robustness and task-specific generalization.

3.5 The integration of drone's sensor data

Modern aerial platforms typically feature various sensors, such as barometers, accelerometers, gyroscopes, and satellite navigation systems. However, despite

their widespread availability, data from these onboard sensors are seldom leveraged in computer vision-based remote sensing algorithms. Among these underutilized data sources, the drone's altitude relative to ground level stands out. While prior research has proposed various strategies to mitigate performance issues caused by scale variation in remote sensing imagery, none have directly exploited altitude measurements from UAV sensors to enhance the detection of tiny objects, despite the accessibility of such data.

Within the domain of remote sensing, extensive work has been conducted on data fusion methodologies. For instance, the study in [127] investigates the integration of imagery from multiple sensors to evaluate the potential of multisource data fusion. A different strategy is employed in [128], where a two-phase approach is proposed. Initially, a deep learning model is trained to estimate the (GSD (Ground Sampling Distance)) from imagery, followed by the fusion of the model's latent feature vectors to improve object detection. Notably, none of these approaches have examined the direct incorporation of altitude and GSD information – despite their availability from onboard systems – which represents an untapped opportunity for improving detection performance.

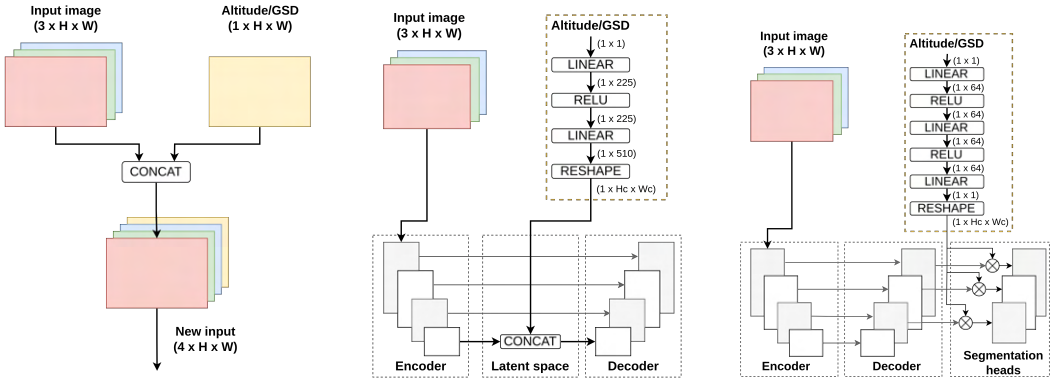
3.5.1 Motivation

Incorporating drone altitude information from sensors directly into the neural network provides valuable features that enhance performance metrics.

3.5.2 Method

In the research [129], three distinct methods for incorporating altitude information into the model are proposed and analyzed. Each method augmented the input images with additional features derived from the drone's altitude sensor. Furthermore, the effectiveness of using both raw altitude data and GSD is evaluated to identify the most efficient fusion strategy. The altitudes range from 26.0 to 101.0 meters, with an average altitude of 60.3 meters. The GSD is calculated based on the drone's camera calibration parameters along with corresponding altitude information. Employing a consistent baseline architecture and training procedure, three different fusion approaches are evaluated, each varying in network architecture and the method of delivering altitude information. A schematic overview of these methods is presented in Figure 3.7, with detailed descriptions provided below.

- **Method A (Figure 3.7a).** This approach incorporates scalar metadata (e.g., normalized altitude or GSD) directly into the input tensor by appending it



(a) Method A – fusion of raw scalar information as an additional channel of an input image.

(b) Method B – fusion of scalar information as learnable features in the latent space.

(c) Method C – fusion of scalar information at the decoder stage via feature modulation.

Figure 3.7: Overview of altitude information fusion strategies [129]. ALT: altitude; GSD: Ground Sampling Distance.

as an additional channel alongside the RGB channels. It represents the most computationally efficient and architecturally simple fusion strategy, enabling early integration of contextual information into the feature extraction process.

- **Method B (Figure 3.7b).** In this method, scalar information is projected into the latent space as a learnable one-channel embedding. This embedding is then spatially upsampled to match the resolution of intermediate latent feature maps and concatenated as an additional feature channel. This mid-level fusion allows the network to learn task-relevant interactions between image features and scalar context representations within the latent space.

- **Method C (Figure 3.7c).** This method introduces scalar information at the decoder level through feature modulation. Specifically, at each stage of the decoder, the feature maps are element-wise multiplied by a learned embedding of the normalized altitude or GSD. This late fusion technique provides a mechanism for adaptive conditioning of the decoder's spatial features based on external metadata, enhancing the network's ability to leverage contextual cues during prediction.

3.5.3 Evaluation

The experimental results are summarized in Table 3.3. When evaluating performance with respect to altitude, all tested methods outperform the baseline, which achieves an F1-score of 0.701. Specifically, Method A attains an F1-score of 0.705, Method B reaches 0.711, and Method C yields 0.705. In contrast, when considering the influence of Ground Sampling Distance, the highest F1-score of 0.712 is achieved

by both Method A and Method B, while Method C follows closely with a score of 0.710. Overall, the use of Ground Sampling Distance consistently outperformed the consideration of raw flight altitude data, indicating its higher influence on the effectiveness of the evaluated methods.

Table 3.3: Methods comparison considering raw altitude and GSD (Ground Sampling Distance) fusion.

	Altitude			GSD		
Method	**Precision(↑)**	**Recall(↑)**	**F1-Score(↑)**	**Precision(↑)**	**Recall(↑)**	**F1-Score(↑)**
Baseline	0.697	0.725	0.701	0.697	0.725	0.701
Method A	0.721	0.701	0.705	0.698	0.737	**0.712**
Method B	0.731	0.703	0.711	0.700	0.734	**0.712**
Method C	0.685	0.737	0.705	0.712	0.718	0.710

Conclusions: All evaluated fusion strategies exhibit an improvement in performance relative to the baseline model, underscoring the effectiveness of incorporating auxiliary scalar information, such as altitude or GSD, into the learning process. These results suggest that enriching the visual input with contextual metadata enhances the model's representational capacity and contributes to improved predictive accuracy. Furthermore, a comparative analysis reveals that the inclusion of GSD yields consistently better performance metrics compared to the use of raw altitude information. This indicates that GSD, as a normalized and task-relevant spatial resolution descriptor, provides more informative and discriminative cues for the model than raw absolute altitude values.

3.6 The importance of the usage of every pixel in an image

As discussed in Section 3.3, the resolution of the input image has a significant impact on the performance of neural network models. Ideally, the network would process full-resolution images, utilizing features from every pixel. However, this approach is constrained by limited computational resources, particularly memory, which is required to store large arrays of model weights. To address these challenges, researchers have proposed various optimizations aimed at reducing memory usage and computational costs, with a particular focus on image super-resolution tasks [130]. These methods often address the difficulty of the increasing image resolution with deep learning capabilities. In many cases, these approaches aim to balance the trade-off between model complexity and performance, leveraging innovations such as dedicated mathematical operations, knowledge distillation, and memory-optimized network architectures. A key focus lies in mitigating the exponential increase in computational demand associated with higher image resolutions,

which places a substantial burden on computing machines that struggle to maintain high processing speeds and efficient memory usage under such conditions.

3.6.1 Motivation

Using full-resolution images in the initial layer of the neural network enables improved accuracy without significantly compromising performance.

3.6.2 Method

In the research [98], a dedicated module is proposed that enables full-resolution image processing, allowing for the processing of high-definition images by taking into account every pixel value. In contrast to a sliding-window approach, which divides images into tiles and processes them separately, and the interpolation approach when the image resolution is reduced, the proposed method processes the entire image at once without reducing its resolution. Thanks to operating on all pixels simultaneously, processing high-resolution images with the proposed module is more computationally efficient, and the neural network considers all the image's pixels. This avoids information loss that occurs with downsampling interpolation, especially in the tiny objects localization task.

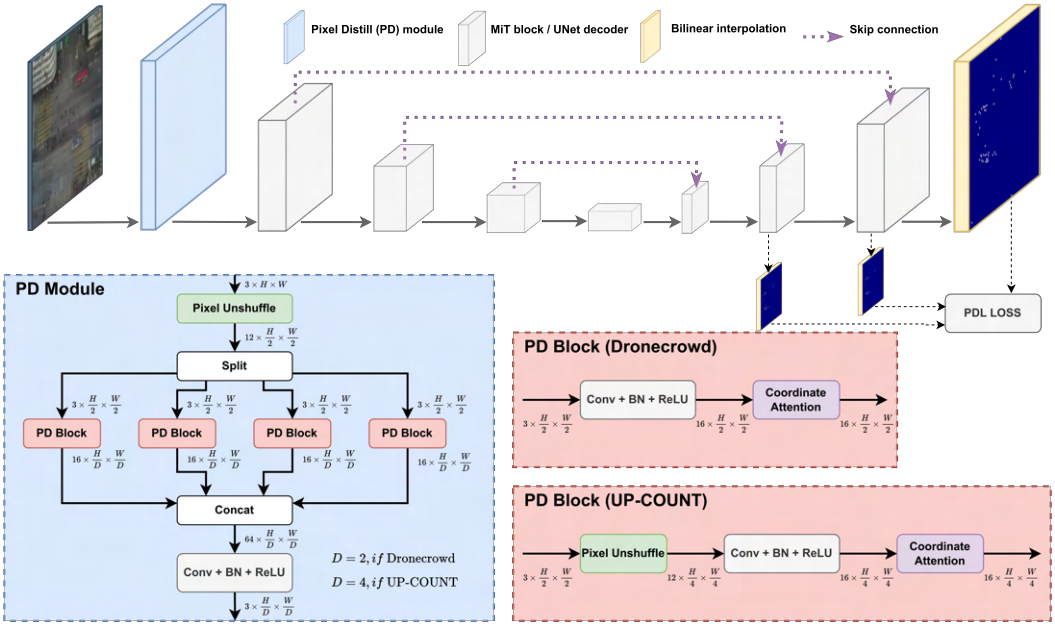


Figure 3.8: The schema of PD (Pixel Distill) module, including blocks comprising it, and the overview of PD in the deep network architecture [98].

The design architecture of the proposed module, called PD (Pixel Distill), is illustrated in Figure 3.8. It represents a versatile architecture for high-resolution

60:1128600889

image processing, effectively extracting relevant features while preserving spatial information. At the initial stage, a PixelUnshuffle operation is employed to halve the input resolution while increasing the number of channels. Originally introduced in [131] for super-resolution applications, this layer enables the use of reduced filter sizes without compromising contextual coverage, thereby improving computational and memory efficiency. Subsequently, the resulting tensor is divided along the channel axis to yield four downsampled sub-images, each independently processed through a dedicated PD block. This partitioning strategy facilitates the extraction of diverse spatial features, enhancing the expressiveness of the final feature representation compared to a single block with increased filter capacity. Two PD block variants have been developed: one optimized for Full HD datasets, such as DroneCrowd, and the other for 4K resolution datasets, like UP-COUNT. These variants standardize the output feature size regardless of input resolution, accommodating the differing spatial reduction requirements to ensure feature extraction occurs at an appropriate scale for each dataset. In the first variant, the sixteen-channel feature maps are extracted from the image using a convolutional operation followed by BN (Batch Normalization), and the ReLU (Rectified Linear Unit) activation. Next, the positional information is extracted and multiplied with feature maps using the Coordinate Attention mechanism [132], which generates direction-aware, position-sensitive attention maps to enrich the semantic representation of target objects. The second variant of the PD block is preceded by an additional PixelUnshuffle layer for further resolution reduction before processing. Following this multi-branch processing, all feature maps are concatenated along the channel axis. A final block, comprising convolution, BN and ReLU layers, then performs the necessary channel reduction to align with the model's architectural requirements.

3.6.3 Evaluation

To evaluate the proposed module, a comparison of the results on DroneCrowd and UP-COUNT datasets is performed. It includes the point-oriented object localization metric – L-AP. Three distinct preprocessing strategies are considered:

- Downsampled Input: The original image is resized using interpolation techniques, resulting in a twofold reduction in resolution for the DroneCrowd dataset and a fourfold reduction for the UP-COUNT dataset. This simulates the most often applied approach, resulting in low-resolution input. It serves as a baseline.

- Sliding Window Tiling: The input image is divided into overlapping patches using a sliding window approach. This strategy preserves local detail but

may introduce boundary artifacts and increased computational cost due to patch-wise processing.

- Proposed PD (Pixel Distill) Module: The input image is preprocessed using the proposed module, designed to preserve spatial details while reducing the computational burden adaptively. This approach aims to maintain high localization accuracy by enhancing feature representation in low-quality or complex regions.

All cases use the exact input resolution for the neural network input, which is 960×544. The results in Table 3.4 show that the proposed PD (Pixel Distill) module outperforms classical processing methods on both datasets. In contrast to conventional image resizing, which often loses fine-grained features, PD enhances performance by preserving detailed spatial information while simultaneously reducing computational overhead. On the DroneCrowd dataset, PD yields modest gains over the sliding-window strategy, achieving L-AP and L-AP@10 scores of 51.00 and 57.06, respectively. The performance gains are more substantial on the UP-COUNT dataset, where PD surpasses both the resizing and sliding-window baselines, attaining the highest L-AP (66.49) and L-AP@10 (75.46) scores. These results highlight the effectiveness of pixel-wise processing via the proposed module in enhancing feature extraction, particularly under high-resolution conditions. Moreover, the findings validate that PD offers a computationally efficient framework for high-resolution image analysis, avoiding the significant memory demands associated with traditional methods.

Table 3.4: Performance comparison of using Pixel Distill against classical processing methods. L-AP and L-AP@10 are evaluation metrics, with higher values indicating better model accuracy.

Method	**DroneCrowd**		**UP-COUNT**	
	L-AP(↑)	**L-AP@10(↑)**	**L-AP(↑)**	**L-AP@10(↑)**
Image scaled down	47.63	53.37	60.66	69.07
Using sliding window	50.73	55.61	54.17	58.74
Using Pixel Distill	51.00	57.06	66.49	75.46

Conclusions: As demonstrated by the proposed PD (Pixel Distill) module, which yields improved performance metrics on both the DroneCrowd and UP-COUNT datasets, it is feasible to enhance localization accuracy, particularly for tiny and densely distributed objects, without incurring substantial computational overhead. By effectively preserving fine-grained spatial details in full-resolution images, the PD module enables more accurate point-based object localization. This highlights its potential as a valuable preprocessing strategy for high-resolution and high-density visual scenes, contributing to advancements in the challenging task of tiny object localization.

