Artificial Intelligence Review (2026) 59:210
https://doi.org/10.1007/s10462-026-11610-w

Advancing integrated sensing and communications with RL/DRL in 6 G networks: a survey

Wali Ullah Khan[1] · Waqas Khalid[2] · Muhammad Iqbal[3] · Chiew Foong Kwong[2] · Manzoor Ahmed[4] · Syed Tariq Shah[5]

Received: 3 October 2025 / Accepted: 28 May 2026

Abstract
Integrated sensing and communication (ISAC) has emerged as a key enabling technology for sixth-generation (6 G) networks, supporting joint environment perception and data transmission with high spectral efficiency and reduced hardware cost. However, the design and deployment of ISAC systems remain challenging due to dynamic wireless environments, sensing–communication trade-offs, and the increasing complexity of large-scale networks. Reinforcement learning (RL) and deep reinforcement learning (DRL) have recently attracted significant attention as data-driven approaches for addressing these challenges through model-free, adaptive, and real-time optimization. This survey first reviews the fundamentals of ISAC technology and then summarizes major RL and DRL algorithms relevant to ISAC design. It further provides a comprehensive overview of RL/DRL methods for ISAC in 6 G networks. Existing approaches are categorized into value-based, policy-based, and hybrid methods, and are further classified according to representative ISAC scenarios, including reconfigurable intelligent surface (RIS)-assisted systems, unmanned aerial vehicle (UAV) and satellite systems, vehicular networks, and other emerging 6 G applications. The survey highlights reported gains in spectral efficiency, sensing accuracy, adaptability, and robustness, while also identifying key limitations of current approaches. Finally, the survey outlines current challenges, future research directions, and key lessons learned.

Keywords 6 G · ISAC · RL/DRL · RIS-assisted systems · UAV and satellite communications · Vehicular networks

1 Introduction

The next generation of wireless communication networks, namely sixth-generation (6 G) networks, is expected to deliver unprecedented improvements in data rate, latency, reliability, connectivity, and intelligence (Zhu et al. 2025; Ahmed et al. 2025). 6 G is envisioned to support a broad range of emerging applications across terrestrial networks (TNs) and non-terrestrial networks (NTNs), including autonomous vehicles, extended reality (XR),

Extended author information available on the last page of the article

Published online: 14 June 2026

smart healthcare, smart cities, the industrial Internet of Things (IIoT), smart agriculture, and unmanned aerial vehicles (UAVs) (Cui et al. 2025; Khan et al. 2021). These applications require not only ultra-high-speed communications but also accurate and real-time awareness of the surrounding environment (Drampalou et al. 2024; Khan et al. 2019). To meet these stringent requirements, future 6 G systems are expected to tightly integrate sensing and communication functionalities. Such integration can enable simultaneous environment perception and data transmission while improving spectrum utilization, reducing hardware cost, and enhancing overall network intelligence (Ding et al. 2026; Khan et al. 2022).

Beyond introducing new services, 6 G is also expected to evolve toward a native "network-of-systems" paradigm, in which different terrestrial and non-terrestrial platforms, such as cellular base stations, UAVs, high-altitude platform stations (HAPS), and satellites, cooperate to provide pervasive coverage and context-aware connectivity Okati et al. (2026). This vision relies on the close integration of communication, sensing, and computation, supported by intelligent and adaptive network components, dynamic spectrum sharing, and artificial intelligence (AI)-driven control. However, such an architecture also introduces substantial complexity due to time-varying channels, heterogeneous quality-of-service (QoS) requirements, user mobility, and strong coupling across communication, sensing, and computing resources Khan et al. (2023). Consequently, data-driven and adaptive decision-making frameworks are becoming increasingly important for managing this complexity. In this context, reinforcement learning (RL) and deep reinforcement learning (DRL) have emerged as promising tools for real-time orchestration in integrated sensing and communication (ISAC) networks and related integrated terrestrial–non-terrestrial intelligent systems involving computation offloading and latency-aware optimization (Li et al. 2023; Pei et al. 2026).

ISAC has recently attracted significant attention as a promising enabling technology for future 6 G TN and NTN systems (Lu et al. 2024; Li et al. 2025; Sheemar et al. 2025). Unlike conventional communication frameworks, ISAC waveforms are designed to support both information transmission and environmental sensing simultaneously He et al. (2024). This integrated design improves spectrum and hardware utilization and can enhance a wide range of emerging applications, including autonomous driving, UAV operations, smart cities, and industrial automation (González-Prelcic et al. 2024; Strinati et al. 2025; Jiang et al. 2025). In such systems, sensing is treated as an inherent network capability rather than as an auxiliary function layered on top of communication services (Magbool et al. 2025; Khan et al. 2021).

Despite its considerable potential, ISAC still faces several important technical challenges. Wireless channels in practical environments vary continuously due to factors such as mobility, multipath fading, and Doppler effects (Sheemar et al. 2025; Wen et al. 2024). These dynamics make the design of ISAC systems significantly more challenging, particularly for resource allocation, beamforming, and waveform optimization Liu et al. (2025). To address these issues, conventional optimization approaches, including convex optimization, heuristic methods, and iterative algorithms, have been widely investigated (Song et al. 2025; Tishchenko et al. 2025; Khan et al. 2021). However, such methods often rely on accurate channel state information (CSI) and can incur high computational complexity, which limits their applicability in large-scale or rapidly time-varying networks (Allu et al. 2025; Khan et al. 2023, 2021).

RL and DRL have emerged as promising approaches for ISAC design and optimization Li et al. (2025). RL provides a decision-making framework in which an agent interacts with the environment, learns from feedback, and optimizes long-term objectives without requiring an explicit model of the system dynamics (Luong et al. 2025; Kaleem et al. 2024). DRL, which combines RL with deep neural networks (NNs), can handle high-dimensional state and action spaces and is therefore well suited to complex ISAC scenarios (Khan et al. 2025; Sheemar et al. 2025). By leveraging RL/DRL, ISAC systems can adapt to dynamic environments, jointly optimize communication and sensing performance, and support real-time decision-making (Gong et al. 2025; Khan et al. 2022). Fig. 1 illustrates representative RL/DRL-enabled ISAC scenarios, including vehicular networks, UAV-assisted ISAC, reconfigurable intelligent surface (RIS)/beyond-diagonal RIS (BD-RIS) architectures, and multiple-input multiple-output (MIMO)-based deployments. As shown, RL/DRL acts as an intelligent control engine for managing communication and sensing resources in dynamic wireless environments.

Several factors motivate the adoption of RL/DRL in ISAC systems. First, resource allocation in ISAC is inherently a multi-objective optimization problem, requiring a careful balance between sensing accuracy and communication efficiency. Second, practical ISAC systems often involve multiple antennas, RISs, and UAVs, leading to high-dimensional control spaces that are difficult to address using conventional optimization techniques. Third, real-time adaptation is essential in mobile and time-varying environments, where transmit power, beam directions, and sensing schedules must be continuously adjusted. In such settings, RL/DRL offers an attractive framework by enabling model-free learning and sequential decision-making under uncertainty.

Table 1 summarizes the main abbreviations used in this paper.

