RSTDE
Remote Sensing Technology in Defense and Environment
RSTDE 2(2): 117–131
ISSN 3062-8970

Institute for Advanced Science, Social and Sustainable Future
MORALITY BEFORE KNOWLEDGE

Utilization of deep learning in PTZ (pan-tilt-zoom) camera control systems for geospatial-based intelligence surveillance

Farhat Bashir[1,*], Syachrul Arief[2]
[1] *Sekolah Tinggi Intelijen Negara, Bogor, West Java 16810, Indonesia;*
[2] *Geospatial Information Agency, Bogor, West Java 16911, Indonesia.*
*Correspondence: farhatbashir1093@gmail.com

Received Date: June 3, 2025 Revised Date: July 27, 2025 Accepted Date: August 31, 2025

ABSTRACT

Background: The rising complexity of threats to public safety and critical infrastructure has highlighted the limitations of conventional human-operated surveillance systems, creating the need for adaptive, intelligent, and real-time monitoring solutions. Advances in artificial intelligence (AI), computer vision, and geospatial technologies provide opportunities to enhance surveillance through automated detection, analysis, and response. This article examines the integration of pan-tilt-zoom (PTZ) cameras with deep learning models, geospatial data, and distributed computing frameworks as the foundation for next-generation intelligent surveillance systems. **Methods:** The study employs a narrative review approach, synthesizing recent developments in PTZ camera calibration, convolutional neural networks (CNN), reinforcement learning for autonomous control, and fog computing for distributed video analysis. Research spanning dual-mode fisheye-PTZ systems, lightweight CNN architectures, geospatial data integration, and Internet of Robotic Things (IoRT) frameworks is analyzed to demonstrate practical applications in smart city, industrial, and defense contexts. **Findings:** Findings reveal that PTZ cameras, when coupled with deep learning and geospatial intelligence, achieve high accuracy in real-time object tracking, small-object recognition, and anomaly detection, with minimal latency under dynamic conditions. Experimental evidence shows error margins below 2% in calibration models and near-perfect accuracy in long-range facial recognition. Integration with fog computing and IoRT enhances responsiveness, scalability, and contextual awareness, while reinforcement learning enables autonomous decision-making for robots and camera networks. **Conclusion:** The article concludes that combining PTZ hardware precision, AI-based visual analysis, and spatial data intelligence transforms surveillance systems from passive observers into proactive, adaptive, and collaborative agents. However, challenges remain in ensuring robustness under real-world conditions, minimizing latency, and addressing operational usability. **Novelty/Originality of this article:** This work presents a holistic synthesis of AI-driven vision, PTZ camera control, geospatial intelligence, and distributed architectures, offering an integrated framework for developing adaptive and context-aware surveillance systems in the digital era.

KEYWORDS: geospatial intelligence; intelligent surveillance; PTZ cameras.

1. Introduction

The increasing complexity of threats to public safety and critical infrastructure necessitates surveillance systems that are not only passive but also adaptive, intelligent, and capable of responding to threats in real-time. Conventional surveillance systems that rely on human operators have proven insufficient in addressing dynamic field situations,

Cite This Article:
Bashir, F., & Arief, S. (2025). Utilization of deep learning in PTZ (pan-tilt-zoom) camera control systems for geospatial-based intelligence surveillance. *Remote Sensing Technology in Defense and Environment, 2*(2), 117-131. https://doi.org/10.61511/rstde.v2i2.2025.2249

particularly in detecting incidents such as acts of violence or security breaches. Therefore, the integration of artificial intelligence (AI) into surveillance systems has emerged as a strategic approach to providing more reliable security solutions. The application of AI in monitoring systems has advanced through the use of facial recognition technologies, automated access control systems, and video surveillance powered by machine learning and computer vision. Studies have shown that integrating AI into perimeter security systems—such as those deployed in data centers—can enhance operational efficiency and intrusion detection through visual authentication methods and automatic monitoring of sensitive areas (Villegas-Ch & García-Ortiz, 2023). This technology not only strengthens access control but also accelerates threat response with greater accuracy compared to manual approaches.

Simultaneously, video-based violence detection technologies have undergone rapid development. Recent research has introduced deep learning models based on U-Net and LSTM architectures capable of recognizing violent actions from surveillance video with high accuracy and low computational cost (Vijeikis et al., 2022). These models enable on-device edge processing, thereby speeding up detection and reducing the bandwidth required to transmit full video streams to central servers. The combination of AI-based access control systems and automated violence behavior detection results in surveillance systems that are more responsive, intelligent, and resource-efficient. In the context of national security and smart cities, such systems hold great potential for delivering proactive protection without compromising efficiency or privacy. Consequently, the development of AI-powered surveillance systems using end-to-end approaches has become a critical priority in efforts to strengthen security resilience in the digital age.

In developing such systems, hardware capable of dynamic movement while maintaining high visual accuracy is essential. This is where pan-tilt-zoom (PTZ) cameras play a crucial role, particularly due to their ability to cover wider areas through rotation and optical zoom. However, continuous changes in viewing angles and zoom levels pose technical challenges, notably in the form of shifting camera calibration parameters, which can reduce detection accuracy. A generalized model has been proposed for dual-PTZ cameras that accounts for optical center displacement and misalignment of rotation axes, linking camera parameters to actual feedback values through mathematical modeling and data fitting. Experimental results show that this method achieves focal length errors below 4% and rotation and translation errors below 1%, making it an efficient and accurate solution for visual and 3D spatial monitoring applications (Mao et al., 2022). This feedback-parameter-based PTZ camera calibration method allows the system to automatically recalibrate intrinsic and extrinsic parameters across various pan, tilt, and zoom settings.

Beyond image acquisition accuracy using hardware such as PTZ cameras, the effectiveness of modern surveillance systems also depends heavily on how video data is analyzed and transmitted across complex networks. To address these challenges, distributed architectural approaches combining deep learning, the Internet of Things (IoT), and fog computing are being increasingly developed. For example, a weapon-detection-based surveillance system was designed using the YOLOv5 model, integrated into an intelligent network architecture based on software-defined networking (SDN) and fog nodes. In this system, surveillance cameras connected to edge devices do not merely record but also perform inference processes locally—if a hazardous object is detected, only metadata and relevant video snippets are sent to the command center via pre-optimized SDN pathways (Fathy & Saleh, 2022). This approach enables data processing to occur near the source (fog layer), reducing reliance on cloud data centers and minimizing transmission latency. Furthermore, the architecture supports system scalability, as communication and network control flows can be flexibly programmed to meet evolving security needs. By integrating this localized data processing strategy with adaptively calibrated PTZ camera systems, surveillance platforms can be transformed into intelligent monitoring solutions that are not only responsive but also context-aware and resource-efficient. This combination is well-suited for modern security systems requiring high reaction speeds, as well as military environments that demand real-time, high-precision spatial analysis.