3.7 Dedicated loss function for the point-oriented localization task

In the research [98], the common MSE (*Mean Square Error*) loss function is observed to lead to overfitting in neural networks when used for point-oriented localization tasks. MSE is sensitive to high-density areas, resulting in large errors and contributing significantly more to the total loss. In that case, the network might focus excessively on minimizing errors in these high-density peaks or specific noisy spots, potentially reducing performance in lower-density regions or failing to generalize in different spatial configurations. Furthermore, when training datasets are insufficient, such as in drone-based crowd counting, the evaluation metric can inadvertently lead the model to focus on specific density patterns because it lacks the data to develop proper, robust structural awareness.

Considering these observations, a loss functions that address the challenges posed by the point-oriented localization task in low-altitude aerial imagery is investigated.

3.7.1 Motivation

Applying a tailored loss function rather than commonly used ones for the point-oriented localization task results in more refined output masks from the neural network, thereby improving the distinction between objects and the background and enhancing performance metrics.

3.7.2 Method

In the research [98], a novel loss function named Point Distance-aware Localization is proposed. The loss is dedicated to drone tiny object localization with pixel-level precision, generating input image-size masks where zero represents the background and one indicates the object position. It is defined as a combination of three factors:

$$L_{total} = 0.25 \cdot L_{neg} + L_{obj} + 2 \cdot L_{reg} \tag{3.1}$$

where L_{neg} is the modified Focal loss; L_{reg} and L_{obj} are respectively objectness and regression losses described below. The constant coefficients were selected empirically.

- **Modified Focal Loss:** object locations are annotated with positive values, whereas background regions are assigned negative values. Following the

approach of [133], a modified variant of the Focal Loss [134] is employed. Specifically, for each positive pixel within the mask, the loss function generates a Gaussian-shaped neighborhood, thereby attenuating the penalty associated with false negative predictions that occur near true positives. During the early training stages, the negative loss component (L_{neg}) dominates and rapidly diminishes over successive epochs, allowing other loss components to gain influence. This strategy is designed to enhance the separation between object and background values in the mask at the initial phase of training, which facilitates convergence and contributes to training stability.

- **Point Localization Loss:** Drawing inspiration from the object detection loss formulation presented in [135], which decomposes the detection task into object presence and bounding box localization, a dual-component loss function tailored for point-based object detection is introduced. The objectness component (L_{obj}) leverages the Binary Cross-Entropy criterion to evaluate the correspondence between the predicted and target masks. This term accentuates the contrast between object and background regions, thereby improving detection confidence. Complementarily, the localization component (L_{reg}) aims to minimize the spatial discrepancy between predicted and ground-truth object locations. This is achieved through the Euclidean distance (L2 norm), calculated between the predicted coordinates $(\hat{x}, \hat{y})$ and the corresponding ground truth (x, y), ensuring precise point-level localization.

To train a neural network for the object localization task using MSE, commonly, the heatmap is generated as a ground-truth mask. It requires the use of a Gaussian filter to reduce localization ambiguity and create a more informative and easier-to-learn target signal for the network, thereby improving the stability of the training process. However, it results in the generation of less precise output masks and an issue of objects overlapping. In contrast, the use of Point Distance-aware Localization loss, which employs modified focal and point localization losses that operate on pixel-level binary masks, enables the generated output mask to be more contrastive, thereby achieving higher precision. The motivation for using the combination of these two loss functions is to maximize the difference between the background space (zeros) and the label space (ones). The example of both the heatmap and the new approach masks is presented in Figure 3.9.

3.7.3 Evaluation

To validate the influence of the proposed loss function, a performance comparison is conducted using various neural network encoders on both the DroneCrowd and UP-COUNT datasets. Its results are presented in Table 3.5. The evaluation uses

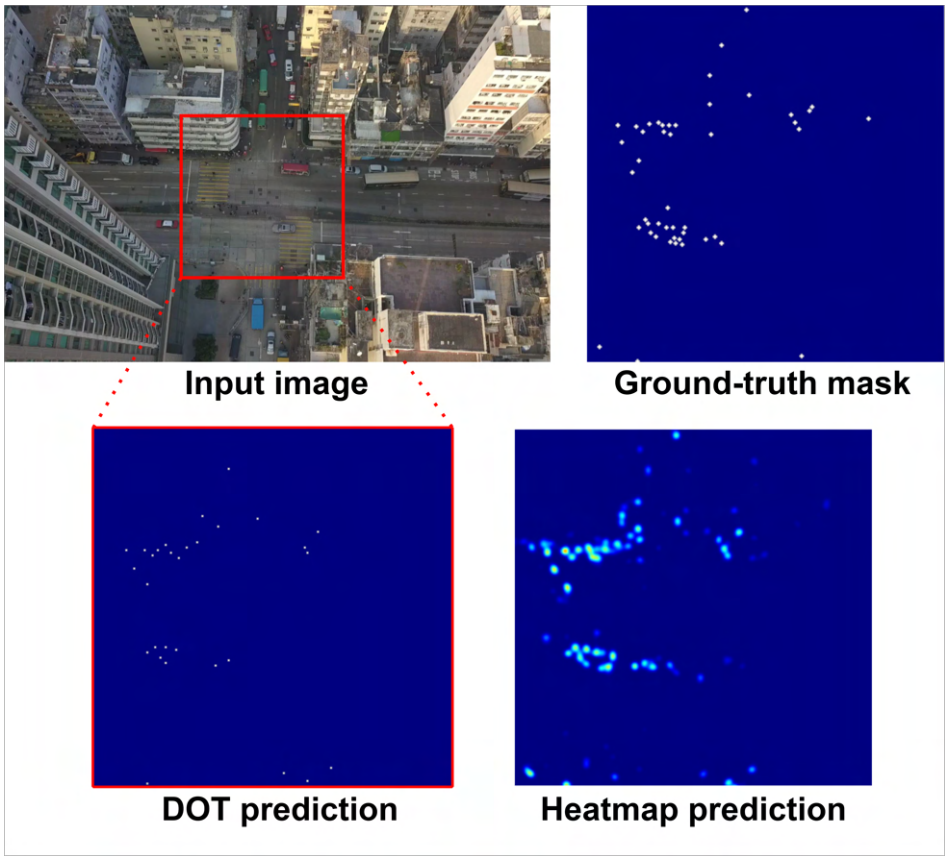


Figure 3.9: Example of mask prediction of a neural network using the proposed approach and heatmap.

L-AP as the default metric for the object localization task, considering a threshold of ten pixels (L-AP@10) and a cumulative metric between 1 and 25 pixels (L-AP). For all evaluated encoders, the proposed loss function consistently yields higher L-AP scores: 45.83 vs. 34.35, 47.20 vs. 36.13, and 47.63 vs. 38.08 on the DroneCrowd dataset; and 60.66 vs. 49.46, 60.34 vs. 50.98, and 62.83 vs. 52.24 on the UP-COUNT dataset. These results demonstrate a significant improvement in the people localization task.

Table 3.5: Evaluation results of different loss functions, considering various networks' encoders, and datasets.

Network's encoder	Loss	DroneCrowd		UP-COUNT	
		L-AP(↑)	L-AP@10(↑)	L-AP(↑)	L-AP@10(↑)
ResNet50	Heatmap	34.35	38.37	49.46	55.29
	Proposed	45.83	50.94	60.66	69.07
EfficientNetB2	Heatmap	36.13	40.30	50.98	57.53
	Proposed	47.20	52.46	60.34	67.63
MiT B2	Heatmap	38.08	42.51	52.24	58.78
	Proposed	47.63	53.37	62.83	70.01

During experimentation, it is also observed that the use of Modified Focal Loss in the early training stages facilitates convergence, as it encourages the network

to start predicting significant points on the mask. This phenomenon is visualized in Figure 3.10. As the influence of this term diminishes, giving way to the Point Localization Loss, the objectness and regression components take precedence, leading to improved performance and higher scores. The inclusion of the stabilization term addresses the observation that, without the Modified Focal Loss, the training process required many epochs before the Point Localization Loss could be effectively reduced.

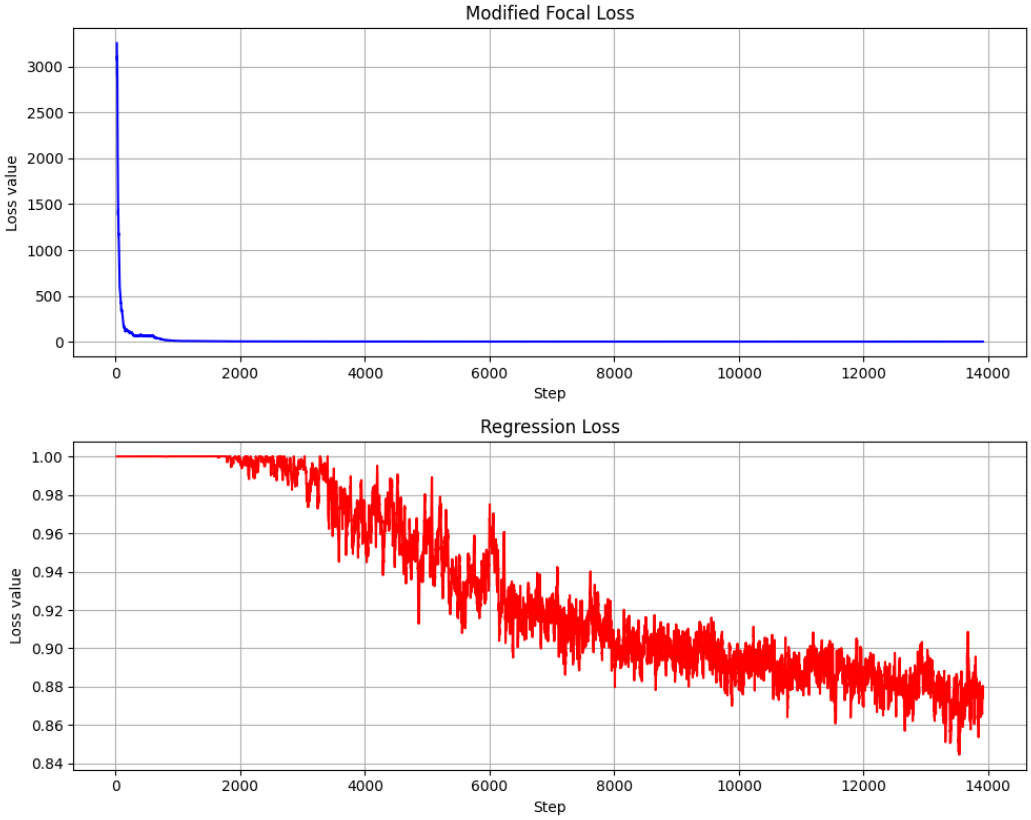


Figure 3.10: The plot illustrates the Modified Focal Loss (first row), demonstrating a rapid decrease in the initial steps, indicative of fast convergence towards predicting key mask points. In contrast, Regression Loss (second row) fluctuates within a narrow range before exhibiting a consistent reduction.

Conclusions: The use of a dedicated loss function that omits the need for generating Gaussian density masks, while directly optimizing object position predictions, has been shown to produce significantly improved performance metrics compared to conventional loss functions commonly employed in individual localization tasks. By focusing explicitly on spatial precision rather than relying on intermediate density estimation, this approach enhances the model's ability to accurately localize discrete objects, particularly in crowded or high-density scenarios.

3.8 Final evaluation of point-oriented localization method

As shown in Table 3.6, the proposed method has been evaluated against three categories of existing approaches. In comparison to STNNet [86] and MFA [87], the former state-of-the-art methods on the DroneCrowd dataset, the proposed approach achieves improvements of 10.55 and 7.57 in L-AP, and 14.31 and 9.92 in L-AP@10, respectively. On the UP-COUNT dataset, STNNet demonstrates considerably lower precision, with declines of 29.29 in L-AP and 46.98 in L-AP@10. It is important to note that the MFA results are directly cited from [87] because the reported performance on DroneCrowd could not be replicated, and the method could not be trained on the UP-COUNT dataset. Although STEERER [78] achieves strong performance on several non-drone-based benchmarks for object localization, it exhibits limited generalization to the task, resulting in L-AP of 38.31 and 40.20 for DroneCrowd and UP-COUNT, respectively. Similarly, although RFLA [84] is explicitly designed for detecting tiny objects, it also reports low scores, resulting in L-APs of 32.05 and 32.41 for DroneCrowd and UP-COUNT, respectively. In contrast, SD-DETR [85], another method within this domain, shows higher robustness. On the DroneCrowd dataset, SD-DETR trails the new method by 2.88 in L-AP and 4.50 in L-AP@10, while on the UP-COUNT dataset, the margins increase to 8.60 and 10.89, respectively.

Qualitative comparisons of the evaluated methods are illustrated in Figure 3.11. These findings demonstrate the superiority of the proposed method over state-of-the-art conventional techniques, underscoring their sensitivity to dataset-specific factors such as image resolution.

Table 3.6: The evaluation results of the people localization task for DroneCrowd and UP-COUNT dataset.