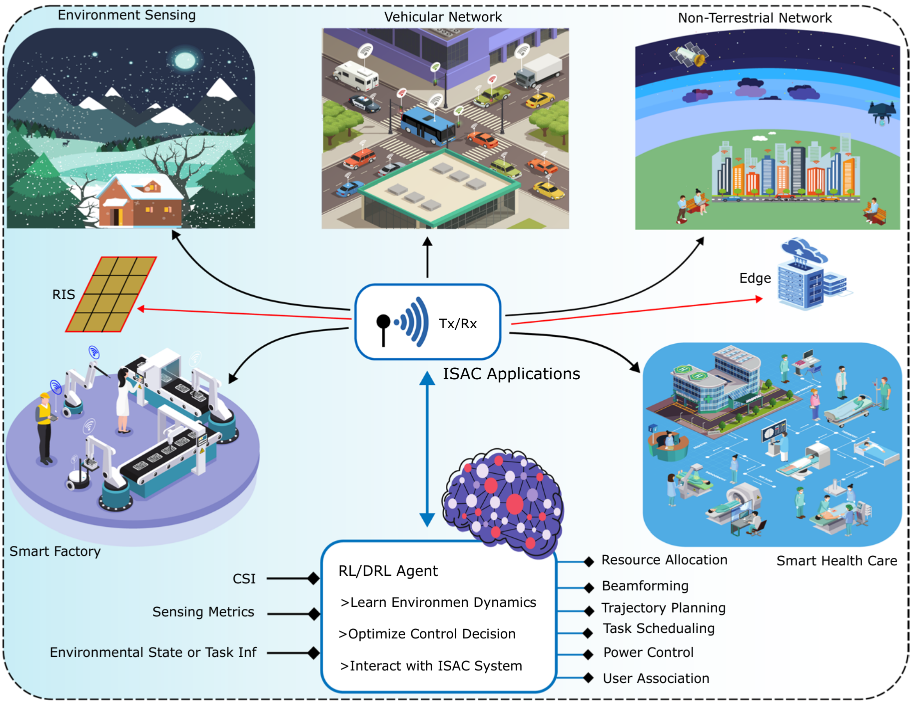


Fig. 1 Potential applications/use cases of RL/DRL algorithms in ISAC systems

Table 1 List of Abbreviations

Abbr.	Definition	Abbr.	Definition
6 G	Sixth-generation wireless networks	A2C	Advantage actor–critic
A3C	Asynchronous advantage actor–critic	AC	Actor–critic
ADMM	Alternating direction method of multipliers	AI	Artificial intelligence
AirComp	Over-the-air computation	AoA	Angle of arrival
ASR	Achievable sum rate	AV	Autonomous vehicle
BD-RIS	Beyond-diagonal reconfigurable intelligent surface	B&B	Branch-and-bound
C-RAN	Cloud radio access network	CR	Cognitive radio
CRB	Cramér–Rao bound	CRLB	Cramér–Rao lower bound
CSI	Channel state information	DDPG	Deep deterministic policy gradient
DDQN	Double deep Q-network	DFRC	Dual-functional radar–communication
DL	Deep learning	DP	Dynamic programming
DRL	Deep reinforcement learning	DQN	Deep Q-network
D-RIS	Diagonal reconfigurable intelligent surface	EE	Energy efficiency
ESN	Echo state network	FDD	Frequency-division duplexing
FEEL	Federated edge learning	FIM	Fisher information matrix
FL	Federated learning	FMCW	Frequency-modulated continuous-wave
HAPS	High altitude platform station	HDRL	Hierarchical deep reinforcement learning
HTD3	Hierarchical twin delayed DDPG	IBTD	DRL-based ISAC beam tracking design
IIoT	Industrial Internet of Things	IoT	Internet of Things
IoV	Internet of Vehicles	ISAC	Integrated sensing and communication
ISCC	Integrated sensing, communication, and computation	JCAS	Joint communication and sensing
KL	Kullback–Leibler divergence	LoS	Line-of-sight
MADDPG	Multi-agent deep deterministic policy gradient	MADRL	Multi-agent deep reinforcement learning
MAE	Mean absolute error	MAPPO	Multi-agent proximal policy optimization
MARL	Multi-agent reinforcement learning	MC	Monte Carlo
MEC	Mobile edge computing	MDP	Markov decision process
MIMO	Multiple-input multiple-output	ML	Machine learning
mmWave	Millimeter wave	MSE	Mean squared error
NLoS	Non-line-of-sight	NSP	Near-field steering processing
NTN	Non-terrestrial network	OFDM	Orthogonal frequency-division multiplexing
PI	Policy iteration	POMDP	Partially observable Markov decision process
PPO	Proximal policy optimization	PRR	Packet reception ratio
QMIX	Q-value mixing network	QoE	Quality of experience
QoS	Quality of service	RCS	Radar cross section
RIS	Reconfigurable intelligent surface	RL	Reinforcement learning
RMSE	Root mean squared error	ROC	Receiver operating characteristic
RSMA	Rate-splitting multiple access	RSU	Roadside unit

Table 1 (continued)

Abbr.	Definition	Abbr.	Definition
SAC	Soft actor–critic	SARSA	State–action–reward–state–action
SCS	Subcarrier spacing	SE	Spectral efficiency
SINR	Signal-to-interference-plus-noise ratio	SLAC	Simultaneous localization and communication
SNR	Signal-to-noise ratio	SPEB	Squared position error bound
SPD	Staged policy distillation	SPR	Staged policy reuse
TD	Temporal difference	TD3	Twin delayed deep deterministic policy gradient
TDD	Time-division duplexing	THz	Terahertz
UAV	Unmanned aerial vehicle	V2I	Vehicle-to-infrastructure
V2P	Vehicle-to-pedestrian	V2V	Vehicle-to-vehicle
V2X	Vehicle-to-everything	VDN	Value-decomposition network
VI	Value iteration	NOMA	Non-orthogonal multiple access

1.1 Previous surveys and tutorials

Several recent studies have reviewed ISAC from different perspectives in the context of future 6 G networks. For example, Luong et al. (2021) examines recent developments in ISAC systems with a particular focus on resource management strategies. It introduces the fundamental concepts, performance metrics, and applications of ISAC, highlights the benefits of ISAC in terms of spectrum efficiency, energy efficiency, system compactness, and cost reduction, and discusses a range of resource management approaches. The study also identifies key design challenges and outlines potential future research directions. Reference Zhang et al. (2022) reviews the evolution from coexisting communication and radar systems to joint communication and sensing (JCAS), presenting system architectures, sensing operations, and required infrastructure changes, while highlighting the differences among various JCAS designs. It also discusses important research topics, including waveform optimization, antenna design, clutter suppression, sensing parameter estimation, and sensing-assisted communications, and concludes with open challenges and lessons learned. In Zhou et al. (2022), the authors survey ISAC as a key enabling technology for 6 G systems capable of supporting sensing and communication on a common hardware platform. That study reviews the motivations, applications, and challenges of ISAC and provides a detailed overview of waveform design approaches, classifying them into communication-centric, sensing-centric, and joint optimization categories.