In intelligent surveillance systems, beyond visual and analytical aspects, mobility also plays a significant role. One relevant approach involves the use of autonomous patrol robots powered by computer vision and artificial intelligence. Zheng et al. (2022) developed an enhanced path-planning method for indoor patrol robots using deep reinforcement learning (DRL). In their research, PTZ cameras were employed to capture spatial information, which was then processed by learning algorithms to automatically determine the robot's movement direction and speed. The main focus of this approach was to refine the reward and punishment functions to allow faster algorithm convergence and produce optimal paths that avoid obstacles and reach targets (Zheng et al., 2022). This implementation demonstrates great potential for active robot-based surveillance, particularly in enclosed environments such as industrial facilities, shopping centers, or data centers. However, in dynamic operational settings, fog computing-based approaches play a critical role in enabling localized decision-making.

Jing & Xue (2024) proposed an IoT optimization method based on fog computing, enhanced by an improved convolutional neural network (CNN). The refined CNN is used to continuously estimate values within the framework of a markov decision process, which is particularly suitable for navigation and robotic decision-making applications. By integrating the CNN architecture into fog nodes, analysis and response processes can be conducted closer to data sources such as robots or surveillance cameras, thereby reducing latency and increasing operational efficiency. The combination of deep reinforcement learning for navigation control and CNN in a fog computing environment enables patrol robots to respond to their surroundings rapidly and adaptively. This system not only optimizes patrol routes but also simultaneously performs visual analysis to detect nearby threats. The integration suggests that AI-powered autonomous robots can become an integral part of intelligent spatial surveillance systems that are efficient, context-aware, and capable of real-time response in modern security scenarios.

When designing surveillance systems that are not only adaptive and intelligent but also spatially contextual, the integration of artificial intelligence with geographic information becomes essential. This approach, known as GeoAI, combines artificial intelligence, machine learning, and deep learning with geographic information systems. According to Choi (2023), GeoAI enables automatic and accurate spatial analysis for purposes such as object detection, land-use change mapping, and location-based tracking. These capabilities greatly support PTZ camera systems in dynamic surveillance and target tracking. The automatic processing of spatial imagery is further reinforced by research from Puttinaovarat & Horkaew (2022), who developed a geospatial platform for classifying green space areas using deep learning. Their system analyzes satellite imagery with high accuracy using a ZFNet-based model and presents results in real-time via a digital platform. This shows that the combination of GIS and deep learning can be employed not only for environmental monitoring but also to support priority area mapping and visual detection in spatial intelligence systems. In the context of PTZ cameras, such spatial understanding provides a foundation for more focused and contextually aware camera movements, ensuring that rotation and zoom target spatially significant areas or objects for further analysis.

The effectiveness of PTZ camera-based monitoring systems depends on both the hardware's ability to capture high-quality imagery and the software's capacity to analyze this information within an accurate spatial context. Supporting this, the selection of camera sensors and the use of deep learning-based processing approaches are key to system success. A study by Llauradó et al. (2023) examined challenges in recognizing human faces from long distances, a common issue in surveillance systems deployed in public smart city spaces. Their method involved developing a performance-based evaluation of image sensors, considering object distance from the camera, lens focal length, and final face resolution in the frame. Using a restructured dataset based on long-range images from the Georgia tech and quality dataset for distance faces databases, they tested several sensor configurations and found that increasing focal length significantly improved face recognition accuracy to over 99% at distances of 15–20 meters. These findings provide valuable insights for determining the optimal positioning and specifications of PTZ cameras

in environments requiring precise identification, such as airports, border zones, or government buildings.

Meanwhile, Huillca & Fernandes (2022) highlighted the role of conventional cameras as active spatial sensors through an innovative approach. They demonstrated that with just a single still image, a ship's speed can be estimated accurately by analyzing wave patterns (Kelvin wakes) and applying principles of projective geometry. By leveraging homography and vanishing line estimation, they reconstructed the water surface in metric space, identified wave points, and calculated the ship's speed. Validation using radar data as ground truth showed the method to be reliable. This confirms that cameras are not merely visual documentation tools but also accurate geospatial data sources that can be processed using AI-based analytical methods. By combining the optical capabilities of cameras with AI-driven spatial approaches, systems can dynamically focus on strategic areas, enhance monitoring accuracy, and generate deep geospatial intelligence—forming a critical pillar of future adaptive and intelligent surveillance systems.