	DroneCrowd			**UP-COUNT**		
Method	**L-AP(↑)**	**L-AP@10(↑)**	**L-AP@15(↑)**	**L-AP(↑)**	**L-AP@10(↑)**	**L-AP@15(↑)**
STNNet [86]	40.45	42.75	50.98	37.20	28.48	50.97
MFA [87]	43.43	47.14	51.58	x	x	x
STEERER [78]	38.31	41.96	46.58	40.20	42.14	50.32
RFLA [84]	32.05	34.41	39.59	32.41	33.27	42.54
SD-DETR [85]	48.12	52.56	57.35	57.89	63.57	75.76
Proposed	**51.00**	**57.06**	**60.45**	**66.49**	**75.46**	**79.57**

To further elucidate the L-AP metric, evaluations are conducted across a range of correctness thresholds (from 1 to 25). As illustrated in Figure 3.12, SD-DETR and the proposed method notably differ from other methods in their performance trends. Lower thresholds correspond to stricter localization accuracy, whereas higher

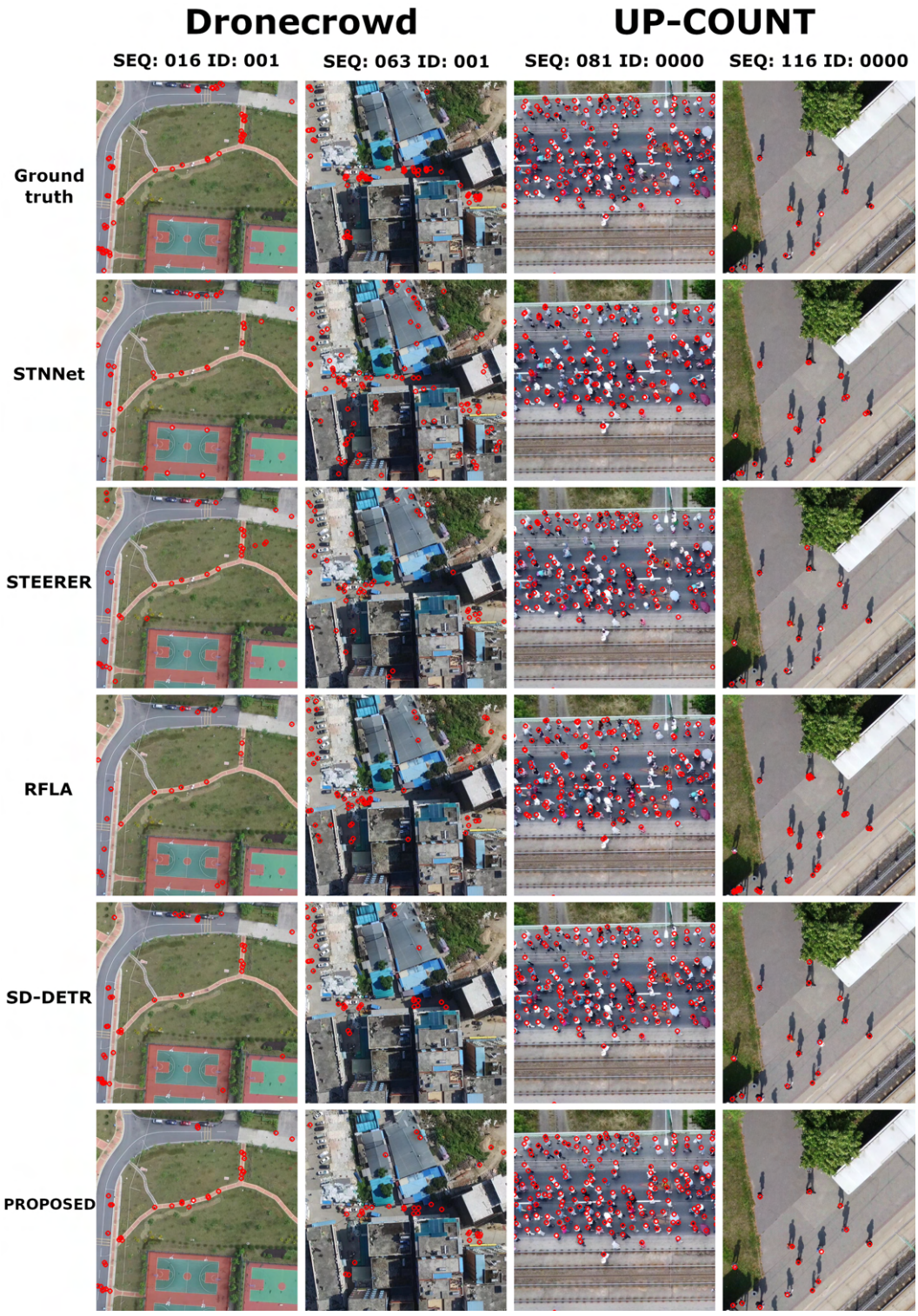


Figure 3.11: Visual comparison of prediction provided by considered methods for DroneCrowd and UP-COUNT datasets [98].

thresholds may allow false positives to be incorrectly accepted as correct matches in scenes where objects are densely packed. On the DroneCrowd dataset, SD-DETR marginally outperforms the new method at the lowest thresholds (below 4); however, at all subsequent threshold levels, the novel method consistently achieves notably

better performance. For the UP-COUNT dataset, the proposed method demonstrates markedly better results for thresholds below 15, while SD-DETR exhibits stronger performance in the higher threshold range. These threshold-dependent analyses offer a more granular understanding of each method's behavior, underscoring the method's advantage in achieving precise localization under stricter evaluation criteria.

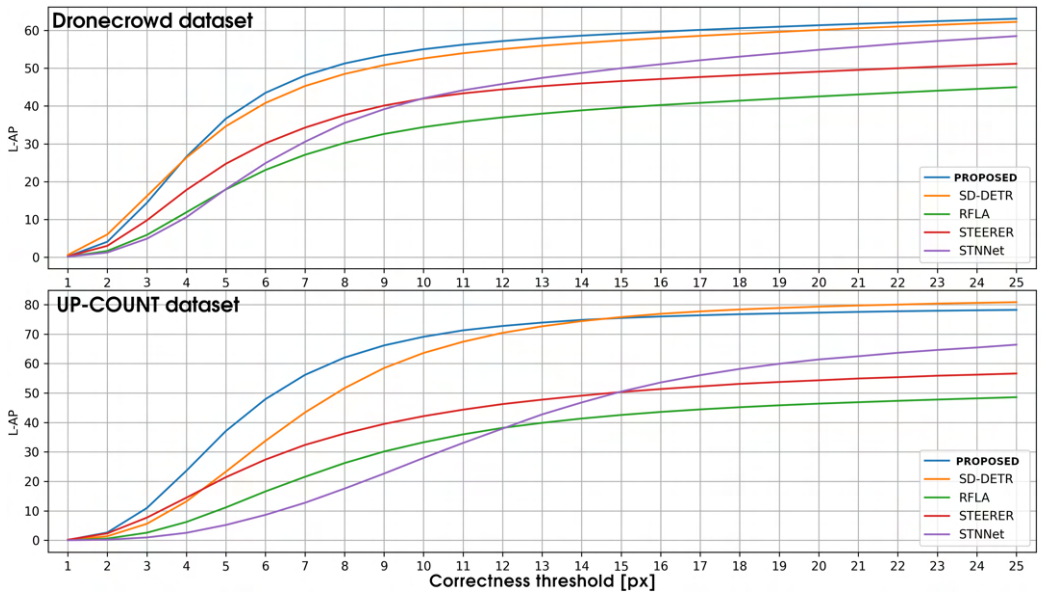


Figure 3.12: The comparison across a range of correctness thresholds, spanning from 1 to 25 pixels, which aligns with the evaluation window used for computing the LmAP metric [98].

3.9 Final conclusions

The series of methodological enhancements has been proposed to address the challenges of point-oriented object localization. These improvements leverage a combination of spatial and temporal features, incorporate drone sensor metadata, and integrate task-specific refinements tailored to the unique characteristics of aerial crowd analysis. The final algorithm, evaluated on the DroneCrowd and UP-COUNT datasets, demonstrates a notable performance gain over previous approaches, achieving state-of-the-art results in terms of localization accuracy under varying scene conditions.

4

Point-oriented object tracking methodology

The previous chapter focused on the point-oriented object localization task, a valuable tool for localizing and counting objects in a single image. However, additional assignment (tracking) methods are required to match the same objects between frames, allowing for their identification and accurate counting in the entire video sequence. Moreover, although object localization methods are considered a central core of tracking methods, their performance remains insufficient for maintaining accurate tracking, thereby decreasing the continuity of trajectories, which is crucial for effective monitoring and counting.

In the following chapter, the tracking methodologies proposed to enhance individual, point-based object tracking are presented and detailed. Figure 4.1 illustrates the complexity of the task, displaying past trajectories of tracked individuals.

Figure 4.1: An example frame demonstrates the task of point-oriented object tracking [99].

4.1 Baseline tracking method

The point-oriented tracking approach presented in this study is based on the SORT algorithm [92]. The introduction of SORT resulted in substantial refinement and broad application in object tracking contexts [136, 91]. It integrates a Kalman filter to model object dynamics and utilizes the Hungarian algorithm for data association, based on a predefined similarity metric. Despite its limitations, such as object motion modeling with a simple linear model, SORT provides a dependable and computationally efficient baseline, serving as the foundational tracking mechanism for the proposed methodology.

The tracking method under consideration adopts a detect-to-track paradigm, utilizing independently predicted object locations in each frame to perform inter-frame association. Employing the method described in the previous chapter, the objects' coordinates are determined, providing a robust foundation for the introduced tracking method. Additionally, its U-type architecture enables the extraction of spatial features at the decoder stage, providing detailed spatial context for accurate visual tracking.

Although SORT has been intensively developed, its performance remains insufficient for maintaining accurate tracking of point-oriented objects, primarily due to the use of default detection-assignment-based methods that are better suited for bounding box tracking. Therefore, in the research [99], a distance-based assignment method is proposed to associate objects with known trajectories. This approach calculates the Euclidean distance between the coordinates of each detected point, which marks the object's location, and the predicted coordinates of the corresponding trajectory object. The distances are measured in pixels. The matching correctness threshold is determined within a circular region centered on the predicted coordinate. A detected point is considered a correct match if it lies within this circle and the distance to the predicted coordinate is less than the circle's radius. By default, the radius is set to ten pixels, which is half of the twenty-pixel object size used in DroneCrowd [86].

Adequately adjusted parameters are essential for the successful performance of tracking algorithms. One key parameter is the minimum number of hits required for a trajectory to be marked as confirmed, representing the minimum trajectory length. Another one is the maximum trajectory age, which specifies how long a missed trajectory is kept in memory before being removed. In the setup, these values are set to 30 and 60 frames, corresponding to approximately 1 and 2 seconds, respectively. These settings make the algorithm robust enough to reduce both false positives and false negative trajectories for long and dynamic sequences.

In addition to the proposed baseline method for point-oriented object tracking, four improvements are proposed and detailed in the following sections. The comprehensive evaluation, along with comparisons to the baseline and the state-of-the-art global optimized method, is presented in Section 4.6. To provide a visual overview of the proposed method, the algorithm flow diagram is presented in Figure 4.2. It illustrates the baseline tracking method combined with proposed improvements, including camera motion compensation (Section 4.2), drone flight altitude adjustment (Section 4.3), enhanced trajectory validation (Section 4.4), and a new method for improving trajectory continuity (Section 4.5).

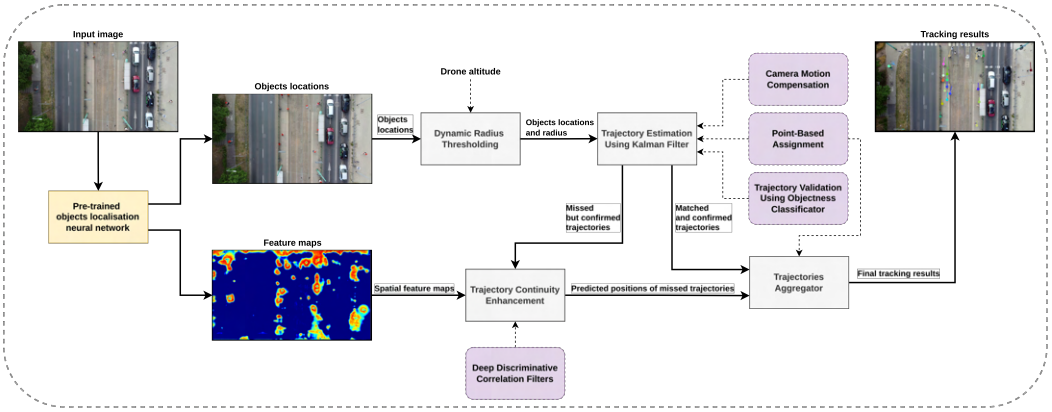


Figure 4.2: Overview of the proposed tracking pipeline [99], integrating point-based object localization, spatial feature maps, and trajectory enhancement methods integration.

4.2 Camera motion compensation

When a drone moves during video recording, algorithms must address noise caused by camera motion. The simultaneous movement of objects and the camera complicates the accurate estimation of trajectories. This movement is not limited to forward motion; it encompasses complex dynamics, including translations, rotations, and often high-frequency vibrations caused by motors or external factors, such as wind. Despite many of these noises being reduced due to the use of vibration-isolated camera gimbals, drone movement can still be considered in object tracking, thereby improving its accuracy. An example motion transformation between two images is presented in Figure 4.3. The example utilizes sparse visual features in both images and a brute-force matcher for assignment, ultimately enabling the calculation of the homography matrix, which contains translation and rotation information in pixels.

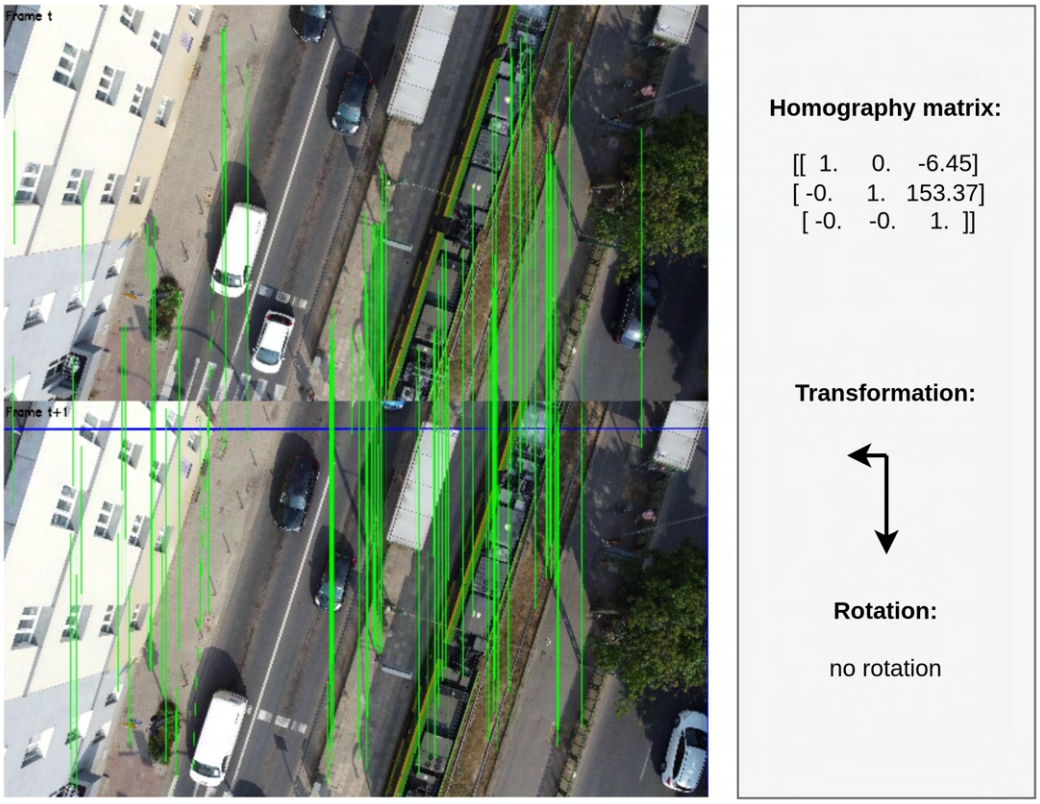


Figure 4.3: An example of in-frame features matched along with calculated transformation and rotation.

4.2.1 Motivation

Using the drone ego-motion calculations in tracking improves its accuracy by reducing movement noise.

4.2.2 Method

Although many drones provide movement estimations from sensors, their usability can be limited due to difficulties in synchronizing this information with camera frames. Therefore, camera compensation methods often focus on estimating movement directly from a frame sequence, extracting visual features, and calculating the translation and rotation between frames. Operating on feature correspondences, this approach remains accurate until most of the image remains motionless, for example, with the ground background.