In addition, Wei et al. (2023) analyzes ISAC signal design in 6 G systems and focuses on its impact on both sensing and communication performance. It discusses signal design, signal processing techniques, and signal optimization methods, and provides useful directions for future ISAC signal design. The work in Wei et al. (2025) presents a review of ISAC channel modeling techniques and emphasizes their importance for evaluating and improving radar sensing performance in 6 G systems. It summarizes methods for modeling target radar cross section (RCS) and clutter RCS, and discusses how these models affect both sensing and communication channels. The study in Jiang et al. (2024) provides a detailed review of terahertz (THz)-enabled ISAC for 6 G networks, covering motivations, applications, historical development, propagation characteristics, channel modeling, transceiver architectures, antennas, and beamforming design, and concludes with open challenges and future directions.

The survey in Ade Krisna Respati and Lee (2024) discusses how machine learning (ML) and deep learning (DL) techniques can be applied to 6 G ISAC systems to enhance sensing performance and improve the joint use of communication and sensing resources. It reviews different ISAC configurations, practical applications, and commonly used ML/DL models, and identifies several promising research directions for ML-driven ISAC. The study in Wen et al. (2024) investigates integrated sensing, communication, and computation (ISCC) as an important step toward future 6 G systems. It reviews the evolution of ISCC, discusses current limitations, and examines signal design, resource management, and system operation for integrated functionality. The work in González-Prelcic et al. (2024) presents a future-oriented vision of 6 G-ISAC, where sensing capabilities embedded in infrastructure, user equipment, and communication signals enhance network resilience and adaptability. It also provides insights into signal processing, optimization, and ML methods for enabling ISAC in 6 G systems, supported by simulation results that combine ray-tracing measurements with mathematical models.

More recently, Zhu et al. (2025) presents an ISAC-oriented unified IoT architecture for 6 G networks that integrates software-defined communication with AI-enabled agents for intelligent connectivity. It reviews the evolution of ISAC over the past decade, categorizes prior work across hardware, architectural, and application layers, and highlights the importance of AI and multimodal design. The study also identifies major technical challenges, security concerns, and future research opportunities for intelligent and secure ISAC-enabled IoT systems. In addition, Luong et al. (2025) reviews the use of learning algorithms in ISAC systems and emphasizes their potential to address resource management problems and environmental dynamics more effectively than traditional optimization methods. It surveys learning-based approaches for different ISAC configurations, summarizes recent advances in learning for wireless sensing, and outlines key challenges and future directions for learning-enabled ISAC.

As shown in Table 2, most existing ISAC surveys focus on waveform design, signal processing, resource management, or general ML-enabled ISAC. In contrast, RL/DRL is often treated only as one component within a broader discussion. This survey specifically focuses on RL/DRL for ISAC and provides a unified taxonomy that connects learning paradigms, including value-based, policy-based, actor–critic, and multi-agent methods, with ISAC tasks and deployment scenarios. It also highlights key performance metrics, technical trends, and future research opportunities.

1.2 Motivation and main contributions

Although ISAC has advanced rapidly in recent years, practical deployments still require repeated decision-making under uncertainty, joint optimization of sensing and communication objectives, and rapid adaptation to environmental dynamics such as mobility, blockage, and interference. These characteristics align naturally with RL/DRL, which can learn control policies through interaction when accurate system models are unavailable or when online optimization is required with manageable complexity. Nevertheless, existing RL/DRL-based ISAC research remains fragmented across different problem settings, learning paradigms, and evaluation criteria, making it difficult to compare methods fairly and

Table 2 Comparison of existing surveys and tutorials on ISAC and related technologies

Reference	Year	Main Focus	Key Technologies/Scenarios	RL/DRL Coverage
Luong et al. (2021)	2021	ISAC resource management	Resource allocation, performance metrics, applications	Limited
Zhang et al. (2022)	2022	Evolution to JCAS	Coexistence to joint design, waveform optimization, antenna design	Not covered
Zhou et al. (2022)	2022	ISAC waveform design	Waveform categorization: communication-centric, sensing-centric, joint	Limited
Wei et al. (2023)	2023	ISAC signal design	Signal processing, signal optimization	Not covered
Wei et al. (2025)	2025	ISAC channel modeling	Channel modeling for sensing, RCS, clutter	Not covered
Jiang et al. (2024)	2024	Terahertz ISAC	THz propagation, channel modeling, transceiver design	Not covered
Ade Krisna Respati and Lee (2024)	2024	ML/DL for ISAC	ML/DL models, system configurations, use cases	Limited RL
Wen et al. (2024)	2024	Integrated sensing, communication, computation	Signal design, resource management, system functionalities	Limited
González-Prelcic et al. (2024)	2024	ISAC vision for 6 G	Network resilience, signal processing, optimization	Limited ML
Zhu et al. (2025)	2025	ISAC-oriented IoT architecture	AI-enabled agents, multi-modal design, security	Limited RL
Luong et al. (2025)	2025	Learning algorithms for ISAC	Resource management, dynamic environments, wireless sensing	General learning focus (not RL/DRL-specific)
This survey	2026	RL/DRL for ISAC in 6 G	Value-based, policy-based, and actor–critic RL; RIS, UAV, vehicular, and other 6 G scenarios	Comprehensive: DQN, DDQN, A2C, DDPG, SAC, PPO, MARL, etc.

to identify the conditions under which learning-based approaches are most effective. This motivates a focused survey that systematically categorizes RL/DRL-based ISAC designs, clarifies the commonly used objectives, metrics, and assumptions, and highlights reproducible insights as well as open research problems.

The main objective of this survey is to review state-of-the-art RL/DRL algorithms for ISAC, classify them according to their learning paradigms and deployment scenarios, and discuss their strengths and limitations. Specifically, the survey covers model-free, model-based, value-based, policy-based, and hybrid RL approaches. It examines a range of representative scenarios, including vehicular networks, UAV-assisted ISAC, MIMO-ISAC systems, and RIS/BD-RIS-assisted architectures. The reviewed studies are analyzed using performance indicators such as achievable sum rate, sensing accuracy, latency, energy efficiency, and reliability. While the primary focus is on RL/DRL-based solutions, conventional optimization methods are also briefly discussed to provide context and to highlight the comparative advantages and limitations of learning-based techniques. The main contributions of this survey are summarized as follows:

- A comprehensive taxonomy of RL/DRL-based approaches for ISAC is provided, highlighting the main differences among model-free, model-based, value-based, policy-based, and hybrid learning methods.

- RL/DRL techniques are critically reviewed across multiple ISAC scenarios, including vehicular networks, UAV-assisted systems, MIMO-ISAC, and RIS/BD-RIS-enabled architectures, with emphasis on performance gains and practical implications.
- Open research challenges and promising future directions are identified, including scalable algorithm design, multi-agent learning, robust operation under uncertainty, and the integration of advanced sensing modalities.
- Key lessons learned from the existing literature are summarized to offer guidance for the design and development of effective RL/DRL-enabled ISAC systems for future 6 G networks.