2. Methods

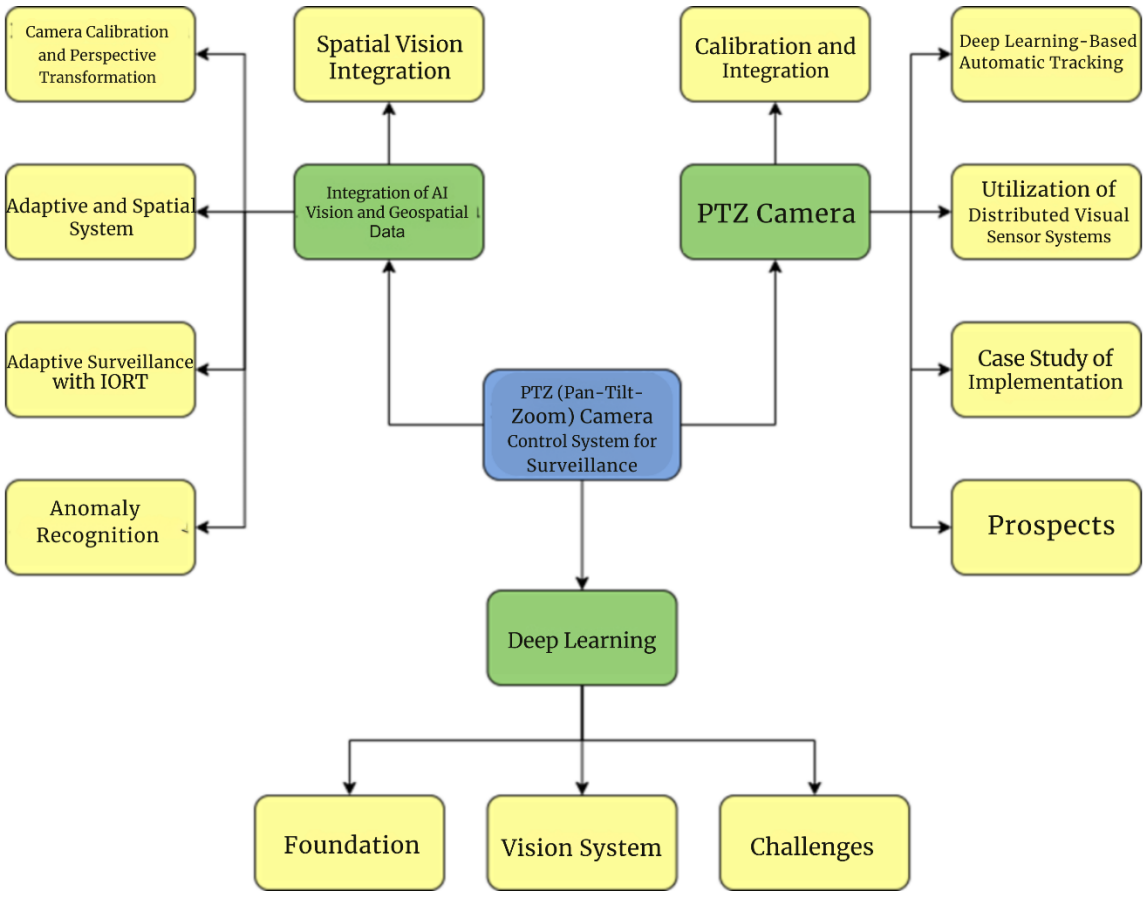

Fig. 1. Brainstorm

The Figure 1 illustrates a conceptual framework of a Pan-Tilt-Zoom (PTZ) camera control system for surveillance, integrated with artificial intelligence, geospatial data, and deep learning technologies. The framework begins with PTZ cameras that undergo calibration and integration, alongside the fusion of AI-based vision and geospatial data to enable spatial vision and adaptive surveillance. These capabilities support key functions such as anomaly recognition, adaptive surveillance with IoRT, and spatially responsive systems. At the core, deep learning serves as the foundation, enhancing the vision system while addressing implementation challenges. Furthermore, the system is evaluated through case studies, the utilization of distributed visual sensor networks, and deep learning–based automatic tracking, while also highlighting its prospects for future surveillance applications.

3. Results and Discussion

3.1 Pan-tilt-zoom (PTZ) camera technology

Pan-tilt-zoom (PTZ) cameras are vital components in modern surveillance systems due to their ability to move horizontally (pan), vertically (tilt), and perform image magnification (zoom). These features allow for flexible monitoring of wide areas using a single camera, which can be directed automatically or manually. In complex and dynamic urban environments, PTZ cameras are particularly useful for tracking object movements or suspicious activities without losing visual detail. In a study by Arroyo et al. (2021), an intelligent surveillance system was constructed by combining a fisheye camera (for full 180° × 360° coverage) with a PTZ camera (for focused observation). When the fisheye camera detects movement or activity in a particular area, the system automatically directs the PTZ camera to that location to capture clearer and more detailed images. This automated control is enabled by the implementation of ONVIF (Open Network Video Interface Forum), an open communication standard that allows various IP-based video surveillance devices to interconnect and interoperate. With ONVIF, PTZ cameras can be controlled through multiple video management systems, regardless of brand. Functions such as pan, tilt, zoom, preset positioning, and automated commands can be accessed over IP networks. ONVIF simplifies the integration and automation of surveillance systems, especially in AI- and geospatial-based intelligent systems. With its adaptive capability and high interoperability, ONVIF-enabled PTZ cameras are not merely auxiliary tools but integral elements of efficient and responsive intelligent monitoring systems.

To ensure that PTZ cameras operate effectively within automated surveillance systems, proper calibration and integration of sensors, control systems, and other visual components are essential. In the study by Arroyo et al. (2021), a dual-mode surveillance system was designed using a fisheye camera for broad monitoring and a PTZ camera for detailed observation. To accurately direct the PTZ camera to targets detected within the fisheye image, the system employed a back-projection-based calibration approach. Unlike traditional re-projection methods, back-projection maps pixels from the fisheye image onto a flat plane (ground surface) where the object is assumed to exist. Assuming all objects move on a planar surface, the system geometrically calculates the target's location and determines the angle required for the PTZ camera to be precisely directed to that point. This allows for continued observation without relying on stereo feature matching or complex triangulation (Arroyo et al., 2021). Additionally, the system uses the ONVIF protocol to integrate IP camera control directly into the software platform. ONVIF enables automated pan, tilt, and zoom controls of the PTZ camera based on detections from the fisheye camera, rendering the system responsive, real-time, and applicable in complex urban or traffic surveillance environments.

Following successful calibration for accurate target localization within the observation field, the next stage involves automatic tracking using deep learning. In research by You et al. (2022), a PTZ-based tracking system was developed using the YOLOv4 algorithm for real-time object detection in video streams. The detection outputs were then converted from image coordinates to real-world coordinates using the perspective-n-point (PnP) algorithm. Subsequently, a Pan-Tilt-Height (PTH) model was utilized to compute PTZ camera movements. This model translates the spatial position of a target into camera control instructions—namely pan (horizontal), tilt (vertical), and height—which correlate with camera positioning or zoom. Through this approach, the system is capable of autonomously maintaining focus on a moving target without human intervention. This method aligns closely with the previously discussed back-projection strategy in dual-camera systems, as both approaches leverage spatial information to guide the camera with high precision. Experiments demonstrated that this method yielded high accuracy, with a maximum position error of only 2.31 cm, an average of 1.245 cm, and a maximum camera direction error of 1.78°, with an average of 0.656°. Using this method, the PTZ camera can

automatically and precisely track moving targets within the monitored area, enhancing the effectiveness of AI-based spatial surveillance systems.