In the research, the Camera Motion Compensation module from BoT-SORT [96] is adopted. This module estimates inter-frame motion by computing sparse optical flow, enabling the determination of average translational and rotational shifts between consecutive frames. Specifically, sparse corner features are detected in

both frames and matched using the Lucas-Kanade tracking algorithm [137]. These matched keypoints are then utilized to derive an affine transformation matrix that geometrically aligns the frames. The resulting transformation is applied to both the Kalman filter's state vector and its associated noise covariance matrix, allowing object positions to be updated in accordance with the estimated camera motion. As long as the frames contain sufficient visual features for reliable keypoint extraction and matching, this compensation strategy effectively mitigates the impact of camera movement, thereby enhancing the accuracy and stability of object trajectory estimation.

4.3 Utilizing drone sensor altitude for dynamic thresholding

As described in the previous chapter, where flight altitude information was delivered directly into the neural network architecture, this sensor data is often omitted. Knowing the drone's height above ground level can provide valuable context for interpreting the visual scene. It can aid in understanding the relationship between the drone camera, sensed objects, and the terrain below, enabling dynamic adaptation of parameters defined in a tracking algorithm. Figure 4.4 shows eight example fragments of images, considering the same size in pixels and different drone flight altitudes above ground. Although the point-oriented object localization ignores the spatial size of objects, it is possible to consider the estimation of this parameter globally in a frame, calculating dynamic thresholds based on the distance between the drone's camera and sensed objects.

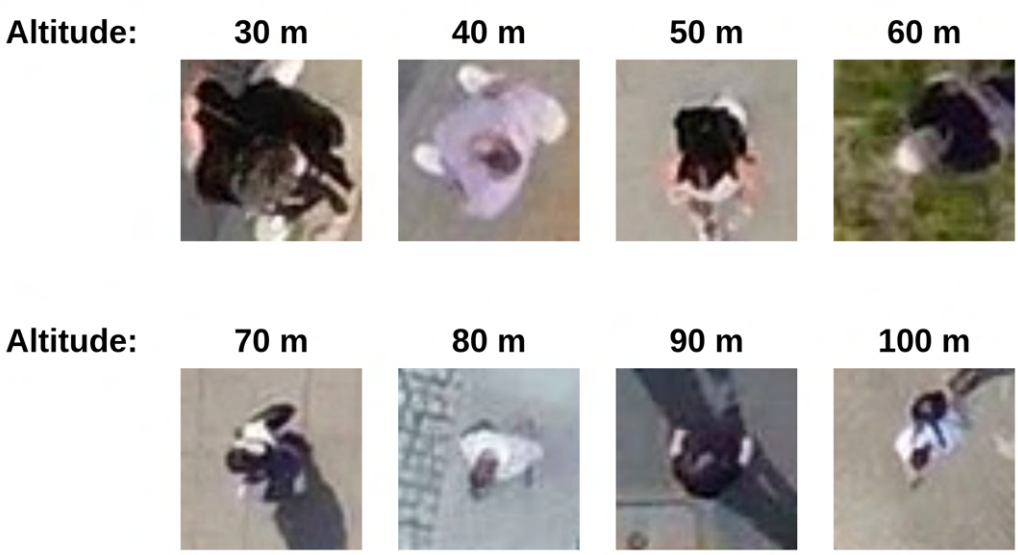


Figure 4.4: The comparison of people's visual appearance depending on drone flight altitude above the ground.

4.3.1 Motivation

Incorporating drone altitude information from sensors into the tracking algorithm enables its parameters to be dynamically adjusted.

4.3.2 Method

Considering flight altitude, in the research [99], this sensor information is proposed to be utilized to calculate a dynamic threshold for assignment correctness, which is defined as:

$$\text{Tr} = max(10, \frac{100}{\text{altitude}} \cdot 10) \tag{4.1}$$

where 10 pixels is half of the default evaluation object size formed in [86].

This dynamic thresholding mechanism influences the assessment of object matching in tracking, changing the circle radius in distance-based trajectory assignment. It enables the tracking algorithm to adapt to fluctuations in object scale and inter-object spacing. At lower altitudes, where objects appear larger in the image due to higher pixel resolution, a greater separation distance is required to ensure accurate assignment. In contrast, at higher altitudes, where objects occupy fewer pixels, a smaller threshold is used to reduce the likelihood of trajectory label switching, thereby maintaining consistency in assignments.

4.4 Additional classification steps to reduce false positives

Tracking algorithms, such as SORT, typically follow the trajectory for a certain duration before it can be confirmed, which means a trajectory is considered valid and included in the evaluation. This requirement reduces the occurrence of very short trajectories and minimizes false positives. It is common to adjust the "trajectory age" in tracking parameters, depending on how sensitive the algorithm should be to new trajectories. In point-oriented object tracking, where sequences are usually long, the value is also appropriately higher. Although this mechanism is designed to reduce false positives, it is insufficient for tasks such as drone crowd counting, as a high number of false positives often characterizes point-oriented localization methods, thereby reducing tracking accuracy and increasing counting errors. Therefore, to further improve trajectory confirmation, an additional classification step is proposed that assesses whether the trajectory's surroundings cover a human instance or not. The need for this step becomes apparent when analyzing challenge cases based on

the localization method. Figure 4.5 contains example regions of interest classified as people by the localization method. The presented true positives represent correct samples with minimal confidence scores, while false positives indicate objects from the background that are wrongly classified as people with higher confidence.

Figure 4.5: Example regions of interest classified as people, including true positives with low confidence and false positives with high confidence.

4.4.1 Motivation

An additional classification step, performed before considering the trajectory as confirmed, reduces false positives, making point-oriented object tracking more accurate.

4.4.2 Method

In the research [99], the trajectory confirmation process is extended by introducing an additional classification step using a neural network. Let the trajectory age

(number of incidences) be defined as (N_{age}), and the minimal trajectory duration as N_{thresh}. If the counter of trajectory age (N_{age}) exceeds a threshold based on the minimal trajectory duration (N_{thresh}), defined as:

$$N_{age} >= (N_{thresh} - 3) \tag{4.2}$$

the object's surroundings are classified until the trajectory is either terminated or confirmed as valid. A trajectory is designated as confirmed once it satisfies two conditions: it reaches the minimum required duration, denoted as N_{thresh}, and the average classification probability, indicating the presence of objects within the localized surroundings, exceeds 80%. This probability threshold was empirically determined through experimental validation, with a focus on reducing false positives.

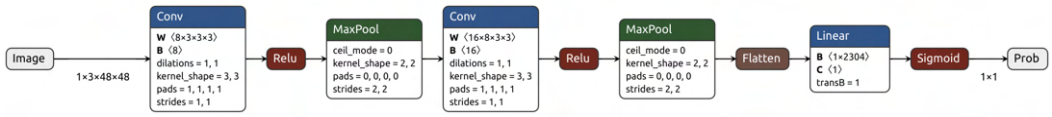


Figure 4.6: Architecture of a simple convolutional network designed to classify if a region of interest contains a person [99].

A lightweight neural network serves as the classification model, with its architecture illustrated in Figure 4.6. The network processes regions of interest extracted from the RGB image, centered on the location of the detected object and bounded by a dynamically defined assignment threshold described in the previous section. These image patches are resized to a fixed resolution of 48×48 pixels prior to inference. The architecture comprises two sequential blocks, each containing a convolutional layer followed by a ReLU activation function and max pooling, facilitating hierarchical feature extraction. The resulting feature maps are subsequently passed through a fully connected layer and transformed via a Sigmoid activation function, producing normalized outputs within the zero-one range. The algorithm is trained based on samples extracted from training images, considering the positions of known objects and randomly selected negative samples.

4.5 Enhancing trajectory continuity

Although point-oriented localization methods provide accurate poses, the resulting estimations can still be significantly noisy. It is particularly pronounced in drone-based crowd tracking, which is commonly affected by both false-negative and false-positive detections. The tiny size and irregular shape of individuals complicate accurate detection, often resulting in missed detections across consecutive frames. These omissions disrupt trajectory continuity and contribute to the increase in counting errors. Conversely, static environmental elements are frequently misclassified

as individuals, producing spurious trajectories that negatively impact performance metrics. Additionally, although videos with high frame rates make tracking easier due to the minimal differences between frames, they can also lead to fragmented trajectories because of the rapid accumulation of frames before the object reappears in the scene. Therefore, methods that enhance trajectory continuity are valuable for maintaining consistent trajectories, especially in long and challenging sequences, such as crowd object tracking.

4.5.1 Motivation

The implementation of the double-behavior tracking method, utilizing classical SORT and Correlation Filters, enhances trajectory continuity, improves tracking abilities, and reduces individual identification switches.

4.5.2 Method

In the research [99], DDCF (Deep Discriminative Correlation Filters) are adapted to maintain trajectories of known objects even in the presence of missed detections by following known and confirmed trajectories. When a confirmed trajectory fails to associate with a detection in the current frame, a correlation filter is initialized using the object's most recent position and visual features. The approach builds on the ECO (Efficient Convolution Operators) algorithm [138], which enables fast object re-localization via correlation computations in the Fourier domain. ECO was designed to track a single object by extracting visual features from the first (Conv-1) and last (Conv-5) convolutional layers of the VGG architecture [139], pretrained for general-purpose use. However, this approach is not practical for multi-objective, densely packed crowds, especially sensed from a drone perspective. Therefore, using the U-style architecture of the point-oriented localization method and its rich spatial feature representation on the decoder stage, a solution aligned with the zero-waste machine learning paradigm [140] is proposed. Rather than employing a separate neural network for feature extraction, the necessary information is derived directly from the people localization method outlined in the preceding chapter. To optimize spatial resolution and enrich feature representation, features are extracted from the penultimate layer of the model's head, yielding a spatial feature map of size $544 \times 940 \times 16$. This global feature map encompasses the entire frame and is subsequently refined to individual objects by isolating regions around their respective locations. These features inherently capture both spatial and semantic (in the sense of describing the type of object) context, which is particularly advantageous in densely populated environments. An example visualization of object-specific features is provided in Figure 4.7. By reusing features already computed during object

localization, the approach adheres to the principles of zero-waste machine learning, minimizing computational redundancy and enabling the tracking of multiple objects simultaneously.

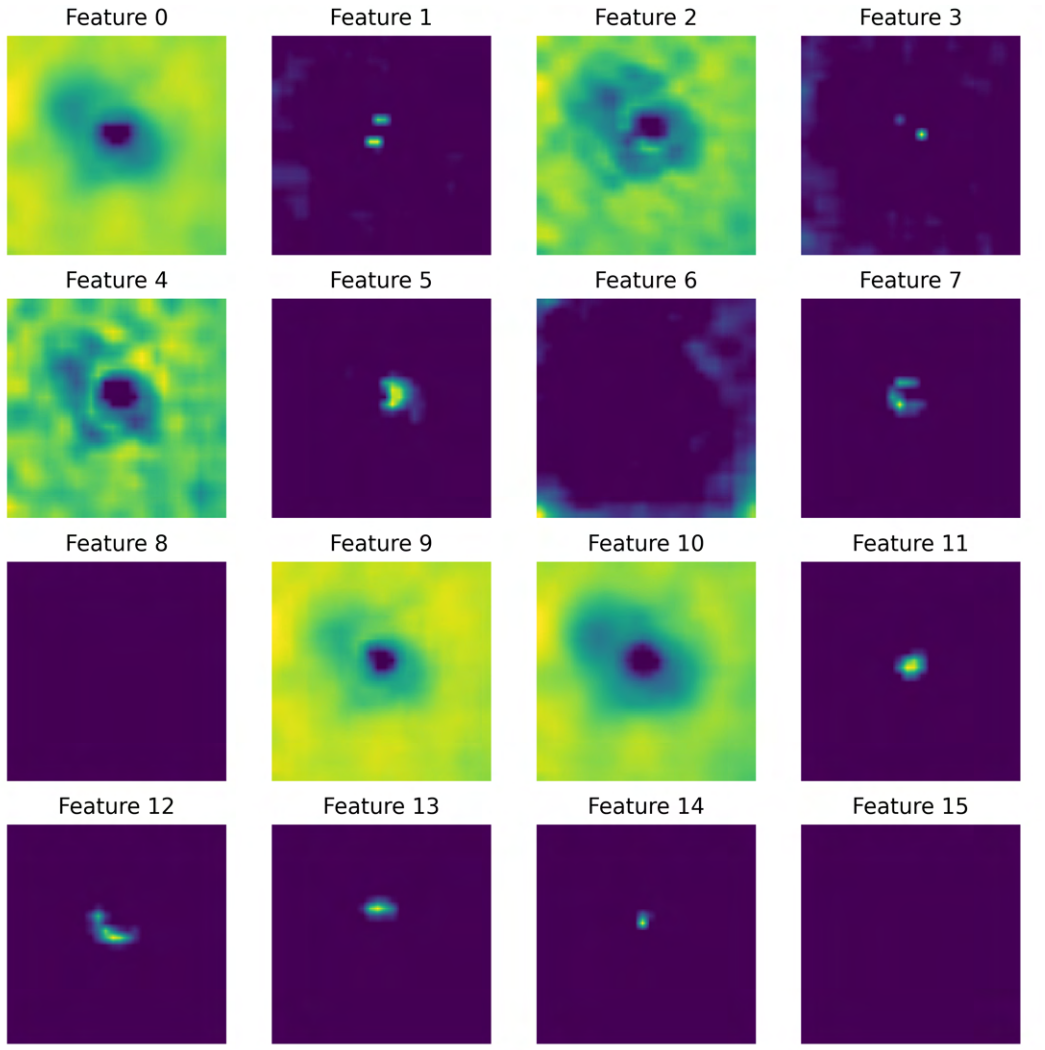


Figure 4.7: Sixteen deep features extracted in the neighborhood of a person [99]. Values are normalized to a zero-one range for visualization purposes.

When confirmed trajectories fail to match with detections in a given frame due to occlusions or missed detections, their positions are estimated using Deep Discriminative Correlation Filters. By leveraging features extracted from the localization network, this mechanism updates the trajectories and integrates them into Kalman filter predictions, thereby enhancing both their continuity and temporal smoothness. The filters use the position of the object as computed in the previous frame and its visual features' maps to initialize tracking. Then, an initial discriminative correlation filter is learned based on these features. Next, in a new frame, a search region is extracted around the predicted target location, feature maps are extracted from this search region, and the learned filter is applied to it. This process generates a response map, where the peak indicates the most likely new target location. Finally,

the filter is periodically updated using the selected training samples to adapt to changes in object appearance. A visual representation of the approach is provided in Figure 4.8, which shows selected tracking frames from sequences (frames 0 to 200) and demonstrates the method's efficacy in tracking individual objects based on spatial features across both the UP-COUNT-TRACK and DroneCrowd datasets.