The remainder of this paper is organized as follows. Section 2 introduces the fundamentals of ISAC technology. Section 3 provides an overview of RL/DRL algorithms for ISAC. Section 4 reviews recent RL/DRL-based ISAC methods according to their learning frameworks and application scenarios. Section 5 discusses current challenges, future research directions, and lessons learned. Finally, Sect. 6 concludes the paper. The overall structure of the survey is illustrated in Fig. 2.

2 Fundamentals of ISAC technology

This section provides an overview of the main idea of ISAC. It examines key aspects, including sensing activities, radar classifications, levels of ISAC integration, joint waveform design, and ongoing efforts to standardize these areas.

2.1 Sensing tasks

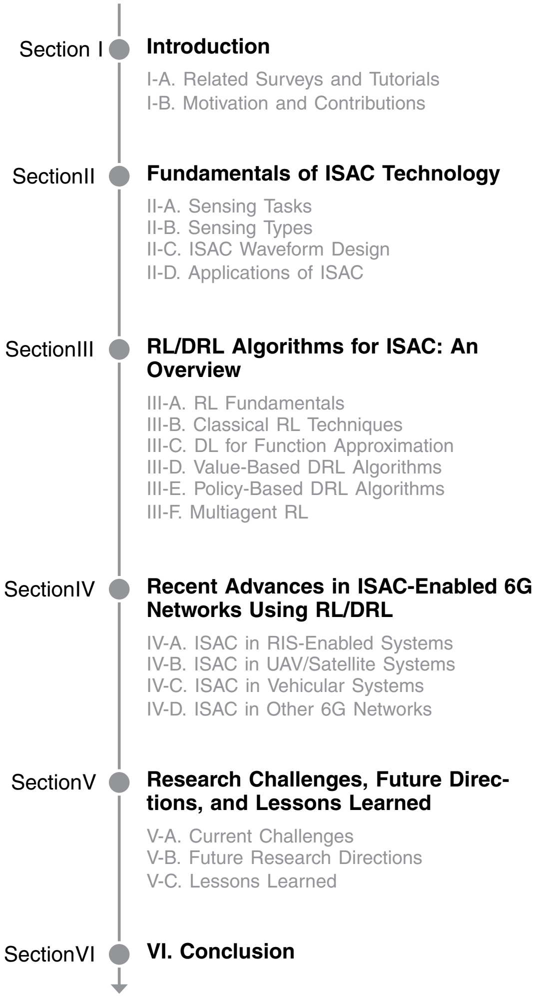
Radar sensing involves transmitting a probing signal and analyzing the returned echoes to infer information about one or more objects. Active localisation (positioning) uses active devices that send and receive signals, while classical radar detection looks at passive objects that just reflect the incoming waveform. In real life, sensing activities in ISAC systems are usually split into three primary groups: parameter estimation, detection, and recognition.

2.1.1 Estimation

Parameter estimate seeks to ascertain unknown variables that define the propagation environment or the target, such range, radial velocity (Doppler effect), angle of arrival/departure, radar cross-section, or reflectivity. In numerous studies, "radar sensing" and "parameter estimation" are regarded as almost synonymous within the framework of localisation and tracking Zhang et al. (2021).

There are a number of ways to measure how well an estimate works. The mean squared error (MSE) is the expected squared difference between the true parameter and its estimate. It is a way to quantify overall accuracy Liu et al. (2022). The Cramér Rao bound (CRB) is a common theoretical lower bound on the variance of any unbiased estimator because it is often hard to get the exact MSE in closed form Liu et al. (2022). Other important numbers are the estimator's bias and variance, the root-MSE (RMSE), and Bayesian bounds like the

Fig. 2 Structure of this survey



Ziv Zakai bound, which show how well the estimator works in situations with low signal-to-noise ratios (SNR) or limited prior knowledge.

2.1.2 Detection

The second sensing task is state detection, which determines which hypothesis about the target is true based on the received echoes. Common instances include determining the presence or absence of a target (binary detection) or categorising a target into one of several distinct states, such as a range or velocity bin Liu et al. (2022).

Several probabilities are often used to describe how well a detection system works. The detection probability tells you how frequently the system correctly says there is a target when there is one, and the false-alarm probability tells you how often the detector wrongly says there is a target when there isn't one Yao et al. (2025). The likelihood of missing a detection is the opposite of the probability of detecting something. Receiver operating characteristic (ROC) curves are a common way to show these numbers. They show how the balance between detection and false alarms changes when the decision threshold increases. In this survey, these probabilities are referred to as performance metrics, while the term "indicator" denotes a specific scalar value of a given metric at a particular operating point.

2.1.3 Recognition

The last sensing task is recognition, which means getting higher-level semantic information about the target or scene from the echoes that were received. There are two main types of recognition problems: event recognition and object recognition Magbool et al. (2025). Event recognition focuses on identifying human activities or events, such as walking, falling, or hand gestures. Object recognition aims to determine the class or type of the observed object. For example, in a situation with cars, it would be able to tell the difference between cars, people, and bikes.

A common tool to assess recognition performance is the confusion matrix, which summarizes how often each true class is assigned to each predicted class by the classifier. In the binary case, the 2×2 confusion matrix contains four entries: true positives, false positives, false negatives, and true negatives. For multi-class problems, the matrix naturally generalizes to a $C \times C$ form, where each row corresponds to the true class and each column to the predicted class. The term "confusion matrix" reflects the fact that off-diagonal entries indicate how frequently the classifier confuses one class with another.

2.2 Sensing types

Sensing architectures can be classified from several perspectives. This survey considers two common classification criteria: i) the relative locations of the transmitter and receiver, and ii) the type of antenna array employed.

2.2.1 Sensing types based on transmitter and receiver location

According to the spatial relationship between the probing transmitter and the receiving array, radar sensing architectures are commonly grouped into mono-static, bi-static, and multi-static configurations.

- *Mono-static radars:* In mono-static sensing, the same physical platform (or co-located transmit and receive arrays) is used to transmit the probing signal and receive the echoes Liu et al. (2022). Such systems involve fewer geometric parameters, and quantities like angle-of-departure and angle-of-arrival are often related through simple geometry. A main challenge is self-interference between the strong transmitted signal and the weak echoes. Time-division duplexing, frequency-division duplexing, and analog cancellation circuits are typically employed to mitigate this self-interference, at the cost of additional hardware or reduced temporal efficiency.
- *Bi-static Radars:* Bi-static radars utilise spatially distinct transmit and receive arrays Kong et al. (2025). They usually need to estimate extra geometric quantities (such bi-static range), which is a drawback of mono-static systems. However, they can cut down on self-interference because the transmitter and receiver are not in the same place. Bi-

static setups are appealing in situations where a high dynamic range or less interference from the transmitter is important, such surveillance or clandestine sensing.

- *Multi-static Radars:* Multi-static radars: In multi-static radar systems, several transmitters and/or receivers that are spread out in space work together to light up and watch the target area Yonel et al. (2024). Multi-static systems can greatly increase the chances of detection, angular resolution, and tracking resilience compared to mono-static and bi-static structures. They can also cover wider areas by taking use of spatial diversity. These benefits come with a cost: the system is more complicated, with stricter synchronisation needs, more signalling overhead between nodes, and more difficult data fusion at a central processor or edge server. The detailed comparison of these three types is summarized in Table ‘3.