Once the PTZ camera has been successfully integrated into an AI-based tracking system, the next step involves its deployment within a distributed camera network, or Visual Sensor Network (VSN). In a study by Giordano et al. (2022), PTZ cameras were used collaboratively alongside static cameras to create a real-time 3D monitoring system for indoor environments. This system is capable of detecting and tracking multiple targets simultaneously, thanks to the PTZ camera's ability to adjust its viewing angle and zoom level based on object movement dynamics. To efficiently manage camera positioning, the system employed a game theory approach that enables each PTZ camera to make optimal decisions. Simulation results showed that the system could reduce overlapping coverage areas between cameras and maintain high tracking accuracy, with an average tracking error of only 0.656°, and a maximum of 1.78°. These advantages were achieved without a centralized control unit, making the system more scalable and adaptive. This integration strongly supports the earlier discussed PTH model, as the spatial positions of targets can be directly utilized by PTZ cameras in the VSN to conduct intelligent, coordinated tracking in real-world environments.

After demonstrating the effectiveness of PTZ cameras in distributed systems like VSNs, field studies illustrate how this technology is applied in real-world scenarios. In the Remote AFIS (Aerodrome Flight Information Service) project, Reuschling & Jakobi (2022) developed a remote tower surveillance system using a combination of panoramic and PTZ cameras. With the aid of a VR headset, operators could intuitively control the PTZ camera; however, challenges such as cybersickness caused by abrupt camera movements and high latency in camera responses were noted. On the other hand, Church et al. (2024) evaluated the effectiveness of zoom cameras mounted on drones, comparing the performance of optical, digital, and hybrid zoom types. Their findings revealed that hybrid zoom cameras with visual computing capabilities could accurately read livestock ear tags from a distance of up to 60 meters—demonstrating a performance closely aligned with PTZ cameras for detailed long-range observation. These studies show that while PTZ cameras are highly reliable, their real-world integration still faces practical challenges such as motion stability, bandwidth limitations, and user comfort.

PTZ cameras play a crucial role in modern surveillance systems through their automatic pan, tilt, and zoom capabilities. When integrated with deep learning and spatial calibration, they enable real-time, high-precision target tracking. In distributed visual networks, PTZ cameras operate collaboratively to ensure efficient coverage. Real-world case studies confirm their field effectiveness, although issues such as latency and motion smoothness require further optimization to enhance system performance.

3.2 Deep learning in vision systems

PTZ (Pan-Tilt-Zoom) cameras require an intelligent visual system to enable automatic and adaptive object tracking. The primary role of deep learning in this system is to serve as a signal processing unit capable of dynamically interpreting and responding to image data. Convolutional neural networks (CNN) are a fundamental pillar of modern computer vision due to their hierarchical and progressive ability to extract spatial features from images. Research by Alzubaidi et al. (2021) affirms that CNN enable richer visual representations without the need for manual feature engineering, laying the groundwork for PTZ cameras to recognize objects with high precision. On the other hand, approaches such as imitation learning pave the way for PTZ cameras to learn from observations, for instance, by tracking keypoints of moving objects. This "learning by observing" mechanism enhances the camera's adaptability, allowing it to continuously evolve from visual data (Sun et al., 2022).

To ensure PTZ cameras can respond to object movement accurately and in real-time, a visual detection system is needed that is not only reliable but also computationally efficient. When surveillance systems must handle continuous video streams—especially in complex environments such as urban traffic or public spaces—processing efficiency becomes as

crucial as accuracy. Therefore, selecting an appropriate CNN architecture plays a strategic role in designing an effective vision system for PTZ cameras. The study by Junos et al. (2022) highlights how lightweight CNN architectures, such as those based on MobileNetV2, can be employed for object recognition in hardware-constrained systems. They demonstrated that the combination of bottleneck convolutions and squeeze-and-excitation (SE) mechanisms not only maintains human detection accuracy in video streams but also significantly reduces processing latency. This approach is particularly relevant for PTZ camera integration, as it allows the system to operate smoothly under real-time constraints.

In addition, advancements in spatial modeling offer further value in enhancing the initial capabilities of vision systems. The research by Fuertes et al. (2022) on human detection using omnidirectional cameras illustrates how a spatial grid approach with a foveatic classifier can focus processing resources on the most relevant areas within an image. This strategy results in much greater processing efficiency, especially when applied as the initial layer in a PTZ camera system. When a fisheye camera captures a wide area and detects movement, such a system can quickly redirect the PTZ camera to the relevant region for detailed observation.

This interplay reinforces the concept of a dual-mode system previously described, in which the combination of a wide-angle camera and a PTZ camera distributes roles between broad detection and focused observation. In this system, detection speed and model efficiency critically determine the PTZ camera's responsiveness in adjusting its pan, tilt, and zoom angles. The use of lightweight CNN and adaptive spatial processing serves as a crucial bridge between initial visual sensors and the camera's motion mechanisms. Thus, accuracy alone is not the primary parameter in a PTZ-based vision system—efficiency in the underlying CNN architecture is equally vital. The combination of intelligent spatial processing and lightweight models opens the door for PTZ systems to operate more adaptively, energy-efficiently, and reliably over time, even when exposed to dynamic operational conditions and massive visual data.

Once the vision system efficiently detects an object, the next challenge is translating the detection results into autonomous PTZ camera control. At this stage, deep learning functions not only as a visual recognizer but also as a motion control mechanism. PTZ cameras, with their pan, tilt, and zoom capabilities, offer substantial flexibility in surveillance, but this potential is only maximized if the system can be intelligently and adaptively controlled. A promising approach is demonstrated by Zhai & Zhang (2022), who developed a method for detecting low-flying UAV (Unmanned Aerial Vehicle) targets using an enhanced YOLOv3 model. In addition to adapting the YOLO architecture for greater sensitivity to small and fast-moving targets, they also integrated a PID controller to directly steer the PTZ camera toward detected targets. This system is capable of autonomously keeping the object centered within the field of view, eliminating the need for manual intervention. This aligns well with the core principle of PTZ systems based on Pan-Tilt-Height (PTH), where the target's position in image coordinates is converted into physical parameters for camera angle control.

A more applied system was developed by Fan et al. (2021), in the context of a UAV-based Explosive Ordnance Disposal (EOD) system. They combined a PTZ camera with a YOLOv5 algorithm trained to detect grenades. The visual detection results were then used to direct the PTZ in real time toward targets detected from the air. Interestingly, this process involved not only image detection but also integration between remote control mechanisms and CNN-based recognition algorithms, showcasing how visual and control components can be merged into a unified intelligent system. In the context of a dual-mode system, this approach mirrors the division of roles where the fisheye camera serves as the initial detector, and the PTZ camera performs detailed tracking. Once an object is imaged, its coordinates are computed and translated into pan-tilt commands via back-projection or the PTH model. Without CNN and detection models like YOLO, this PTZ control process would be significantly slower or potentially non-automated. Thus, the integration of CNN (as detectors) with the camera control system (as actuators) positions deep learning not just as a recognition tool but also as a precise motion controller. This synergy enables the

development of adaptive vision systems capable of tracking objects in dynamic environments without human intervention.