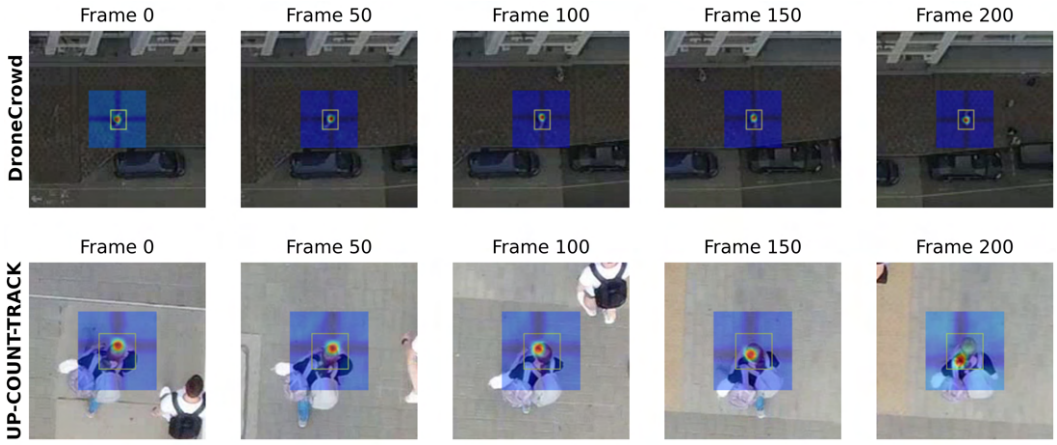


Figure 4.8: Usage of Deep Discriminative Correlation Filters in drone-based people tracking for sample images [99]. The heatmap indicates the locations with the highest correlation within the analyzed image area.

4.6 The complete tracking pipeline validation

The evaluation process covers tracking and trajectory counting metrics on the UP-COUNT-TRACK and DroneCrowd datasets. It compares a baseline tracking method with incremental additions of Camera Motion Compensation (CMC), dynamic thresholding considering flight altitude (ALT), additional classification in the trajectory confirmation step (CLS), and the enhanced Deep Discriminative Correlation Filters (DDCF). Additionally, the method is evaluated against the current state-of-the-art, the globally optimal greedy (GOG) approach, demonstrating the robustness of the online tracking method, even when compared to an offline solution that considers the entire video recording and operates in a dual, forward-backward manner. The results of the evaluation are presented in Table 4.1 and Table 4.2, assessing the impact of proposed improvements on the UP-COUNT-TRACK and DroneCrowd datasets, respectively.

Focusing on UP-COUNT-TRACK (Table 4.1), the baseline method achieves a HOTA of 0.63 and a T-AP of 37.04, with 3305 ID switches and a Tr-nMAE of 0.37 ± 0.34. Adding CMC provides marginal benefits, slightly increasing T-AP (37.36) and reducing ID switches (3180). Incorporating ALT yields further improvements, notably boosting T-AP to 38.68, T-AP@10 to 40.90, and slightly increasing HOTA to 0.64, while also reducing counting errors (Tr-nMAE 0.33 ± 0.29). The addition of CLS

Table 4.1: Tracking results for the UP-COUNT-TRACK dataset, considering proposed improvements. *Globally-optimal greedy (GOG) algorithm is an offline method.

Method	HOTA(↑)	T-AP(↑)	T-AP@10(↑)	ID-SW(↓)	Tr-MAE(↓)	Tr-nMAE(↓)
Baseline	0.63	37.04	39.34	3305	49.13 ± 117.22	0.37 ± 0.34
Baseline + CMC	0.63	37.36	39.59	3180	48.84 ± 114.43	0.38 ± 0.32
Baseline + CMC + ALT	0.64	38.68	40.90	3126	44.97 ± 103.25	0.33 ± 0.29
Baseline + CMC + ALT + CLS	0.63	38.72	40.95	2943	41.81 ± 96.76	0.30 ± 0.26
Baseline + CMC + ALT + CLS + DDCF	**0.63**	**44.35**	**45.88**	**287**	**20.45 ± 44.81**	**0.15 ± 0.11**
GOG*	0.42	36.21	37.63	1868	64.19 ± 110.24	0.57 ± 0.36

primarily contributes to reducing ID switches further (2943) and improving counting accuracy (Tr-nMAE 0.30 ± 0.26), while maintaining HOTA and T-AP levels. The most significant gains are observed with the final addition of DDCF. The method achieves the best results across most metrics. While HOTA remains at 0.63, T-AP increases substantially to 44.35, and T-AP@10 reaches 45.88. Most notably, ID switches are drastically reduced to 287, and counting performance improves significantly, with the absolute trajectory counting error, Tr-MAE, dropping to 20.45 ± 44.81, and the relative counting error, Tr-nMAE, to 0.15 ± 0.11. These results demonstrate the cumulative effectiveness of the proposed components, with DDCF having the most substantial impact on association quality and counting accuracy. For reference, the offline GOG method shows comparatively lower performance on this dataset across these metrics.

In contrast, on DroneCrowd (Table 4.2), both the baseline and proposed simple improvements (CMC, ALT, CLS) perform worse than the Globally-optimal greedy algorithm. While baseline achieves T-AP of 46.27, ID-SW of 6290, and Tr-nMAE of 0.32 ± 0.15, the GOG results in 50.47, 1326, and 0.24 ± 0.17. Although CMC, ALT, and CLS improvements improve T-AP, Tr-MAE, and Tr-nMAE metrics slightly, they tend to generate more ID switches. However, the use of the trajectory continuity enhancement with correlation filters (DDCF) outperforms GOG on all metrics, reducing ID-SW to 388, improving tracking metrics to 0.54 (HOTA) and 54.59 (T-AP), and minimizing trajectory counting errors to 37.60 ± 25.78 (Tr-MAE) and 0.23 ± 0.16 (Tr-nMAE).

Table 4.2: Tracking results for the DroneCrowd dataset, considering proposed improvements. *Globally-optimal greedy (GOG) algorithm is an offline method.

Method	HOTA(↑)	T-AP(↑)	T-AP@10(↑)	ID-SW(↓)	Tr-MAE(↓)	Tr-nMAE(↓)
Baseline	0.53	46.27	49.99	6290	57.47 ± 37.13	0.32 ± 0.15
Baseline + CMC	0.53	46.87	50.37	6231	55.83 ± 35.53	0.31 ± 0.14
Baseline + CMC + ALT	0.52	47.44	51.13	6771	52.90 ± 31.37	0.30 ± 0.13
Baseline + CMC + ALT + CLS	0.52	47.32	50.88	6645	52.10 ± 30.36	0.30 ± 0.14
Baseline + CMC + ALT + CLS + DDCF	**0.54**	**54.59**	**57.04**	**388**	$\mathbf{37.60 \pm 25.78}$	$\mathbf{0.23 \pm 0.16}$
GOG*	0.51	50.47	53.21	1326	38.10 ± 31.14	0.24 ± 0.17

Conclusions: While the proposed method achieves trajectory counting metrics comparable to GOG on the DroneCrowd dataset, it consistently outperforms GOG across all other evaluation metrics, most notably in reducing ID switches. On the UP-COUNT-TRACK dataset, the enhanced online tracking approach significantly surpasses the offline GOG algorithm. This performance gap likely arises from the greater visual variability and extended sequence lengths in UP-COUNT-TRACK, which introduce more severe tracking challenges in the absence of continuity enhancements. Overall, these results consistently demonstrate the robustness and effectiveness of the proposed online tracking method across diverse datasets and challenging drone-based crowd scenarios.

To illustrate the tracking performance across both datasets, Figure 4.9 shows examples from diverse environments and flight altitudes. Ground-truth trajectories are marked in green, while the predicted trajectories generated by the model are shown in red. This visualization highlights the model's ability to maintain accurate and continuous trajectory estimations.

4.7 Statistical results analysis

To assess the algorithm's robustness across different conditions, a statistical analysis is performed on the UP-COUNT-TRACK dataset (Figure 4.10) that contains additional flight metadata. Specifically, Pearson correlation coefficients are computed between trajectory counting errors (Tr-nMAE) and three sequence-level attributes: sequence duration, the number of distinct trajectories, and mean flight altitude.

Figure 4.9: Visual comparison of ground-truth trajectories (green) and model-predicted trajectories (red) across both datasets [99].

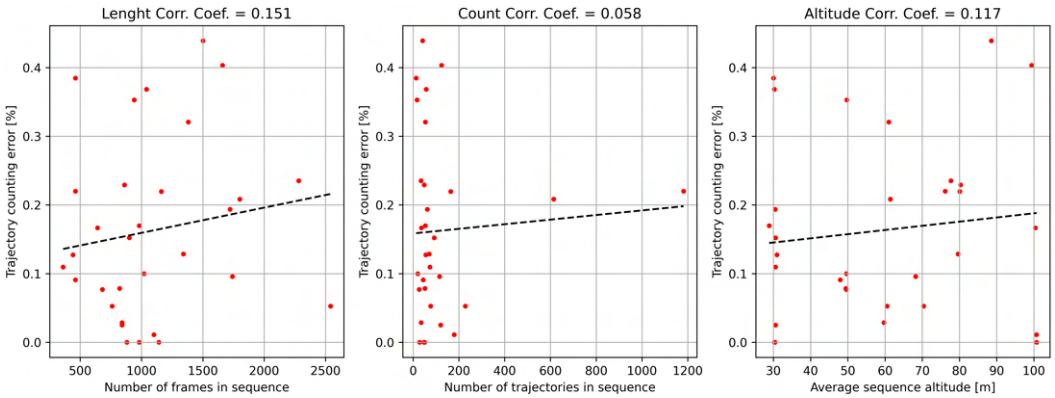


Figure 4.10: Statistical analysis of counting trajectory error and three sequence characteristics: sequence length, number of unique trajectories in a sequence, and average flight altitude during recording.

The resulting correlation values – 0.151, 0.058, and 0.117, respectively – indicate a minimal linear association. These findings imply that the algorithm's performance remains stable, regardless of variations in tracking duration, object density, or average object scale.

4.8 Final conclusions

To address the challenges associated with tracking tiny, densely packed objects in low-altitude aerial imagery, a point-oriented tracking pipeline is proposed. Its architecture employs three improved modules that aim to improve the general detection-to-tracking approach of point-based objects. Additionally, the trajectory continuity enhancement module is proposed and tailored to improve the spatial-temporal association, significantly improving tracking robustness. The evaluation on the DroneCrowd and UP-COUNT-TRACK datasets demonstrates a notable performance gain over the previously used GOG approach, achieving state-of-the-art results in terms of tracking and counting accuracy.

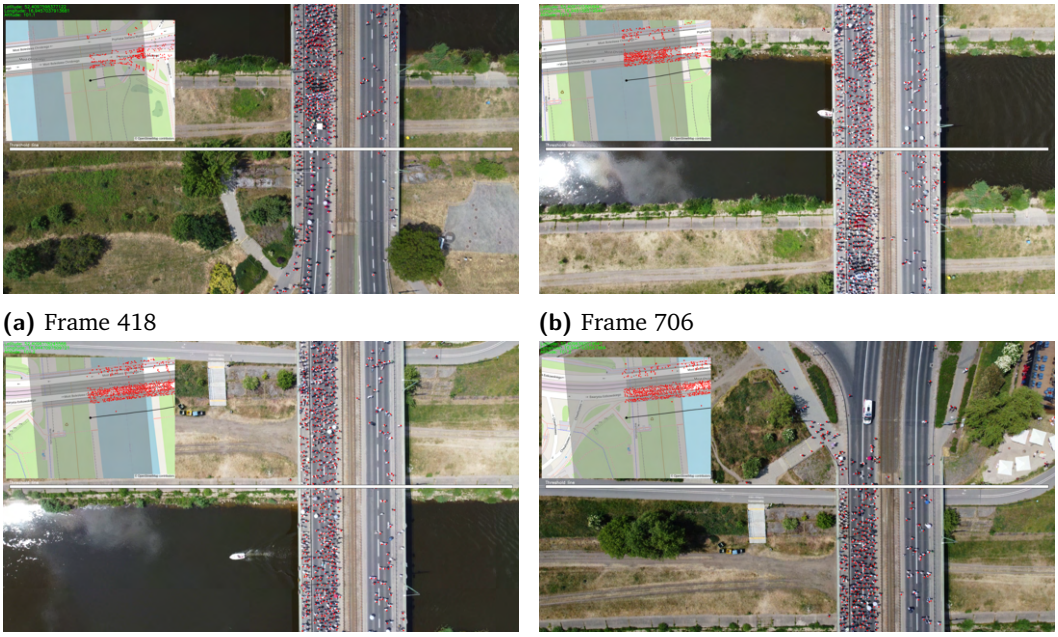
5 Real-world usage scenarios

Although precise point-oriented object localization and tracking algorithms have broad potential applications, they are particularly valuable in domains such as crowd analysis and management due to their ability to identify individuals in complex environments. This capability is particularly significant in high-resolution video contexts, such as drone recordings, where identifying and tracking individuals can provide critical insights. Below, several promising applications in this domain are explored in more detail.

Accurate object identification within an image can be successfully integrated with data from the drone's onboard sensors to estimate global coordinates, such as GPS (Global Positioning System) positions defined by latitude and longitude. By incorporating the camera's intrinsic parameters, for example, focal length, sensor size, and principal point, typically obtained through a calibration process, it is possible to transform image coordinates into a local metric coordinate system, and subsequently into a global reference frame [141]. The precision of this transformation depends not only on the quality of the camera calibration but also on the accuracy of the sensor data – particularly the altitude above ground level (commonly measured relative to the takeoff elevation), the drone's orientation, and the position of the camera (or gimbal) relative to the drone's frame of reference. When these conditions are satisfied, triangulation techniques can be employed to estimate the spatial positions of detected individuals, thereby enabling 3D localization from 2D image data.

Due to legal and safety regulations, flying over a crowd is considered a high-risk operation that requires special precautions and, in many cases, dedicated flight authorization, particularly during mass events. Therefore, the proposed applications and use cases are intended to demonstrate potential capabilities and are primarily aimed at public services such as law enforcement, fire departments, and municipal authorities. Additionally, the use of a constructed platform, which includes a hardware-optimized embedded device for neural network deployment, usually involves a greater weight of the drone, resulting in a higher hazard compared to a consumer-grade aerial platform. This is particularly important during large-scale events and gatherings of people.

5.1 Crowd counting in dynamic environments



(a) Frame 418 **(b)** Frame 706

(c) Frame 993 **(d)** Frame 1280

Figure 5.1: Example frames from a video recording of a march. Each frame contains people detections marked in red, the white threshold line indicating the consideration cut-off, and a mini-map displaying the global coordinates of the detected people [98].