2.2.2 Sensing types based on array type

Sensing systems can also be categorized according to the structure of the antenna array.

Table 3 Comparison of mono-static, bistatic, and multistatic sensing configurations

Aspect	Monostatic	Bistatic	Multistatic	Implication for ISAC
Geometry	Co-located Tx/Rx; single viewpoint	Separated Tx and Rx; two viewpoints	Multiple spatially separated Tx/Rx nodes	Spatial diversity increases with separation/number of nodes
Parameter estimation	Strong SNR when co-located; limited angular/geometry diversity	Improved geometric diversity; bistatic range/Doppler effects	Highest diversity; improved observability and ambiguity reduction	Multistatic can improve estimation robustness, especially in clutter/blockage
Interference resilience	Sensitive to self-interference (full-/half-duplex issues)	Lower self-interference coupling; requires coordination	Diversity helps mitigate localized interferers; coordination overhead grows	Distributed sensing can improve resilience but needs coordination
Coverage	Limited by single-site LoS and shadowing	Better coverage due to separated nodes	Broadest coverage via distributed sites	Multistatic is attractive for intersection/canyon coverage
Synchronization	Simplest (single clock)	Requires time/frequency sync across Tx/Rx	Strongest synchronization and calibration needs	Synchronization becomes a key practical constraint
Hardware/complexity	Lowest hardware and infrastructure cost	Moderate (two nodes + backhaul)	Highest (multi-node deployment + backhaul)	Complexity–performance trade-off guides deployment choice

- *Phased Array radars:* Phased-array systems apply a transmit beamformer so that all antennas radiate a single baseband signal with different phase shifts. This coherently steers the main lobe of the radiation pattern in a desired direction and increases the beamforming gain Xu et al. (2024). Functionally, phased-array radars resemble communication transmitters that implement analog beamforming to send a single data stream, enabling fast electronic beam steering without mechanical movement.
- *MIMO radars:* In contrast, MIMO radars transmit distinct, typically orthogonal, waveforms from different antenna elements. Because the echoes of these signals experience independent fading, MIMO radars can maintain a relatively stable received SNR and achieve improved parameter identifiability Kalkan (2024). This comes at the cost of higher hardware complexity, as each antenna requires its own RF chain and more demanding baseband processing.
- *Hybrid radars:* Hybrid (phased MIMO) radars provide a trade-off between performance and complexity Wei et al. (2025). The full antenna array is partitioned into several subarrays. Each sub-array implements phased-array transmission to steer a beam, while different sub-arrays transmit independent signals. This structure combines the coherent gain of phased arrays with the diversity of MIMO, similar to the “array-of-subarrays” architecture in communication systems, and allows designers to balance beamforming capability against hardware cost. Table 4 summarizes the comparison of these types.

2.3 ISAC waveform design

ISAC systems can be classified based on the nature and purpose of waveform utilization into three categories.

a) Communication-centric ISAC: In this approach, standard communication waveforms, such as orthogonal frequency-division multiplexing (OFDM), or signals in mmWave bands / massive-MIMO systems, are repurposed for sensing functionalities Zhang et al. (2024). This enables seamless integration with existing communication infrastructure, ensuring backward compatibility. The transmitted signal can be modeled as:

$$x(t) = \sum_{n=0}^{N-1} \sum_{k=0}^{K-1} s_{k,n} e^{j2\pi k \Delta f (t-nT)}, \tag{1}$$

where $s_{k,n}$ is the modulated symbol on subcarrier k during time slot n, Δf denotes subcarrier spacing, and T is the symbol duration. The reflected and delayed versions of $x(t)$ are processed to extract range, Doppler, and angle information.

b) Sensing-centric ISAC: Originally designed radar or sensing waveforms—such as frequency-modulated continuous-wave (FMCW), pulse-Doppler, or chirp signals—are adapted to carry information Yang et al. (2025). Information can be modulated onto these waveforms using techniques like phase or pulse position modulation Leyva et al. (2024). The baseband signal is expressed as:

$$x(t) = \sum_{m} a_m p(t - mT), \tag{2}$$

Table 4 Comparison of phased-array, MIMO radar, and hybrid radar architectures for ISAC

Aspect	Phased Array Radar	MIMO Radar	Hybrid Radar	Typical ISAC Use
Beam steering	Analog/digital phase shifts form a directive beam	Waveform diversity across antennas; virtual aperture	Combines beamforming + waveform diversity	Hybrid offers flexible sensing–comms trade-offs
SNR behavior	High coherent beamforming gain (narrow beams)	Lower coherent gain per beam; diversity gain possible	Tunable: partial coherent gain + diversity	Phased array suits long-range/high-SNR sensing
Resolution/ambiguity	Good angular resolution with large aperture; narrower FoV	Improved parameter identifiability via diversity; virtual array can improve resolution	Balances resolution and coverage through mixed design	MIMO/hybrid suits rich-scattering, multi-target scenarios
Hardware cost	Often lower RF-chain count (analog beamforming)	Higher RF-chain and waveform-generation demands	Intermediate; configurable RF-chain count	Hybrid is practical for mmWave/THz constraints
Complexity	Beam management and tracking; simpler signal model	More complex processing (waveform separation, estimation)	Moderate; design complexity but scalable	Trade-off between performance and processing overhead
Application scenarios	Long-range tracking, narrow-beam sensing, mmWave beam alignment	Multi-target, rich scattering, distributed sensing	Adaptive ISAC with dynamic QoS and sensing needs	Vehicular, UAV, RIS-assisted, and dense urban deployments

where a_m is the information-bearing symbol and $p(t)$ denotes the radar pulse shape. This method is particularly effective in short-range and high-resolution environments (e.g., automotive radar).

c) Dual-function ISAC: This design entails crafting a single waveform that simultaneously enables both sensing and communication Jiang et al. (2025). The waveform is co-optimized to balance trade-offs between detection accuracy, latency, and throughput. The design objective is often modeled as a multi-objective optimization problem:

$$\min_{x(t)} \quad \mathrm{MSE}_{\mathrm{sensing}} - \lambda R_{\mathrm{comm}}, \tag{3}$$

where $\mathrm{MSE}_{\mathrm{sensing}}$ denotes the sensing mean squared error, R_{comm} is the communication data rate, and λ is a trade-off control parameter. This approach is highly flexible and considered foundational for 6 G wireless systems.

2.4 Use cases for different ISAC waveforms in industry and research

In order to connect the waveform taxonomy with practice, communication-centric ISAC is usually used in short-term deployments where standardised communication frames are reused and sensing is added as an extra feature. For example, V2X connectivity with opportunistic sensing for environmental awareness, beam management, and position/link assistance under strict compatibility constraints Zhao et al. (2024). In contrast, sensing-centric ISAC is often used with radar-driven platforms that focus on high-resolution ranging and velocity estimation and often prefer radar-like signalling with strict transmit requirements (like low-PAPR/constant-envelope properties). In these cases, communication is built in as a low-rate overlay, and the main problems are coexistence and interference mitigation Ren et al. (2026). Finally, dual-function ISAC targets integrated platforms that must provide both sensing and communication QoS at the same time. For example, roadside infrastructure that supports cooperative perception and data dissemination in busy intersections. As a result, research in this area focuses on co-designed waveforms and receivers that explicitly balance rate and sensing accuracy (e.g., CRB/MSE-related metrics) and looks at how to operate with multiple users and targets as well as how to be strong when CSI/target uncertainty is present.