In developing autonomous PTZ camera systems, the performance of the vision system cannot rely solely on sophisticated deep learning architectures. More critically, the CNN model must adapt to real-world visual conditions, such as low lighting, low-contrast objects, or image noise from high zoom levels. To accurately follow targets in suboptimal conditions, the CNN model must demonstrate strong generalization capabilities. This means the model must remain stable even when exposed to environmental variations not explicitly represented in its training data. Two relevant field studies show how CNN can be effectively implemented in real-world visual systems. In a study by García-Segura et al. (2023), CNN were applied to road inspection videos to detect surface damage such as cracks and wear. The images were captured from moving vehicles under diverse road conditions, including direct sunlight, tree shadows, and heavy traffic. The model successfully recognized damage patterns with high accuracy, illustrating that CNN can function robustly even under varying textures, contrasts, and visual noise—conditions commonly encountered in PTZ systems.

Meanwhile, Ren et al. (2022) investigated the detection of residential energy systems (e.g., solar panels) from drone imagery using a U-Net-based CNN with a ResNet50 backbone. The main challenges included unstable image quality, low resolution, extreme viewing angles, and shadow interference. However, experimental results showed that the model maintained high accuracy even under far-from-ideal visual conditions. This success was attributed to training on diverse and realistic datasets rather than synthetic or overly structured data. Both studies provide concrete evidence that CNN models trained with varied and realistic data possess high adaptability—an essential quality for PTZ cameras, which must track objects not in laboratory settings but in real-world scenarios: in public spaces, under changing lighting conditions, and with rapidly moving objects. The effectiveness of PTH-based PTZ tracking systems hinges on initial visual precision. If a CNN lacks generalization, minor angular errors from misdetection could cause loss of tracking focus.

Moreover, the findings from these two studies underscore that CNN model robustness cannot be separated from training strategies. Architectural depth alone is insufficient; data diversity, visual augmentation, and contextual validation are critical to ensuring the model is field-deployable. Therefore, in PTZ camera systems, building representative datasets and conducting comprehensive validation are not optional additions—they are fundamental requirements. In conclusion, CNN generalization is not merely an added advantage but a foundational pillar of reliable PTZ vision systems across diverse operational conditions. Drawing from empirical field evidence, this approach offers a practical solution to bridge the gap between technical capabilities and real-world application needs.

Finally, after discussing how CNN can be integrated, optimized, and generalized within PTZ camera vision systems, one crucial aspect remains: model validation. Deep learning models will only be effective in autonomous surveillance if they can maintain stable performance under real-world conditions—not just during laboratory trials. Kattenborn et al. (2022) warned that many CNN models that appear accurate during internal evaluation often fail in deployment because they are validated on data too similar to the training set. In PTZ systems, even minor detection errors can result in incorrect pan or tilt directions, causing the camera to lose track of its target. Research by Li et al. (2022) reinforces the importance of testing under various image conditions. Models must be validated on visual variations that approximate field scenarios: low lighting, fast-moving objects, and complex backgrounds. Without this, PTZ systems that seem precise during testing may become unresponsive in deployment. Rigorous validation is not merely about numerical accuracy—it involves building trust that the PTZ camera can truly "see and react" correctly, under any circumstances and at any time.

3.3 Integration of AI vision and geospatial data in PTZ camera control

The integration of visual artificial intelligence (AI vision) and geospatial data forms a fundamental basis for modern surveillance systems based on PTZ (Pan-Tilt-Zoom) cameras. These devices no longer serve merely as passive observers; they have evolved into active systems capable of understanding visual-spatial contexts and automatically adjusting monitoring direction and focus. Google earth engine (GEE) serves as the backbone for large-scale spatial data processing as well as a platform for training and deploying cloud-based machine learning algorithms. Yang et al. (2022) demonstrated how GEE provides access to multi-temporal satellite imagery and integrates AI techniques for automated spatial analysis, including land classification, change detection, and spatial object tracking.

Meanwhile, the integration of GIS technology and augmented reality (AR) introduces a new dimension in spatial interaction, especially for PTZ camera control interfaces. In a study by Bazargani et al. (2022), it was shown how AR-GIS systems enable users to visualize spatial objects directly within real-world environments through digital overlays, facilitating positioning and location-based visual decision-making. However, achieving intelligent spatial decision-making presents significant challenges in managing and integrating geospatial big data. Al-Yadumi et al. (2021) highlighted that data source diversity, heterogeneous data formats, and semantic integration requirements pose major obstacles to consolidating spatial information, particularly under demands for real-time analytics and system interoperability. In the context of PTZ cameras, successful geospatial data integration directly influences the effectiveness of camera orientation based on priority zones or high-risk points.

In intelligent PTZ camera surveillance systems, the camera's ability to respond to the target's position in the real world is a critical requirement. The mechanisms controlling camera direction and zoom must be closely linked with spatial data and object behavior to enable adaptive monitoring, in line with the pan-tilt-height (PTH) control principle. Fahim et al. (2023) introduced AcTrak, an automatic PTZ camera control framework based on reinforcement learning. The camera dynamically zooms in on detected objects and zooms out to detect new objects appearing within the surveillance area. AcTrak is capable of selecting the optimal camera configuration based on the target's current status, optimizing the trade-off between object tracking and area coverage.

In traffic monitoring contexts, PTZ cameras are also employed to accurately measure vehicle speed and position using the inverse perspective mapping (IPM) method. This method performs real-time camera calibration using road markings and vanishing points, transforming camera views into actual road coordinates (Shi et al., 2023). From this, vehicle positions and potential risks can be directly computed. This technique enables PTZ cameras to function as active spatial sensors in complex traffic surveillance. The system's accuracy heavily depends on object detection capability, especially for small objects in open environments or remote sensing scenarios. Shivappriya et al. (2021) developed the AAF-Faster RCNN method, which integrates additive activation functions into cascade detection layers. This model enhances small object detection with clearer boundaries and faster convergence, making it highly suitable to support PTZ systems that require accuracy even at small scales or long distances.