Large-scale outdoor events, such as music festivals, marches, or public demonstrations, require dedicated tools to accurately monitor their size, discover critical crowd issues, and ensure situational awareness for effective crowd control and safety management. Fluctuating crowd densities, constant pedestrian flow, unpredictable movement patterns, and large crowd spans make the real-world monitoring challenging, especially when using traditional methods.

The proposed method enables precise headcounts by detecting and tracking individuals not only within a scene, but in entire recordings captured by a drone. The algorithm localizes people in each video frame and integrates this information with data provided by drone sensors. This enables accurate crowd size estimation and behavior analysis, providing valuable feedback for public services. Additionally, thanks to combining individuals' positions in recording with the drone's sensor information, including its global coordinates, orientation, and altitude above ground, it is possible to map individuals' locations accurately onto a global reference frame. An example video processing workflow is illustrated in Figure 5.1. It shows four video frames from different timestamps and positions. In each frame, detected individuals are marked in red, while a white line indicates the boundary for objects considered in the mapping process. Each frame also features a mini-map displaying the drone's position and trajectory (black line), camera field of view (gray area), and

the global coordinates of detected individuals (red markings). When the algorithm finishes processing a sequence, the final crowd heatmap is also generated as an output. An example heatmap is presented in Figure 5.2.

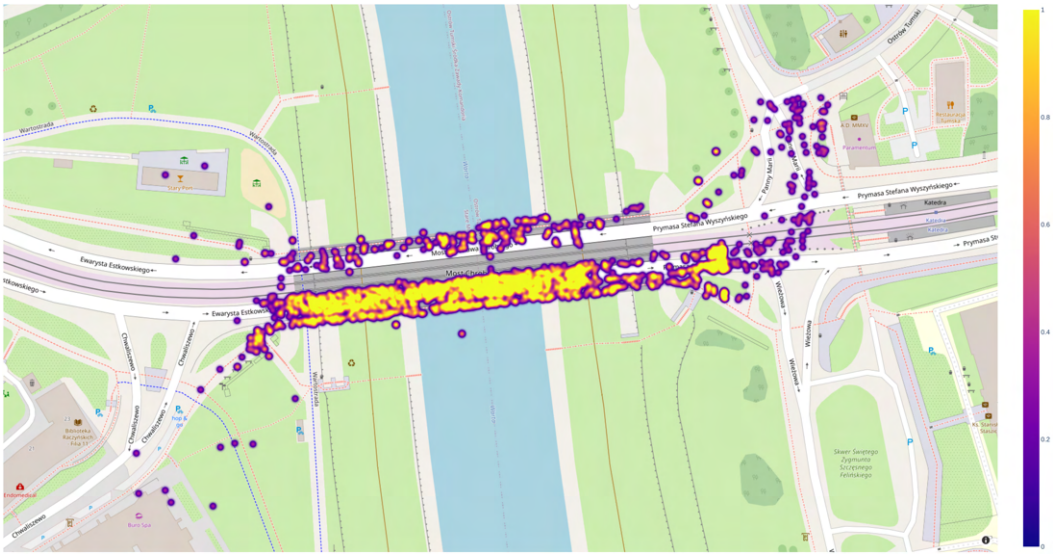


Figure 5.2: An example crowd density heatmap [98].

The primary objective of this application is individual identification and counting, enabling the estimation of crowd size for statistical analysis and the identification of high-density areas. Venue operators can optimize staffing levels based on occupancy data, proactively identify potentially dangerous overcrowding situations, and facilitate efficient emergency response and evacuation procedures.

5.2 Crowd monitoring through individual tracking

The advanced tracking of individuals in drone-recorded videos enables the analysis of individual movement, including monitoring crowd dynamics, capturing motion patterns, and analyzing behavior. An example frame with top-view trajectory analysis is presented in Figure 5.3. The well-prepared trajectory analysis, including object positions, historical movements, and speeds, facilitates further analysis by scientists in different domains.

The proposed algorithm's robustness in maintaining identities during complex interactions generates consistent trajectories, serving as a valuable baseline for individual tracking. By combining the algorithm outcomes with drone sensors data, it is possible to produce a map with the global coordinates of objects. Example results of individuals tracking and mapping in video recordings are presented in Figure 5.4. Every example displays a side-by-side comparison of the processed video

Figure 5.3: An example frame with top-view trajectory analysis.

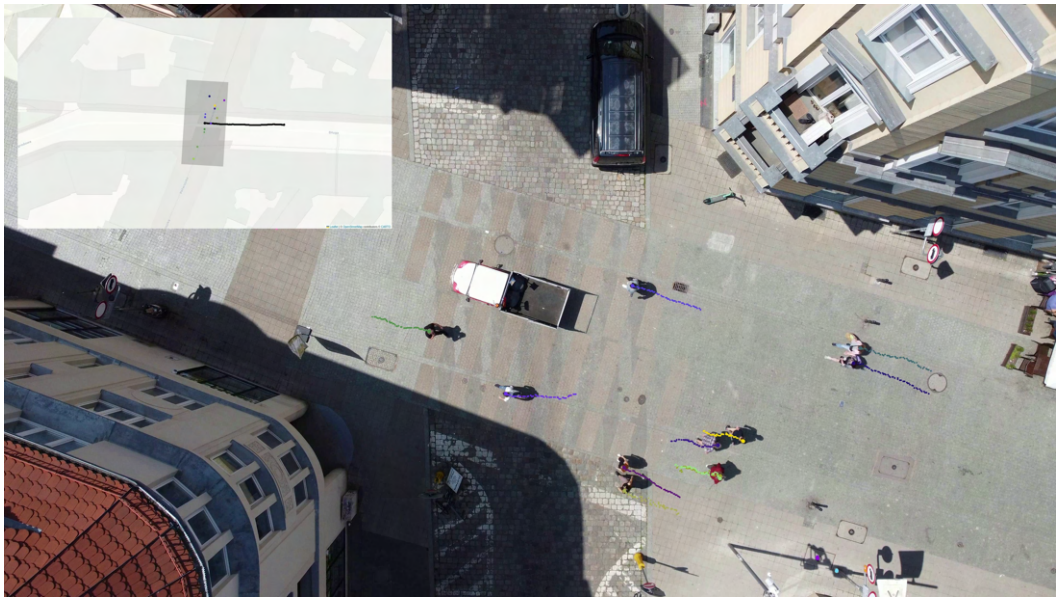
and the corresponding map, both of which highlight detected objects along with their movement histories.

The primary objective of this application is to track individuals, enabling the determination and capture of their trajectories. Such information facilitates data-driven decision-making in urban planning, including the optimization of public spaces to enhance pedestrian flow, reduce congestion, and identify areas that require infrastructure enhancements.

5.3 Conclusions

The real-world applications discussed in this chapter highlight the practical significance and potential to drive advancements in accurate object localization and tracking techniques, particularly in the context of aerial video analysis. Two primary use cases – crowd counting and crowd monitoring – demonstrate the integration of visual data with complementary sensor measurements. In both scenarios, the capability to georeference detected individuals substantially improves situational awareness, supports public safety efforts, and facilitates effective mass event management.

90:7467204667



(a) Example frame from video recorded at low flight altitude.

(b) Example frame from video recorded at medium flight altitude.

Figure 5.4: Visualization of people tracking, along with mapping their global coordinates [99].

6 Discussion

This dissertation addresses the challenges of point-oriented object localization and tracking in low-altitude aerial recordings. This environment poses particular challenges for computer vision tasks due to the continuous movement of tiny objects, simultaneous camera motion, noise, and variations in altitude and perspective. However, advancements in this field can have tangible impacts on crowd management and public safety applications, thereby increasing the social and practical relevance of the research. The primary objective of this thesis was to investigate how spatial and temporal features extracted from high-resolution drone video recordings can be leveraged to close existing research gaps. The proposed algorithms were evaluated on both the publicly available DroneCrowd dataset and a newly introduced dataset, with their utility demonstrated in two real-world use cases. Based on these contributions, the main goals of this dissertation are considered to have been successfully achieved.

6.1 Conclusions

The key contributions of this work, aligned with the initial research hypotheses, are summarized as follows:

- This work introduces several improvements in point-oriented object localization in drone imagery. Key contributions include a motion-enhanced image analysis method that captures object motion across sequential frames and fuses these temporal features with visual features, leading to improved counting accuracy. A specialized neural network module is proposed to process high-resolution images by leveraging spatial features, thereby mitigating information loss that occurs during traditional image downscaling. The thesis also investigates the impact of input resolution on neural network performance in crowd counting tasks, explores the use of task-specific synthetic data during training instead of general-purpose pretrained models, and incorporates drone altitude directly into the network. Additionally, a novel loss function tailored for point-oriented object localization is proposed, enhancing object prominence in the output masks. These contributions are comprehensively

evaluated and achieve state-of-the-art performance on the DroneCrowd and UP-COUNT datasets.

- A novel point-oriented tracking method is presented, adapting the linear tracking approach from the SORT algorithm by modifying the assignment process for point-defined objects. Building on this method, three mechanisms are introduced to improve the algorithm's stability and reduce tracking errors. First, a camera motion compensation module is implemented to correct object positions in response to drone movement. Second, a lightweight neural network classifier filters out false-positive trajectories, increasing robustness. Third, algorithm parameters are dynamically adjusted according to drone altitude, accounting for changes in ground sampling distance. Furthermore, the tracking method has been improved by introducing a novel trajectory continuity and consistency-enhancing method. It effectively extracts spatial features from the localization method and uses these deep visual features to track missing objects with deep discriminative correlation filters, strongly reducing identity switches and increasing trajectory counting accuracy. Validation on both the DroneCrowd and UP-COUNT-TRACK datasets was performed, achieving outstanding results in the point-oriented object tracking task.

- Two new datasets are released to fulfill the requirements of modern drone-based video analysis. While existing datasets have advanced research in this area, the new datasets offer extended sequences in dynamic urban environments with moving drones, complemented by metadata such as global positioning and flight altitude. The UP-COUNT dataset supports localization research, comprising 202 unique sequences and 10000 high-resolution frames with 352487 manually annotated object instances. The UP-COUNT-TRACK dataset provides tracking annotations for 3807 unique trajectories and 1360547 object instances. These publicly available benchmarks promote further research and support a variety of aerial robotics applications.

- To demonstrate the real-world applicability of the proposed localization and tracking pipeline, two proof-of-concept scenarios are presented. The first focuses on dynamic crowd counting, enabling accurate population estimation and detection of high-density regions. The second demonstrates individual-based crowd monitoring, providing insights into pedestrian flow and behavioral patterns through continuous tracking and the integration of drone sensors.

6.2 Limitations

Every method and approach has some limitations, and the ones proposed in this dissertation are no exception. Thus, in this section, the most prominent limitations of the introduced approaches are listed and discussed:

- Although the proposed point-oriented localization methods achieve high accuracy, they still suffer from false positives and negatives. These issues are primarily due to the dynamic drone perspective, the small size of objects, and the diverse appearances of objects. Even human annotators struggle with consistently identifying individuals in single frames. These challenges necessitate ongoing improvements in drone-view object detection.

- Although the tracking algorithm follows an online processing paradigm, the current pipeline is not optimized for direct deployment on drones. Processing high-resolution video streams in real time using multi-stage pipelines exceeds the capabilities of lightweight, power-efficient Edge AI hardware typically available on aerial platforms.

- The tracking method operates on a frame-by-frame basis and is limited in handling long-term occlusions or re-entry of objects into the scene. In such cases, objects are assigned new identities, resulting in identity switches and inaccuracies in the counting process.

- Although the datasets span various conditions, their generalizability may still be limited. Both DroneCrowd and UP-COUNT focus on urban settings. DroneCrowd was collected in China and UP-COUNT in Poland, which may potentially reduce model effectiveness in rural areas or other cultural and geographical contexts.

6.3 Future work

To address these limitations and further advance the field, several promising directions for future research are proposed:

- Future work should explore domain adaptation and transfer learning techniques to enhance the generalizability of localization and tracking models across diverse environments, including varying weather conditions (e.g., snow, fog, nighttime) and different aerial platforms (e.g., satellite or airplane-based imaging).

- The research should also be directed toward self-supervised or weakly supervised learning methods, which address the limitations of dataset availability, particularly in scaling to underrepresented regions or uncommon environmental conditions.

- Optimizing the tracking pipeline for deployment on embedded systems would enable real-time processing directly on drones. It should specifically focus on the trajectory consistency enhancement module, which, despite notable metric improvements, has high operational demands. Achieving this would open new possibilities for real-time crowd monitoring and emergency response.

- The research can also be directed toward the possibility of re-identifying small and densely packed objects. It could improve general tracking and counting accuracy, minimizing the need for individual identification switches. This challenging task can be especially valuable in cases where objects are temporarily occluded or exit the camera's view. Furthermore, integrating re-identification into a multi-drone collaborative configuration could support wide-area surveillance in large-scale events or emergencies.

6.4 Ethics statement

The UP-COUNT and UP-COUNT-TRACK datasets maintain data privacy by design. The aerial perspective captures individuals at a resolution where identifiable features are minimized, reducing the risk of personal identification and mitigating privacy concerns. The dataset does not include biometric data, facial features, or personally identifiable information, ensuring compliance with ethical standards for data collection and usage.

This research is intended for civilian applications, particularly in areas such as crowd monitoring for public safety, urban planning, and resource allocation in large gatherings. However, similar methodologies could be adapted for surveillance or target identification. This highlights the importance of responsible development, transparent governance, and the ethical deployment of computer vision technologies. Researchers and practitioners who utilize proposed algorithms and shared datasets must adhere to ethical guidelines and legal frameworks to ensure that the technology serves beneficial and non-invasive purposes.

6.5 Data availability statement

The data supporting this research are publicly available at:

- UP-COUNT: `https://doi.org/10.5281/zenodo.12683104`
- UP-COUNT-TRACK: `https://doi.org/10.5281/zenodo.13829572`

6.6 Declaration of code availability

Addressing the reproducibility of the results presented in the thesis, the code repositories are provided:

- `https://github.com/up-count/uav-dot/`
- `https://github.com/up-count/uav-dot-track/`
- `https://github.com/PUTvision/DronePeopleCounting/`

98:9784173258

List of Figures

List of Tables

Bibliography

[1] A. Otto, N. Agatz, J. Campbell, B. Golden, and E. Pesch. „Optimization approaches for civil applications of unmanned aerial vehicles (UAVs) or aerial drones: A survey“. In: *Networks* 72.4 (2018), pp. 411–458 (cit. on p. 5).

[2] H. Zhang, C. Wang, S. T. Turvey, et al. „Thermal infrared imaging from drones can detect individuals and nocturnal behavior of the world's rarest primate“. In: *Global Ecology and Conservation* 23 (2020), e01101 (cit. on p. 5).