2.5 Necessity of ISAC

As the requirement for spectrum efficiency, lower hardware costs, and seamless integration of perception skills in new applications like smart cities, autonomous vehicles, and industrial IoT grows, ISAC has emerged as a promising enabling technology for future 6 G networks. Having distinct systems for sensing and communication causes hardware duplication and spectrum congestion. These kinds of situations can't last because the number of connected devices and the amount of data traffic are growing quickly Jiang et al. (2025). ISAC, on the other hand, avoids these problems by sharing the spectrum and hardware. It also keeps communication reliable with low latency while providing real-time awareness of the environment. For instance, automobiles in a vehicular network that have ISAC-enabled devices can exchange data and find obstacles at the same time, making the trip safer and more efficient. There have been important technological advances that have made ISAC possible Du et al. (2025). First, you can employ communication signals as radar probes by changing waveforms like OFDM to work for both purposes. Also, systems with many antennas and RISs have the big benefit of being able to change how the signal spreads to serve both functions at the same time. Also, the combination of advanced beamforming and machine learning-based parameter estimate algorithms can very accurately tell the difference between the sensing echoes and the transmission data. Ongoing standardization efforts, including those related to 3GPP Release 18 and beyond, indicate growing practical feasibility and industrial interest in integrating sensing capabilities into future wireless systems.

2.6 Applications of ISAC

ISAC systems, through their joint support of perception and data exchange functionality, can be tailored to meet the specific needs of different application scenarios. Below are key domains where ISAC has transformative potential in 6 G networks.

2.6.1 Vehicular networks

In vehicular-to-everything (V2X) communications, ISAC can be employed to simultaneously support high-speed data transmission and real-time environment perception Khan et al. (2024). For example, an ISAC transceiver vehicle can use a single waveform to not only send safety-critical messages to nearby vehicles (V2V), road infrastructure (V2I), and pedestrians (V2P) but also detect objects (e.g., pedestrians, other vehicles) and measure their speed, range, and direction Khan et al. (2021). This feature is the basis of advanced driver assistance systems (ADAS), autonomous driving, and collision avoidance in a quickly changing road scene Du et al. (2025).

2.6.2 Industrial internet of things (IIoT):

In a smart industrial setting, ISAC enables the highly precise and efficient operation of different factory components and equipment, such as robotic arms, automated guided vehicles (AGVs), and machinery, through simultaneous sensing, communication, and control Liu et al. (2025). The devices using ISAC can detect whatever is around them (e.g., obstacles, human workers) while at the same time carrying out communication with the control unit or sensor data broadcasting Khan et al. (2020). The ISAC technology is focused on delivering ultra-low latency and high reliability, the two key elements for time-sensitive industrial automation and safety assurance. Recent perception-centric 6 G frameworks also indicate that future ISAC systems will increasingly rely on multi-technology, multi-sensor fusion, multi-band, and multi-static perception to improve environmental awareness and sensing reliability across practical deployment scenarios Li et al. (2025).

2.6.3 Indoor mapping and localization

In complex indoor settings, such as airports, shopping centers, and hospitals, the deployment of ISAC has been instrumental in maximizing user localization, movement tracking, and map creation capabilities, while at the same time, it has strengthened wireless communication Xu et al. (2024). In this regard, ISAC-based access points will not only precisely ascertain the user's position with great accuracy and identify different gestures but also simultaneously provide high-speed Wi-Fi or 6 G connection Ferrari et al. (2025), Ferrari et al. (2025). This dual functionality greatly improves navigation and elevates the overall user experience, opening avenues for the development of context-aware services, including AR/VR applications and effective asset tracking.

2.6.4 Non-terrestrial networks (NTNs)

ISAC allows satellite and aerial communication systems to combine Earth observation and communication functionalities Sheemar et al. (2025). A satellite or UAV can not only send high data rate signals to the ground terminals, but also, at the same time, sense a variety of features of the earth. Such features can include the use of land, environmental conditions, and the areas most affected by the disasters Khan et al. (2023). This strategy not only reduces payload redundancy but also offers significant benefits for remote sensing and environmental monitoring. In addition, it strengthens the integrated communications offering, especially in the areas that are difficult to reach Khan et al. (2024).

2.6.5 Smart healthcare and surveillance

Through ISAC, non-invasive monitoring of vital signs, such as respiration and heart rate, is made possible in healthcare scenarios besides improved data connectivity Zhang et al. (2025). An example is radar-based ISAC systems that can identify micro-movements associated with breathing or falls. At the same time, they send the patient's health data either to a cloud server or a local care facility. ISAC in public surveillance is employed for detecting suspicious movements or actions without interrupting the streaming of video and sensor data to the control centers. This feature significantly cuts down the risk of harm and increases the efficiency of the response.

3 RL/DRL algorithms for ISAC: an overview

RL/DRL, from the psychology of behavior, aims to find the right mix of exploration and exploitation Farhadi et al. (2025). One of the greatest issues with ISAC is the extremely fluctuating and unpredictable wireless environment Luo et al. (2025), where the performance of conventional algorithms keeps deteriorating in terms of accuracy and robustness. RL/DRL presents a viable alternative by empowering agents to make smart decisions via interaction with the environment, hence dealing with resource allocation and coordination problems efficiently. A good example is the case of sensing and communication data, which, being continuously collected and analyzed in real time, allow RL/DRL-based methods to tweak transmission parameters or sensing strategies to bring out the best in ISAC systems.

To support a deeper understanding of how AI can enhance ISAC, this section provides a brief overview of RL and DRL techniques. This section begins with an introduction to RL and highlights the most commonly used RL algorithms that have demonstrated effectiveness in ISAC scenarios. It then provides a concise overview of DL, followed by a discussion of DRL, including both value-based and policy-based approaches.

3.1 RL Fundamentals

The Markov Decision Process (MDP) is commonly used to model RL problems Alsheikh et al. (2015). It represents a stochastic control process over discrete time steps, involving a sequence of interactions between an agent and its environment. An MDP typically consists of four key components: 1) state, 2) action, 3) transition probability, and 4) reward Khan et al. (2021).

For better illustration, Fig. 3 depicts a representative ISAC scenario, where the environment state at time slot t, denoted as s_t, includes parameters such as the positions of sensing nodes, available communication resources, and signal quality indicators. Based on this state, the agent selects an action a_t to interact with the ISAC environment. The environment then transitions to a new state s_{t+1} in the next time slot according to a defined transition probability, and provides a reward r_{t+1} as feedback to the agent. The agent updates its action strategy based on the received reward, aiming to learn an optimal policy π that maximizes the expected cumulative long-term reward.