The ability of PTZ cameras to detect, track, and understand the contextual activities of objects is central to adaptive spatial-visual intelligence. Particularly in dense and dynamic environments, rapid small object detection and behavior interpretation become increasingly important. The latest detection models, such as DC-YOLOv8, introduce significant improvements in small object detection through architectural modifications ranging from novel downsampling methods to enhanced feature fusion. Lou et al. (2023) demonstrated that this algorithm improves detection accuracy on small-object datasets such as VisDrone and TinyPerson, making DC-YOLOv8 highly relevant for directing PTZ cameras toward small objects in open or public spaces. Meanwhile, model selection must also consider efficiency. Nepal & Eslamiat (2022) compared YOLOv3, YOLOv4, and YOLOv5 in the context of UAV emergency landing systems. Their results show that YOLOv5 excels in

accuracy, albeit with slight latency compared to YOLOv3. This combination of speed and precision supports its application in PTZ camera systems for real-time responses to critical events such as intrusions or evacuations.

However, AI vision capabilities extend beyond object recognition. Another crucial aspect is human activity classification. Arshad et al. (2022) noted that most current human activity recognition (HAR) systems rely on visual data and utilize CNN and LSTM architectures. PTZ cameras integrating these models can identify activities such as walking, running, or aggressive behavior, enabling automatic camera redirection based on activity intensity rather than mere physical location. Advanced feature fusion approaches are also explored by Khan et al. (2021), who proposed serial-based extended fusion and weighted classification methods (kurtosis-controlled KNN) for human action recognition. This approach maintains high accuracy with low computational overhead, making it highly suitable for edge-device-based PTZ cameras that require fast, energy-efficient, and memory-efficient responses. By combining lightweight object detection models, real-time tracking, and efficient human action recognition, PTZ camera systems transform from passive observers into active surveillance agents that intelligently respond to spatial presence, movement, and behavior according to visual-spatial cues. Rather than merely following objects based on visual detection alone, modern PTZ camera systems are designed to operate based on comprehensive spatial understanding. This means camera decisions to highlight or monitor an area consider not only object presence but also spatial structure, environmental risks, and threat potentials grounded in geospatial context.

For this purpose, point cloud-based spatial segmentation has become an important approach, especially in complex areas such as dense urban zones or disaster terrains. However, point cloud data tend to be unstable due to sparsity, noise, and class imbalance. Grilli et al. (2023) proposed the use of knowledge enhanced neural networks (KENN), which combine neural network learning with symbolic logic rules. This system enhances spatial segmentation with semantic context, enabling PTZ cameras to more accurately direct their views based on real spatial structures rather than mere image pixels.

Furthermore, in managing disaster-prone or priority monitoring zones, risk-based decision-making approaches are necessary. Gohil et al. (2024) demonstrated that fuzzy logic can be used to build multi-hazard models based on environmental parameters such as rainfall, elevation, and land use. Fuzzy logic mimics human reasoning by accommodating uncertainty and ambiguity; rather than rigid binary categories (true/false), fuzzy values can range continuously (e.g., "moderately high" or "slightly vulnerable"). These fuzzy values are then classified into different vulnerability levels, enabling PTZ cameras to automatically prioritize their monitoring focus on high-risk zones. Such spatial integration has also been validated in field implementations, as shown by Whitehurst et al. (2021), who employed drones to rapidly map disaster-affected areas, including assessing building damage and flood potential. These data were subsequently integrated into Geographic Information System (GIS) platforms to support layered monitoring. In PTZ camera contexts, drone mapping results can be directly linked to camera control systems, expanding surveillance coverage based on spatial conditions identified from aerial data. The integration of 3D shape understanding, fuzzy risk assessment, and drone data opens the path for PTZ cameras to act not only as "observers" but as "interpreters" of space, risk, and scenarios. Such systems evolve beyond passive monitoring into spatially intelligent decision-making networks.

In today's intelligent surveillance ecosystem, PTZ cameras no longer function as static monitoring devices but as active components within distributed networks known as the Internet of Robotic Things (IoRT). IoRT is an integrative concept combining sensors, robots, monitoring devices, and cloud- or edge-based control systems to create responsive and adaptive environments. Within such systems, PTZ cameras serve as "eyes" that not only see but also understand and react to the environment based on collective spatial and visual inputs. According to Andronie et al. (2023), the success of IoRT in surveillance depends on components such as spatial big data management, sensor fusion, and deep learning algorithms for object and event detection. Sensor fusion allows the system to unify data from various devices—for example, UAV imagery, motion detectors, and fixed cameras—to

provide a comprehensive understanding of spatial situations. In these systems, PTZ cameras can autonomously adjust direction and zoom based on information from other nodes, such as when a UAV detects suspicious movement in areas beyond fixed camera coverage.

To ensure responsiveness to extraordinary events, anomaly detection capabilities are required. Visual anomalies refer to behavioral patterns or events in video that deviate from the norm, such as sudden fights, fleeing, or contextually inappropriate activities. Ullah et al. (2021) developed an anomaly detection framework based on MobileNetV2 and residual LSTM augmented with attention mechanisms. This combination efficiently recognizes temporal patterns in video with low inference time and high accuracy. Tests using public video datasets (e.g., UCF-Crime and UMN) showed this system outperforms conventional anomaly detection models in speed and precision. For edge-based implementations, as needed in PTZ cameras, this model is ideal since it does not require high-power hardware to operate. This means PTZ cameras installed at strategic locations can locally process captured video to detect anomalies, then send alerts or signals to control centers or other IoRT nodes. Responses may include redirecting other cameras to the incident location or activating public warning systems. Thus, PTZ cameras within IoRT frameworks become not just observation tools but integral parts of collaborative intelligent systems combining spatial processing, machine learning, and inter-device communication. This creates surveillance systems that are not only reactive but also predictive and context-aware.

4. Conclusions

The utilization of deep learning technology in pan-tilt-zoom (PTZ) camera control systems marks a new chapter in the development of intelligent geospatial surveillance systems. As threats to critical infrastructure and public safety become increasingly complex, the integration of artificial intelligence vision (AI vision), PTZ cameras, and spatial data serves as a vital foundation for building visual intelligence systems capable of detecting, analyzing, and responding to situations in real time. PTZ cameras, with their capabilities for horizontal rotation (pan), vertical movement (tilt), and optical zoom, serve as central components for monitoring wide areas with a single device. However, their effectiveness depends on precise hardware calibration, accurate image processing, and automated machine learning-based control mechanisms. The implementation of a dual-mode system, which combines fisheye lenses for wide coverage with PTZ cameras for focused observation, significantly enhances object detection accuracy.