[3] A. B. Giles, R. E. Correa, I. R. Santos, and B. Kelaher. „Using multispectral drones to predict water quality in a subtropical estuary“. In: *Environmental technology* 45.7 (2024), pp. 1300–1312 (cit. on p. 5).

[4] D. Fawcett, C. Panigada, G. Tagliabue, et al. „Multi-scale evaluation of drone-based multispectral surface reflectance and vegetation indices in operational conditions“. In: *Remote sensing* 12.3 (2020), p. 514 (cit. on p. 5).

[5] M. Kirsch, S. Lorenz, R. Zimmermann, et al. „Integration of terrestrial and drone-borne hyperspectral and photogrammetric sensing methods for exploration mapping and mining monitoring“. In: *Remote Sensing* 10.9 (2018), p. 1366 (cit. on p. 5).

[6] P. Royo, À. Asenjo, J. Trujillo, E. Çetin, and C. Barrado. „Enhancing drones for law enforcement and capacity monitoring at open large events“. In: *Drones* 6.11 (2022), p. 359 (cit. on p. 5).

[7] O. Alon, S. Rabinovich, C. Fyodorov, and J. R. Cauchard. „Drones in firefighting: A user-centered design perspective“. In: *Proceedings of the 23rd international conference on mobile human-computer interaction*. 2021, pp. 1–11 (cit. on p. 5).

[8] M. I. Jordan and T. M. Mitchell. „Machine learning: Trends, perspectives, and prospects“. In: *Science* 349.6245 (2015), pp. 255–260 (cit. on p. 5).

[9] A. Ramachandran and A. K. Sangaiah. „A review on object detection in unmanned aerial vehicle surveillance“. In: *International Journal of Cognitive Computing in Engineering* 2 (2021), pp. 215–228 (cit. on pp. 6, 21).

[10] Y. LeCun, Y. Bengio, and G. Hinton. „Deep learning“. In: *nature* 521.7553 (2015), pp. 436–444 (cit. on pp. 6, 17).

[11] A. Gohari, A. B. Ahmad, R. B. A. Rahim, et al. „Involvement of surveillance drones in smart cities: A systematic review“. In: *IEEE Access* 10 (2022), pp. 56611–56628 (cit. on p. 6).

[12]P. Mittal, R. Singh, and A. Sharma. „Deep learning-based object detection in low-altitude UAV datasets: A survey“. In: *Image and Vision computing* 104 (2020), p. 104046 (cit. on p. 6).

[13]A. Kos, D. Belter, and K. Majek. „Deep Learning for Small and Tiny Object Detection: A Survey“. In: *Pomiary Automatyka Robotyka* 27.3 (2023) (cit. on pp. 7, 20).

[14]B. Li, H. Huang, A. Zhang, P. Liu, and C. Liu. „Approaches on crowd counting and density estimation: a review“. In: *Pattern Analysis and Applications* 24 (2021), pp. 853–874 (cit. on pp. 7, 33).

[15]A. B. Chan and N. Vasconcelos. „Counting people with low-level features and Bayesian regression“. In: *IEEE Transactions on image processing* 21.4 (2011), pp. 2160–2177 (cit. on p. 7).

[16]K. H. Cheong, S. Poeschmann, J. W. Lai, et al. „Practical automated video analytics for crowd monitoring and counting“. In: *IEEE access* 7 (2019), pp. 183252–183261 (cit. on p. 7).

[17]M. Cruz, J. J. Keh, R. Deticio, et al. „A people counting system for use in CCTV cameras in retail“. In: *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM)*. IEEE. 2020, pp. 1–6 (cit. on p. 7).

[18]Y. Pang, Z. Ni, and X. Zhong. „Federated learning for crowd counting in smart surveillance systems“. In: *IEEE Internet of Things Journal* 11.3 (2023), pp. 5200–5209 (cit. on p. 7).

[19]O. Elharrouss, N. Almaadeed, K. Abualsaud, et al. „Drone-SCNet: Scaled cascade network for crowd counting on drone images“. In: *IEEE Transactions on Aerospace and Electronic Systems* 57.6 (2021), pp. 3988–4001 (cit. on p. 7).

[20]I. Bakour, H. N. Bouchali, S. Allali, and H. Lacheheb. „Soft-CSRNet: Real-time dilated convolutional neural networks for crowd counting with drones“. In: *2020 2nd International Workshop on Human-Centric Smart Environments for Health and Well-being (IHSH)*. IEEE. 2021, pp. 28–33 (cit. on p. 7).

[21]C. J. Hong and M. H. Mazlan. „Development of Automated People Counting System using Object Detection and Tracking“. In: *Inter. Journal of Online & Biomedical Engineering* 19.6 (2023) (cit. on p. 8).

[22]G. Tang, J. Ni, Y. Zhao, Y. Gu, and W. Cao. „A Survey of Object Detection for UAVs Based on Deep Learning“. In: *Remote Sensing* 16.1 (2023), p. 149 (cit. on pp. 8, 18).

[23]A. Bakhtiarnia, Q. Zhang, and A. Iosifidis. „Efficient high-resolution deep learning: A survey“. In: *ACM Computing Surveys* (2022) (cit. on p. 9).

[24]M. A. Husman, W. Albattah, Z. Z. Abidin, et al. „Unmanned aerial vehicles for crowd monitoring and analysis“. In: *Electronics* 10.23 (2021), p. 2974 (cit. on p. 9).

[25]M. Ghamari, P. Rangel, M. Mehrubeoglu, G. S. Tewolde, and R. S. Sherratt. „Unmanned aerial vehicle communications for civil applications: A review“. In: *IEEE Access* 10 (2022), pp. 102492–102531 (cit. on p. 9).

[26]H. Xu, L. Wang, W. Han, et al. „Assistance of UAVs in the Intelligent Management of Urban Space: A Survey“. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2023) (cit. on p. 9).

[27]N. H. Motlagh, M. Bagaa, and T. Taleb. „UAV-based IoT platform: A crowd surveillance use case“. In: *IEEE Communications Magazine* 55.2 (2017), pp. 128–134 (cit. on p. 9).

[28]R. Li, X. Sun, K. Yang, et al. „A lightweight wheat ear counting model in UAV images based on improved YOLOv8“. In: *Frontiers in Plant Science* 16 (2025), p. 1536017 (cit. on p. 10).

[29]P. N. Chowdhury, P. Shivakumara, L. Nandanwar, et al. „Oil palm tree counting in drone images“. In: *Pattern Recognition Letters* 153 (2022), pp. 1–9 (cit. on p. 10).

[30]P. Zhu, T. Peng, D. Du, et al. „Graph regularized flow attention network for video animal counting from drones“. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 5339–5351 (cit. on p. 10).

[31]Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou. „A survey of convolutional neural networks: analysis, applications, and prospects“. In: *IEEE transactions on neural networks and learning systems* 33.12 (2021), pp. 6999–7019 (cit. on p. 17).

[32]S. Khan, M. Naseer, M. Hayat, et al. „Transformers in vision: A survey“. In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41 (cit. on p. 17).

[33]R. Wu, L. Sun, Z. Ma, and L. Zhang. „One-step effective diffusion network for real-world image super-resolution“. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 92529–92553 (cit. on p. 17).

[34]G. Parmar, K. Kumar Singh, R. Zhang, et al. „Zero-shot image-to-image translation“. In: *ACM SIGGRAPH 2023 conference proceedings*. 2023, pp. 1–11 (cit. on p. 17).

[35]M. Hamazaspyan and S. Navasardyan. „Diffusion-enhanced patchmatch: A framework for arbitrary style transfer with diffusion models“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 797–805 (cit. on p. 17).

[36]K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. „Momentum contrast for unsupervised visual representation learning“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9729–9738 (cit. on p. 17).

[37]H. Zhang, F. Li, S. Liu, et al. „Dino: Detr with improved denoising anchor boxes for end-to-end object detection“. In: *arXiv preprint arXiv:2203.03605* (2022) (cit. on p. 17).

[38]A. Radford, J. W. Kim, C. Hallacy, et al. „Learning transferable visual models from natural language supervision“. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763 (cit. on p. 17).

[39]M. Awais, M. Naseer, S. Khan, et al. „Foundation Models Defining a New Era in Vision: a Survey and Outlook“. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025) (cit. on p. 17).

[40]A. Kirillov, E. Mintun, N. Ravi, et al. „Segment anything“. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 4015–4026 (cit. on p. 17).

[41]Y. Chang, Y. Cheng, U. Manzoor, and J. Murray. „A review of UAV autonomous navigation in GPS-denied environments“. In: *Robotics and Autonomous Systems* (2023), p. 104533 (cit. on p. 17).

[42]D. Pieczyński, B. Ptak, M. Kraft, M. Piechocki, and P. Aszkowski. „A fast, lightweight deep learning vision pipeline for autonomous UAV landing support with added robustness“. In: *Engineering Applications of Artificial Intelligence* 131 (2024), p. 107864 (cit. on p. 17).

[43]A. Savva, A. Zacharia, R. Makrigiorgis, et al. „ICARUS: Automatic autonomous power infrastructure inspection with UAVs“. In: *2021 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE. 2021, pp. 918–926 (cit. on p. 17).

[44]P. Aela, H.-L. Chi, A. Fares, T. Zayed, and M. Kim. „UAV-based studies in railway infrastructure monitoring“. In: *Automation in Construction* 167 (2024), p. 105714 (cit. on p. 17).

[45]B. Ptak and M. Kraft. „Mapping urban large-area advertising structures using drone imagery and deep learning-based spatial data analysis“. In: *Transactions in GIS* n/a.n/a (). eprint: `https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.13208` (cit. on p. 17).

[46]M. Lyu, Y. Zhao, C. Huang, and H. Huang. „Unmanned aerial vehicles for search and rescue: A survey“. In: *Remote Sensing* 15.13 (2023), p. 3266 (cit. on p. 17).

[47]M. Kraft, M. Piechocki, B. Ptak, and K. Walas. „Autonomous, onboard vision-based trash and litter detection in low altitude aerial images collected by an unmanned aerial vehicle“. In: *Remote Sensing* 13.5 (2021), p. 965 (cit. on p. 17).

[48]K. Li, G. Wan, G. Cheng, L. Meng, and J. Han. „Object detection in optical remote sensing images: A survey and a new benchmark“. In: *ISPRS journal of Photogrammetry and Remote Sensing* 159 (2020), pp. 296–307 (cit. on p. 18).

[49]J. Ding, N. Xue, G.-S. Xia, et al. „Object Detection in Aerial Images: A Large-Scale Benchmark and Challenges“. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1 (cit. on pp. 18, 20).

[50]D. C. Tsouros, S. Bibi, and P. G. Sarigiannidis. „A review on UAV-based applications for precision agriculture“. In: *Information* 10.11 (2019), p. 349 (cit. on p. 18).

[51]A. H. Buckman, M. Mayfield, and S. BM Beck. „What is a smart building?“ In: *Smart and sustainable built environment* 3.2 (2014), pp. 92–109 (cit. on p. 18).

[52]L. G. Anthopoulos. „Understanding the smart city domain: A literature review“. In: *Transforming city governments for successful smart cities* (2015), pp. 9–21 (cit. on p. 18).

[53]M. Kraft, P. Aszkowski, D. Pieczyński, and M. Fularz. „Low-cost thermal camera-based counting occupancy meter facilitating energy saving in smart buildings“. In: *Energies* 14.15 (2021), p. 4542 (cit. on p. 18).

[54]Q. Huang, Z. Ge, and C. Lu. „Occupancy estimation in smart buildings using audio-processing techniques“. In: *arXiv preprint arXiv:1602.08507* (2016) (cit. on p. 18).

[55]L. Wang, Y. Lu, Z. Gao, et al. „Berp: A blind estimator of room acoustic and physical parameters for single-channel noisy speech signals“. In: *arXiv preprint arXiv:2405.04476* (2024) (cit. on p. 18).

[56]S. Di Domenico, M. De Sanctis, E. Cianca, and G. Bianchi. „A trained-once crowd counting method using differential wifi channel state information“. In: *Proceedings of the 3rd International on Workshop on Physical Analytics*. 2016, pp. 37–42 (cit. on p. 18).

[57]G. Murali Krishna, A. Natarajan, and V. Krishnasamy. „Device-free crowd counting using multi-link WiFi CSI for occupant-driven energy management of HVAC systems“. In: *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects* 46.1 (2024), pp. 14318–14333 (cit. on p. 18).

[58]S. T. Kouyoumdjieva, P. Danielis, and G. Karlsson. „Survey of non-image-based approaches for counting people“. In: *IEEE Communications Surveys & Tutorials* 22.2 (2019), pp. 1305–1336 (cit. on p. 18).

[59]X. Zhang and G. Sexton. „A new method for pedestrian counting“. In: *Fifth International Conference on Image Processing and its Applications, 1995*. IET. 1995, pp. 208–212 (cit. on p. 18).

[60]M. Han, W. Xu, H. Tao, and Y. Gong. „An algorithm for multiple object trajectory tracking“. 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 (cit. on p. 19).

[61]P. Kilambi, E. Ribnick, A. J. Joshi, O. Masoud, and N. Papanikolopoulos. „Estimating pedestrian counts in groups“. In: *Computer Vision and Image Understanding* 110.1 (2008), pp. 43–59 (cit. on p. 19).

[62]A. Senior et al. „Tracking people with probabilistic appearance models“. In: *ECCV workshop on Performance Evaluation of Tracking and Surveillance Systems*. Citeseer. 2002, pp. 48–55 (cit. on p. 19).

[63]L. Snidaro, C. Micheloni, and C. Chiavedale. „Video security for ambient intelligence“. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 35.1 (2005), pp. 133–144 (cit. on p. 19).

[64]E. S. Wahyuni, R. R. Alinra, and H. Setiawan. „People counting for indoor monitoring“. In: *2017 International Conference on Computing, Engineering, and Design (ICCED)*. IEEE. 2017, pp. 1–5 (cit. on p. 19).

[65]R. Perko, T. Schnabel, A. Almer, and L. Paletta. „Towards view invariant person counting and crowd density estimation for remote vision-based services“. In: *23rd International Electrotechnical and Computer Science Conference, Portorož, Slovenia. Submitted for review*. 2014 (cit. on p. 19).

[66]Y.-L. Hou and G. K. Pang. „People counting and human detection in a challenging situation“. In: *IEEE transactions on systems, man, and cybernetics-part a: systems and humans* 41.1 (2010), pp. 24–33 (cit. on p. 19).

[67]X. Liu, P. H. Tu, J. Rittscher, A. Perera, and N. Krahnstoever. „Detecting and counting people in surveillance applications“. In: *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.* IEEE. 2005, pp. 306–311 (cit. on p. 19).