In an MDP framework, the return is defined as $G_t = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{T-t-1} R_T$, which represents the cumulative sum of future rewards from time step t onward. Here, $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards. To evaluate how desirable each state is, the state-value function is introduced as $v(s) = \mathbb{E}[G_t \mid S_t = s]$, which denotes the expected return starting from state s and following a given policy thereafter. Based on this, the well-known Bellman equation is formulated as:

$$V(s) = R(s) + \gamma \sum_{s' \in S} P(s' \mid s) V(s'), \tag{4}$$

where S denotes the set of possible next states, and $P(s' \mid s)$ is the transition probability from state s to s'. According to the Bellman equation, the value function $V(s)$ consists of two parts: 1) the immediate reward $R(s)$ obtained in the current state, and 2) the expected value of the next state, discounted by γ. Furthermore, the policy is represented by $\pi(a \mid s) = P(A_t = a \mid S_t = s)$, indicating the probability of selecting action a when in state s. The policy π encompasses the agent's entire decision-making strategy, mapping each state to a distribution over possible actions. In the Markov framework, the policy depends solely on the current state, independent of past trajectories, and can adapt over time. Simi-

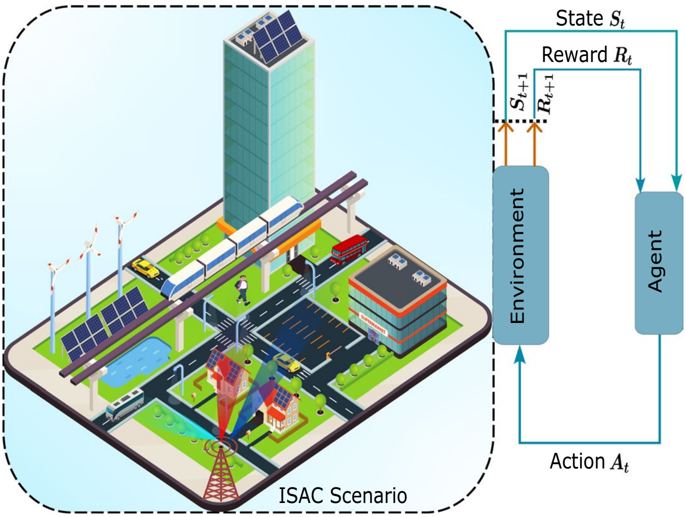


Fig. 3 RL model illustrating state, action, and reward in the context of ISAC

larly, the action-value function is defined as $q^{\pi}(s,a)$, which represents the expected return of taking action a in state s and then following policy π thereafter.

3.2 Classical RL techniques

Solving an RL problem involves identifying an optimal policy that enables the agent to consistently achieve better outcomes than alternative policies during interactions with the environment. The optimal state-value function is defined as the maximum among all state-value functions derived from different policies:

$$v^{*}(s) = \max_{\pi} v_{\pi}(s), \tag{5}$$

where $v^{*}(s)$ denotes the optimal value for state s.

1) Value Iteration (VI):

One intuitive way to determine the optimal policy is through a greedy strategy, where the agent always selects actions yielding the highest expected return. As seen in Equation (1), computing the value of a current state requires the value of subsequent states. Given the recursive nature and overlapping subproblems, dynamic programming (DP) techniques are well-suited for solving this problem. The core idea of VI is to repeatedly apply the Bellman optimality update O'donoghue et al. (2018) and use a greedy approach to improve the policy until convergence. The update rule for value iteration is given by:

$$V_{i+1}(s) \leftarrow \max_{a} \left(R(s,a) + \gamma \sum_{s' \in S} P(s' \mid s,a) V_i(s') \right). \tag{6}$$

Both value iteration and policy iteration fall under DP-based methods, which rely on iterative computations. However, their computational complexity grows significantly in environments with large or continuous state spaces, such as dynamic ISAC environments with fluctuating channel and sensing conditions. Moreover, VI is a model-based method Luo et al. (2020), assuming full knowledge of the transition probability matrix P. When such knowledge is unavailable, model-free approaches are employed instead.

2) Q-Learning

In model-free scenarios, where the transition probabilities are unknown, the agent learns directly through environment interaction. By collecting a large number of trajectories τ , the agent can estimate the true reward and transition dynamics using either Monte Carlo (MC) or Temporal Difference (TD) learning methods. While MC requires complete episodes before updating values, TD methods update the estimates online, making them more suitable for real-time ISAC applications. The TD update rule for state-value estimation is:

$$v(S_t) \leftarrow v(S_t) + \alpha \left(R + \gamma v(S') - v(S_t) \right), \tag{7}$$

where $R + \gamma v(S')$ is the TD target and $R + \gamma v(S') - v(S_t)$ is the TD error. This update replaces the current state's value with an estimate based on the immediate reward and the next state's predicted value. Q-Learning Yang et al. (2024) is a widely used off-policy TD control algorithm. In each step, the agent selects an action A in state S using an ϵ-greedy strategy. After transitioning to the next state S' and receiving a reward R, the Q-value is updated by:

$$Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_{a'} Q(S', a') - Q(S, A) \right]. \tag{8}$$

Here, the action a' is selected based on the greedy strategy, choosing the action that maximizes $Q(S', a')$. Note that this selected action is not executed but is used solely to update the Q-table. As reflected in (8), Q-Learning distinguishes between two policies: the target policy π (e.g., greedy) and the behavior policy μ (e.g., ϵ-greedy) used for exploration. This separation characterizes off-policy learning, enabling the reuse of past experiences to improve learning efficiency. In contrast, on-policy methods use the same policy for both acting and updating. Due to its ability to balance exploration and exploitation and leverage historical data, Q-Learning is well-suited for dynamic ISAC environments, where conditions frequently change, and global optimization is critical. As indicated by Fig. 4, the Q-Learning process illustrates the state, action, and reward flow with a focus on the iterative update procedure.

3.3 DL for function approximation

DL Menghani (2023) is a subfield of ML that focuses on algorithms inspired by the architecture and functionality of the human brain, particularly artificial NN. A typical NN consists of an input layer, one or more hidden layers, and an output layer. Conceptually, an NN acts as a mathematical model that takes input data X and produces an estimated output $\hat{Y}$. This transformation is carried out through a series of matrix operations in each layer, a process known as forward propagation. During the learning phase, the network iteratively updates its internal parameters θ to generate predicted outputs $\hat{Y}$ that closely approximate the true outputs Y. The discrepancy between $\hat{Y}$ and Y is measured using a loss function, generally expressed as $L(\theta) = f(\hat{Y}(\theta), Y)$, where f may represent metrics such as MSE, mean