Within this system, deep learning plays a pivotal role in visual analysis. Convolutional neural networks (CNN) are employed to extract visual features hierarchically and adaptively, reducing reliance on manual feature engineering. Lightweight architectures such as MobileNetV2 enable efficient image processing without compromising accuracy. Moreover, approaches based on spatial grids and foveatic classifiers accelerate object detection by focusing on salient areas within the image. The integration of AI-based vision systems with spatial data enriches the operational context of PTZ cameras. Platforms like google earth engine (GEE) support large-scale spatial analysis, while augmented reality geographic information systems (AR-GIS) enhance real-world camera control accuracy. Segmentation based on knowledge-enhanced neural networks (KENN) applied to point cloud data enables the interpretation of complex spatial structures, improving spatially-informed surveillance capabilities.

In dynamic surveillance environments, the ability to recognize human activity and detect visual anomalies is crucial. The combination of Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks and attention mechanisms allows systems to understand complex behavioral patterns, such as aggressive actions, escape attempts, or other abnormal activities. This integration transforms PTZ cameras into active surveillance agents capable of responding rapidly and adaptively to situational changes. Further development of the internet of robotic things (IoRT) expands the scope of surveillance systems by positioning PTZ cameras as part of a collaborative network of

intelligent sensors. In this ecosystem, PTZ cameras coordinate with unmanned aerial vehicles (UAVs), patrol robots, and fixed sensors, collectively forming a data-driven, adaptive monitoring system. The use of sensor fusion techniques enhances situational awareness and improves decision-making accuracy.

Nevertheless, challenges such as latency in remote control, cybersickness effects from virtual reality (VR) integration, and the need to validate deep learning models against real-world visual conditions demand serious consideration. Therefore, system development must focus not only on technical reliability but also on the adaptability to operational dynamics in the field. Overall, the integration of deep learning, PTZ cameras, and geospatial data presents significant potential for building next-generation surveillance systems that are proactive, highly precise, responsive, and spatially analytical. These systems not only improve monitoring effectiveness in public spaces, industrial zones, and defense facilities but also strengthen the development of spatial intelligence—an increasingly critical component of national security in the evolving digital era.

Acknowledgement
The authors would like to express their sincere gratitude to all parties who contributed to the completion of this research.

Author Contribution
All authors contributed equally to the conception, design, analysis, and writing of this manuscript.

Funding
This research received no external funding.

Ethical Review Board Statement
Not available.

Informed Consent Statement
Not available.

Data Availability Statement
Not available.

Conflicts of Interest
The authors declare no conflict of interest.

References
Al-Yadumi, S., Xion, T. E., Wei, S. G. W., & Boursier, P. (2021). Review on integrating geospatial big datasets and open research issues. *IEEE Access, 9*, 10604–10620. https://doi.org/10.1109/ACCESS.2021.3051084

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data, 8*(1). https://doi.org/10.1186/s40537-021-00444-8

Andronie, M., Lăzăroiu, G., Iatagan, M., Hurloiu, I., Ştefănescu, R., Dijmărescu, A., & Dijmărescu, I. (2023). Big data management algorithms, deep learning-based object detection technologies, and geospatial simulation and sensor fusion tools in the Internet of Robotic Things. *ISPRS International Journal of Geo-Information, 12*(2). https://doi.org/10.3390/ijgi12020035

Arroyo, S., Garcia, L., Safar, F., & Oliva, D. (2021). Urban dual mode video detection system based on fisheye and PTZ cameras. *IEEE Latin America Transactions*, *19*(9), 1537-1545. https://doi.org/10.1109/TLA.2021.9468447

Arshad, M. H., Bilal, M., & Gani, A. (2022). Human activity recognition: Review, taxonomy and open challenges. *Sensors, 22*(17). https://doi.org/10.3390/s22176463

Bazargani, J.S., Zafari, M., Sadeghi-Niaraki, A., & Choi, S. M. (2022). A survey of GIS and AR integration: Applications. *Sustainability (Switzerland), 14*(16). https://doi.org/10.3390/su141610134

Choi, Y. (2023). GeoAI: Integration of artificial intelligence, machine learning, and deep learning with GIS. *Applied Sciences, 13*(6). https://doi.org/10.3390/app13063895

Church, J. S., Gegout, M., & Adams, P. J. (2024). Effectiveness of optical, digital, and hybrid zoom equipped drones for use in reading livestock ear tags for individual animal identification. *Drone Systems and Applications, 12*, 1–7. https://doi.org/10.1139/dsa-2023-0041

Fahim, A., Papalexakis, E., Krishnamurthy, S. V., Roy Chowdhury, A. K., Kaplan, L., & Abdelzaher, T. (2023). AcTrak: Controlling a steerable surveillance camera using reinforcement learning. *ACM Transactions on Cyber-Physical Systems, 7*(2). https://doi.org/10.1145/3585316

Fan, J., Lu, R., Yang, X., Gao, F., Li, Q., & Zeng, J. (2021). Design and implementation of intelligent EOD system based on six-rotor UAV. *Drones, 5*(4). https://doi.org/10.3390/drones5040146

Fathy, C., & Saleh, S. N. (2022). Integrating deep learning-based IoT and fog computing with software-defined networking for detecting weapons in video surveillance systems. *Sensors, 22*(14). https://doi.org/10.3390/s22145075

Fuertes, D., del-Blanco, C. R., Carballeira, P., Jaureguizar, F., & García, N. (2022). People detection with omnidirectional cameras using a spatial grid of deep learning foveatic classifiers. *Digital Signal Processing, 126*. https://doi.org/10.1016/j.dsp.2022.103473

García-Segura, T., Montalbán-Domingo, L., Llopis-Castelló, D., Sanz-Benlloch, A., & Pellicer, E. (2023). Integration of deep learning techniques and sustainability-based concepts into an urban pavement management system. *Expert Systems with Applications, 231*. https://doi.org/10.1016/j.eswa.2023.120851

Giordano, J., Lazzaretto, M., Michieletto, G., & Cenedese, A. (2022). Visual sensor networks for indoor real-time surveillance and tracking of multiple targets. *Sensors, 22*(7). https://doi.org/10.3390/s22072661

Gohil, M., Mehta, D., & Shaikh, M. (2024). An integration of geospatial and fuzzy-logic techniques for multi-hazard mapping. *Results in Engineering, 21*. https://doi.org/10.1016/j.rineng.2024.101758