[68]V. B. Subburaman, A. Descamps, and C. Carincotte. „Counting people in the crowd using a generic head detector“. In: *2012 IEEE ninth international conference on advanced video and signal-based surveillance*. IEEE. 2012, pp. 470–475 (cit. on p. 19).

[69]Y. Hu, H. Chang, F. Nian, Y. Wang, and T. Li. „Dense crowd counting from still images with convolutional neural networks“. In: *Journal of Visual Communication and Image Representation* 38 (2016), pp. 530–539 (cit. on p. 19).

[70]D. Onoro-Rubio and R. J. López-Sastre. „Towards perspective-free object counting with deep learning“. In: *European conference on computer vision*. Springer. 2016, pp. 615–629 (cit. on p. 19).

[71]L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang. „Multi-scale convolutional neural networks for crowd counting“. In: *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2017, pp. 465–469 (cit. on p. 19).

[72]S. Kumagai, K. Hotta, and T. Kurita. „Mixture of counting CNNs: Adaptive integration of CNNs specialized to specific appearance for crowd counting“. In: *arXiv preprint arXiv:1703.09393* (2017) (cit. on p. 19).

[73]J. Shao, C. C. Loy, K. Kang, and X. Wang. „Crowded scene understanding by deeply learned volumetric slices“. In: *IEEE transactions on circuits and systems for video technology* 27.3 (2016), pp. 613–623 (cit. on p. 19).

[74]Y. Li, X. Zhang, and D. Chen. „CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1091–1100 (cit. on p. 19).

[75]Y. Hu, X. Jiang, X. Liu, et al. „NAS-count: Counting-by-density with neural architecture search“. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*. Springer. 2020, pp. 747–766 (cit. on p. 19).

[76]D. Liang, W. Xu, and X. Bai. „An end-to-end transformer model for crowd localization“. In: *European Conference on Computer Vision*. Springer. 2022, pp. 38–54 (cit. on p. 19).

[77]S. Wu and F. Yang. „Boosting detection in crowd analysis via underutilized output features“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 15609–15618 (cit. on p. 19).

[78]T. Han, L. Bai, L. Liu, and W. Ouyang. „STEERER: Resolving scale variations for counting and localization via selective inheritance learning“. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2023, pp. 21848–21859 (cit. on pp. 19, 33, 55).

[79]D. Du, Y. Qi, H. Yu, et al. „The unmanned aerial vehicle benchmark: Object detection and tracking“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 370–386 (cit. on p. 20).

[80]Y. Cao, Z. He, L. Wang, et al. „VisDrone-DET2021: The vision meets drone object detection challenge results“. In: *Proceedings of the IEEE/CVF International conference on computer vision*. 2021, pp. 2847–2854 (cit. on p. 20).

[81]J. Wang, W. Yang, H. Guo, R. Zhang, and G.-S. Xia. „Tiny object detection in aerial images“. In: *Int. Conf. Pattern Recog.* IEEE. 2021, pp. 3791–3798 (cit. on p. 20).

[82]G. Cheng, X. Yuan, X. Yao, et al. „Towards large-scale small object detection: Survey and benchmarks“. In: *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2023) (cit. on p. 20).

[83]J. Wang, C. Xu, W. Yang, and L. Yu. „A normalized Gaussian Wasserstein distance for tiny object detection“. In: *arXiv preprint arXiv:2110.13389* () (cit. on p. 20).

[84]C. Xu, J. Wang, W. Yang, et al. „RFLA: Gaussian receptive field based label assignment for tiny object detection“. In: *Eur. Conf. Comput. Vis.* Springer. 2022, pp. 526–543 (cit. on pp. 20, 55).

[85]Y.-K. Liao, G.-S. Lin, and M.-C. Yeh. „A Transformer-Based Framework for Tiny Object Detection“. In: *Asia Pacific Signal and Inf. Proc. Association Annual Summit and Conf. (APSIPA ASC)*. IEEE. 2023, pp. 373–377 (cit. on pp. 20, 55).

[86]L. Wen, D. Du, P. Zhu, et al. „Detection, tracking, and counting meets drones in crowds: A benchmark“. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7812–7821 (cit. on pp. 21–23, 29, 30, 33, 55, 60, 64).

[87]T. Asanomi, K. Nishimura, and R. Bise. „Multi-Frame Attention with Feature-Level Warping for Drone Crowd Tracking“. In: *Winter Conf. on Applications of Computer Vision*. 2023, pp. 1664–1673 (cit. on pp. 21, 22, 33, 55).

[88]O. Ronneberger, P. Fischer, and T. Brox. „U-net: Convolutional networks for biomedical image segmentation“. In: *Inter. conf. on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241 (cit. on pp. 21, 35).

[89]C. Feichtenhofer, A. Pinz, and A. Zisserman. „Detect to track and track to detect“. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 3038–3046 (cit. on p. 21).

[90]Z. Li, Y. Dong, L. Shen, et al. „Development and challenges of object detection: A survey“. In: *Neurocomputing* (2024), p. 128102 (cit. on p. 21).

[91]B. Mirzaei, H. Nezamabadi-Pour, A. Raoof, and R. Derakhshani. „Small object detection and tracking: a comprehensive review“. In: *Sensors* 23.15 (2023), p. 6887 (cit. on pp. 21, 60).

[92]A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. „Simple online and realtime tracking“. In: *2016 IEEE international conference on image processing (ICIP)*. IEEE. 2016, pp. 3464–3468 (cit. on pp. 22, 60).

[93]N. Wojke, A. Bewley, and D. Paulus. „Simple online and realtime tracking with a deep association metric“. In: *2017 IEEE international conference on image processing (ICIP)*. IEEE. 2017, pp. 3645–3649 (cit. on p. 22).

[94]Y. Du, Z. Zhao, Y. Song, et al. „StrongSORT: Make deepsort great again“. In: *IEEE Transactions on Multimedia* 25 (2023), pp. 8725–8737 (cit. on p. 22).

[95]J. Cao, X. Weng, R. Khirodkar, J. Pang, and K. Kitani. „Observation-centric SORT: Rethinking sort for robust multi-object tracking“. In: *arXiv preprint arXiv:2203.14360* () (cit. on p. 22).

[96]N. Aharon, R. Orfaig, and B. Bobrovsky. „BoT-SORT: Robust associations multi-pedestrian tracking“. In: *arXiv preprint arXiv:2206.14651* () (cit. on pp. 22, 62).

[97]H. Pirsiavash, D. Ramanan, and C. C. Fowlkes. „Globally-optimal greedy algorithms for tracking a variable number of objects“. In: *CVPR 2011*. IEEE. 2011, pp. 1201–1208 (cit. on p. 22).

[98]B. Ptak and M. Kraft. *Enhancing people localisation in drone imagery for better crowd management by utilising every pixel in high-resolution images*. 2025. arXiv: 2502.04014 [cs.CV] (cit. on pp. 24, 25, 33, 48, 51, 56, 57, 76, 77).

[99]B. Ptak and M. Kraft. *Improving trajectory continuity in drone-based crowd monitoring using a set of minimal-cost techniques and deep discriminative correlation filters*. 2025. arXiv: 2504.20234 [cs.CV] (cit. on pp. 26, 59–61, 64–69, 72, 79).

[100]J. Luiten, A. Osep, P. Dendorfer, et al. „HOTA: A higher order metric for evaluating multi-object tracking". In: *International journal of computer vision* 129 (2021), pp. 548–578 (cit. on p. 29).

[101]K. Bernardin and R. Stiefelhagen. „Evaluating multiple object tracking performance: the clear mot metrics". In: *EURASIP Journal on Image and Video Processing* 2008 (2008), pp. 1–10 (cit. on p. 29).

[102]Y. Li, C. Huang, and R. Nevatia. „Learning to associate: HybridBoosted multi-target tracker for crowded scene". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 2953–2960 (cit. on p. 29).

[103]W. Liu, M. Salzmann, and P. Fua. „Context-aware crowd counting". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 5099–5108 (cit. on p. 33).

[104]Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. „Single-image crowd counting via multi-column convolutional neural network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 589–597 (cit. on pp. 34, 35).

[105]X. He, Y. Zhou, J. Zhao, et al. „Swin transformer embedding UNet for remote sensing image semantic segmentation". In: *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), pp. 1–15 (cit. on p. 35).

[106]K. He, X. Zhang, S. Ren, and J. Sun. „Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778 (cit. on pp. 35, 39, 43).

[107]A. Howard, M. Sandler, G. Chu, et al. „Searching for mobilenetv3". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 1314–1324 (cit. on p. 35).

[108]M. Tan and Q. Le. „Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR. 2019, pp. 6105–6114 (cit. on pp. 35, 39, 43).

[109]E. Xie, W. Wang, Z. Yu, et al. „SegFormer: Simple and efficient design for semantic segmentation with transformers". In: *Advances in neural information processing systems* 34 (2021), pp. 12077–12090 (cit. on p. 35).

[110]T. Kroeger, R. Timofte, D. Dai, and L. Van Gool. „Fast optical flow using dense inverse search". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer. 2016, pp. 471–488 (cit. on p. 35).

[111]J. Hur and S. Roth. „Optical flow estimation in the deep learning age". In: *Modelling human motion: from human perception to robot design* (2020), pp. 119–140 (cit. on p. 35).

[112]B. Ptak, D. Pieczyński, M. Piechocki, and M. Kraft. „On-board crowd counting and density estimation using low altitude unmanned aerial vehicles—looking beyond beating the benchmark". In: *Remote Sensing* 14.10 (2022), p. 2288 (cit. on pp. 36, 37, 39, 40).

[113]F. Chollet. „Xception: Deep learning with depthwise separable convolutions“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258 (cit. on p. 38).

[114]O. Ronneberger, P. Fischer, and T. Brox. „U-net: Convolutional networks for biomedical image segmentation“. In: *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer. 2015, pp. 234–241 (cit. on pp. 39, 43).

[115]Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang. „Unet++: A nested u-net architecture for medical image segmentation“. In: *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*. Springer. 2018, pp. 3–11 (cit. on p. 39).

[116]L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam. „Encoder-decoder with atrous separable convolution for semantic image segmentation“. In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818 (cit. on p. 39).

[117]R. Wightman, H. Touvron, and H. Jégou. „Resnet strikes back: An improved training procedure in timm“. In: *arXiv preprint arXiv:2110.00476* (2021) (cit. on p. 39).

[118]M. Tan, B. Chen, R. Pang, et al. „Mnasnet: Platform-aware neural architecture search for mobile“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 2820–2828 (cit. on pp. 39, 43).

[119]M. Tan and Q. V. Le. „Mixconv: Mixed depthwise convolutional kernels“. In: *arXiv preprint arXiv:1907.09595* (2019) (cit. on p. 39).

[120]J. Deng, W. Dong, R. Socher, et al. „ImageNet: A large-scale hierarchical image database“. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255 (cit. on p. 41).

[121]T.-Y. Lin, M. Maire, S. Belongie, et al. „Microsoft COCO: Common objects in context“. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer. 2014, pp. 740–755 (cit. on p. 41).

[122]S. Xiang, P. M. Blok, J. Burridge, H. Wang, and W. Guo. „DODA: Diffusion for Object-detection Domain Adaptation in Agriculture“. In: *arXiv preprint arXiv:2403.18334* (2024) (cit. on p. 41).

[123]Q. Wang, J. Gao, W. Lin, and Y. Yuan. „Learning from synthetic data for crowd counting in the wild“. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 8198–8207 (cit. on p. 41).

[124]B. Ptak and D. Pieczynski. „CountingSim: Synthetic Way To Generate a Dataset For The UAV-view Crowd Counting Task“. In: (2022) (cit. on pp. 41, 42).

[125]*Unity - game engine*. `https://unity.com/`. Accessed: 2025-02-20 (cit. on p. 41).

[126]S. Borkman, A. Crespi, S. Dhakad, et al. „Unity Perception: Generate Synthetic Data for Computer Vision“. In: *CoRR* abs/2107.04259 (2021). arXiv: `2107.04259` (cit. on p. 41).

[127]P. Ghamisi, B. Rasti, N. Yokoya, et al. „Multisource and multitemporal data fusion in remote sensing: A comprehensive review of the state of the art“. In: *IEEE Geoscience and Remote Sensing Magazine* 7.1 (2019), pp. 6–39 (cit. on p. 45).

[128]Y. Yang, C. Wang, Z. Cai, et al. „GSDDet: Ground sample distance guided object detection for remote sensing images“. In: *IEEE Transactions on Geoscience and Remote Sensing* (2023) (cit. on p. 45).

[129]M. Wilinski, B. Ptak, and M. Kraft. „Elevating point-based object detection in UAVs: A deep learning method with altitude fusion“. In: () (cit. on pp. 45, 46).

[130]D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal. „Image super-resolution: A comprehensive review, recent trends, challenges and applications“. In: *Information Fusion* 91 (2023), pp. 230–260 (cit. on p. 47).

[131]W. Shi, J. Caballero, F. Huszár, et al. „Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network“. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2016, pp. 1874–1883 (cit. on p. 49).

[132]Q. Hou, D. Zhou, and J. Feng. „Coordinate attention for efficient mobile network design“. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2021, pp. 13713–13722 (cit. on p. 49).

[133]H. Law and J. Deng. „Cornernet: Detecting objects as paired keypoints“. In: *Proc. of the European conf. on computer vision (ECCV)*. 2018, pp. 734–750 (cit. on p. 52).

[134]T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. „Focal loss for dense object detection“. In: *Int. Conf. Comput. Vis.* 2017, pp. 2980–2988 (cit. on p. 52).

[135]C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao. „YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors“. In: *IEEE Conf. Comput. Vis. Pattern Recog.* 2023, pp. 7464–7475 (cit. on p. 52).

[136]J. C. Miranda, J. Gené-Mola, M. Zude-Sasse, et al. „Fruit sizing using AI: a review of methods and challenges“. In: *Postharvest Biology and Technology* 206 (2023), p. 112587 (cit. on p. 60).

[137]J.-Y. Bouguet et al. „Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm“. In: *Intel corporation* 5.1-10 (2001), p. 4 (cit. on p. 63).

[138]M. Danelljan, G. Bhat, F. Shahbaz Khan, and M. Felsberg. „ECO: Efficient convolution operators for tracking“. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 6638–6646 (cit. on p. 67).

[139]K. Chatfield. „Return of the devil in the details: Delving deep into convolutional nets“. In: *arXiv preprint arXiv:1405.3531* (2014) (cit. on p. 67).

[140]T. Trzcinski, B. Twardowski, B. Zieliński, K. Adamczewski, and B. Wójcik. „Zero-Waste Machine Learning“. In: *ECAI 2024*. IOS Press, 2024, pp. 43–49 (cit. on p. 67).

[141]R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003 (cit. on p. 75).