Grilli, E., Daniele, A., Bassier, M., Remondino, F., & Serafini, L. (2023). Knowledge enhanced neural networks for point cloud semantic segmentation. *Remote Sensing, 15*(10). https://doi.org/10.3390/rs15102590

Huillca, J. L., & Fernandes, L. A. F. (2022). Using conventional cameras as sensors for estimating confidence intervals for the speed of vessels from single images. *Sensors, 22*(11). https://doi.org/10.3390/s22114213

Jing, B., & Xue, H. (2024). IoT fog computing optimization method based on improved convolutional neural network. *IEEE Access, 12*. https://doi.org/10.1109/ACCESS.2023.3348133

Junos, M. H., Mohd Khairuddin, A. S., & Dahari, M. (2022). Automated object detection on aerial images for limited capacity embedded device using a lightweight CNN model. *Alexandria Engineering Journal, 61*(8). https://doi.org/10.1016/j.aej.2021.11.027

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing, 5*. https://doi.org/10.1016/j.ophoto.2022.100018

Khan, S., Khan, M. A., Alhaisoni, M., Tariq, U., Yong, H. S., Armghan, A., & Alenezi, F. (2021). Human action recognition: A paradigm of best deep learning features selection and serial based extended fusion. *Sensors, 21*(23). https://doi.org/10.3390/s21237941

Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation, 112*. https://doi.org/10.1016/j.jag.2022.102926

Llauradó, J. M., Pujol, F. A., Tomás, D., Visvizi, A., & Pujol, M. (2023). Study of image sensors for enhanced face recognition at a distance in the smart city context. *Scientific Reports, 13*(1). https://doi.org/10.1038/s41598-023-40110-y

Lou, H., Duan, X., Guo, J., Liu, H., Gu, J., Bi, L., & Chen, H. (2023). DC-YOLOv8: Small-size object detection algorithm based on camera sensor. *Electronics (Switzerland), 12*(10). https://doi.org/10.3390/electronics12102323

Mao, K., Xu, Y., Wang, R., & Pan, S. (2022). A general calibration method for dual-PTZ cameras based on feedback parameters. *Applied Sciences (Switzerland), 12*(18). https://doi.org/10.3390/app12189148

Nepal, U., & Eslamiat, H. (2022). Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. *Sensors, 22*(2). https://doi.org/10.3390/s22020464

Puttinaovarat, S., & Horkaew, P. (2022). A geospatial platform for crowdsourcing green space area management using GIS and deep learning classification. *ISPRS International Journal of Geo-Information, 11*(3). https://doi.org/10.3390/ijgi11030208

Ren, S., Malof, J., Fetter, R., Beach, R., Rineer, J., & Bradbury, K. (2022). Utilizing geospatial data for assessing energy security: Mapping small solar home systems using unmanned aerial vehicles and deep learning. *ISPRS International Journal of Geo-Information, 11*(4). https://doi.org/10.3390/ijgi11040222

Reuschling, F., & Jakobi, J. (2022). Remote AFIS: Development and validation of low-cost remote tower concepts for uncontrolled aerodromes. *CEAS Aeronautical Journal, 13*(4), 1067–1083. https://doi.org/10.1007/s13272-022-00613-2

Shi, Q., Chen, Y., & Liang, H. (2023). Real-time risk assessment of road vehicles based on inverse perspective mapping. *Array, 20*. https://doi.org/10.1016/j.array.2023.100325

Shivappriya, S. N., Priyadarsini, M. J. P., Stateczny, A., Puttamadappa, C., & Parameshachari, B. D. (2021). Cascade object detection and remote sensing object detection method based on trainable activation function. *Remote Sensing, 13*(2). https://doi.org/10.3390/rs13020200

Sun, Y. T. A., Lin, H. C., Wu, P. Y., & Huang, J. T. (2022). Learning by watching via keypoint extraction and imitation learning. *Machines, 10*(11). https://doi.org/10.3390/machines10111049

Ullah, W., Ullah, A., Hussain, T., Khan, Z. A., & Baik, S. W. (2021). An efficient anomaly recognition framework using an attention residual LSTM in surveillance videos. *Sensors, 21*(8). https://doi.org/10.3390/s21082811

Vijeikis, R., Raudonis, V., & Dervinis, G. (2022). Efficient violence detection in surveillance. *Sensors, 22*(6). https://doi.org/10.3390/s22062216

Villegas-Ch, W., & García-Ortiz, J. (2023). Authentication, access, and monitoring system for critical areas with the use of artificial intelligence integrated into perimeter security in a data center. *Frontiers in Big Data, 6*. https://doi.org/10.3389/fdata.2023.1200390

Whitehurst, D., Friedman, B., Kochersberger, K., Sridhar, V., & Weeks, J. (2021). Drone-based community assessment, planning, and disaster risk management for sustainable development. *Remote Sensing, 13*(9). https://doi.org/10.3390/rs13091739

Yang, L., Driscol, J., Sarigai, S., Wu, Q., Chen, H., & Lippitt, C. D. (2022). Google Earth Engine and artificial intelligence (AI): A comprehensive review. *Remote Sensing, 14*(14). https://doi.org/10.3390/rs14143253

You, J., Hu, Z., Xiao, H., & Xu, C. (2022). Cell-based target localization and tracking with an active camera. *Applied Sciences (Switzerland), 12*(6). https://doi.org/10.3390/app12062771

Zhai, H., & Zhang, Y. (2022). Target detection of low-altitude UAV based on improved YOLOv3 network. *Journal of Robotics, 2022*. https://doi.org/10.1155/2022/4065734

Zheng, J., Mao, S., Wu, Z., Kong, P., & Qiang, H. (2022). Improved path planning for indoor patrol robot based on deep reinforcement learning. *Symmetry, 14*(1). https://doi.org/10.3390/sym14010132

Biographies of Authors

Farhat Bashir, Sekolah Tinggi Intelijen Negara, Bogor, West Java 16810, Indonesia.
- Email: farhatbashir1093@gmail.com
- ORCID: N/A
- Web of Science ResearcherID: N/A
- Scopus Author ID: N/A
- Homepage: N/A

Syachrul Arief, Geospatial Information Agency, Bogor, West Java 16911, Indonesia.
- Email: syachrul.arief@big.go.id
- ORCID: 0000-0002-1839-6301
- Web of Science ResearcherID: N/A
- Scopus Author ID: 57522236500
- Homepage: https://scholar.google.com/citations?user=ubcIvs4AAAAJ&hl=en