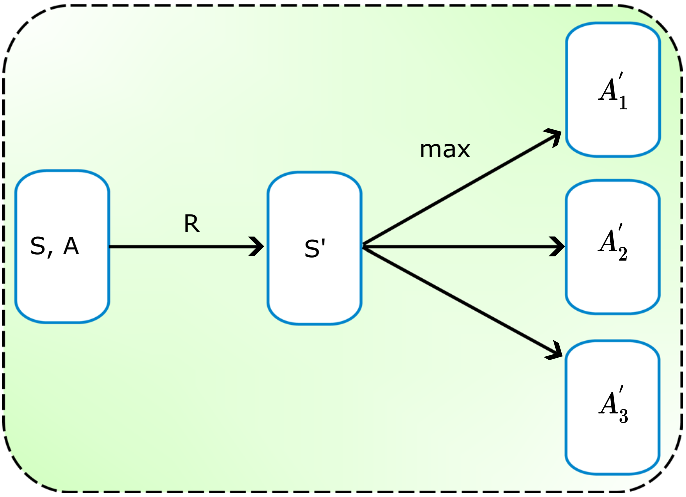


Fig. 4 Illustration of state, action, and reward flow in Q-learning

absolute error (MAE), cross-entropy, or Kullback–Leibler (KL) divergence van Erven and Harremos (2014). The goal of training is to minimize this loss function using the gradient descent optimization technique. This involves computing the gradient of the loss with respect to the network parameters:

$$\theta \leftarrow \theta - \alpha \, \Delta_\theta L(\theta), \tag{9}$$

where α is a hyperparameter known as the learning rate (or step size), and $\Delta_\theta L(\theta)$ denotes the gradient of the loss with respect to θ, indicating the direction of steepest ascent. Initially, the parameters θ_0 are randomly initialized, and then repeatedly updated using (9) until convergence to a local minimum of the loss function.

Among the various NN architectures, convolutional NN Li et al. (2022) and recurrent NN are particularly noteworthy. Convolutional NN are widely used for processing grid-like data such as images, where they apply convolution and filtering operations to extract spatial features while reducing the number of trainable parameters. recurrent NN, on the other hand, are specialized for sequential data and incorporate memory mechanisms to retain information from previous time steps Yu et al. (2019). They are commonly applied in natural language processing (NLP) tasks Fanni et al. (2023) but can also be relevant in ISAC systems that involve temporal or sequential signal processing.

3.4 Value-based DRL algorithms

Q-learning is a model-free RL approach capable of estimating the value of taking a specific action in a given state without requiring knowledge of the underlying environment model. While effective in simple environments, traditional Q-learning becomes impractical when applied to complex scenarios with large or continuous state and action spaces. In such cases, maintaining a Q-table for all state action pairs becomes memory-intensive and computationally inefficient. Furthermore, Q-learning is limited to discrete state spaces, making it unsuitable for continuous-state ISAC environments. To address these challenges, NN can be integrated into the Q-learning framework to approximate the action-value function:

$$\hat{Q}(s, a, w) \approx Q_\pi(s, a), \tag{10}$$

where w represents the parameters of the NN, and $\hat{Q}(s, a, w)$ denotes the approximated Q-value produced by the network. This approach eliminates the need to store discrete Q-values explicitly and allows for generalization across continuous state spaces. A key advantage of using NNs is that the Q-values for all actions in a given state can be computed through a single forward pass, significantly improving efficiency.

1) Deep Q-Network (DQN)

To scale Q-learning for complex tasks, Mnih et al. (2015) proposed the Deep Q-Network (DQN) algorithm, which achieved outstanding performance in Atari video game environments. The authors introduced two techniques to stabilize training: experience replay and fixed target networks. In DQN, experiences $\langle s, a, r, s' \rangle$ are stored in a replay buffer, from which mini-batches are sampled randomly for training. This breaks the temporal correlations between sequential samples and improves data efficiency (Chung 2013; Lei et al. 2020). The DQN parameter update rule is defined as:

$$w \leftarrow w + \alpha \left(Y^{\mathrm{DQN}} - \hat{Q}(s,a,w) \right) \nabla_w \hat{Q}(s,a,w), \tag{11}$$

where the TD target is $Y^{\mathrm{DQN}} = r + \gamma \max_{a'} \hat{Q}(s',a',w)$ and serves as the label. Since the same network is used for both action selection and evaluation, changes in parameters during training can destabilize the learning process. To address this, DQN employs a separate target network with parameters w^-, which is periodically synchronized with the estimation network to compute stable TD targets. Specifically, the estimation network is used to predict $\hat{Q}(s,a,w)$, while the target network computes Y^{DQN}. The loss is then minimized via gradient descent, and after several updates, the target network is updated by copying weights: $w^- \leftarrow w$.

2) Double DQN (DDQN)

Although DQN achieves high performance, it suffers from the problem of overestimating action values due to the $\max_a Q(s',a)$ operation used during target computation. This overestimation may lead to suboptimal policy learning, particularly during the early training stages. To overcome this, Lei et al. (2020) proposed Double DQN (DDQN), which decouples action selection and evaluation by using two separate networks. The TD target in DDQN is formulated as:

$$Y^{\mathrm{DDQN}} = r + \gamma Q \left(s', \arg\max_a \hat{Q}(s',a,w), w^- \right). \tag{12}$$

Here, the estimation network with parameters *w* is used to select the action *a* that maximizes the *Q*-value, and the target network with parameters w^- evaluates the selected action. This separation significantly reduces the overestimation bias and enhances the stability of the deep RL model Van Hasselt et al. (2016). The DDQN algorithm retains the overall structure of DQN while benefiting from more accurate value estimation.

3.5 Policy-based DRL algorithms

Analogous to approximating the value function using NN in Q-learning, i.e., $\hat{Q}(s,a,w) \approx Q_\pi(s,a)$, policy-based RL methods approximate the policy itself. This is typically represented as $\pi_\theta(s,a)$, where θ are the parameters of the NN used to model the policy. The agent interacts with the environment and gathers multiple trajectories τ via MC sampling, aiming to learn a policy that maximizes the cumulative reward over time. The policy gradient estimator using MC sampling is expressed as:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta} \left[\sum_{t=0}^{T-1} G_t \nabla_\theta \log \pi_\theta(a_t \mid s_t) \right], \tag{13}$$

where $\nabla_\theta \log \pi_\theta(a_t \mid s_t)$ is the score function, and G_t is the return from time step *t*, indicating the favorability of a_t.

The optimization encourages the policy to shift towards regions yielding higher rewards. A well-known algorithm based on this formulation is REINFORCE Williams (1992).

1) Actor Critic (AC)

While REINFORCE is straightforward, it suffers from high variance and inefficient sampling. To address this, the Actor Critic (AC) architecture replaces MC sampling with temporal difference (TD) sampling and substitutes G_t with an estimated Q-value. The resulting policy gradient is:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} Q_w(s_t, a_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)\right], \tag{14}$$

here, the *actor* refers to the policy network parameterized by θ, responsible for selecting actions, while the *critic* is the value-estimating network parameterized by *w*, which evaluates the action taken by the actor. To further reduce variance, the *advantage function* is introduced:

$$A_{\pi,\gamma}(s, a) = Q_{\pi,\gamma}(s, a) - V_{\pi,\gamma}(s). \tag{15}$$

This measures how much better an action is compared to the average value of the current state. The Q-value in (15) can thus be replaced by the advantage function, leading to the Advantage Actor Critic (A2C) method Wang et al. (2016). In A2C, the critic network only estimates $V(s)$, and its loss is defined as $L_{\text{Critic}}(w) = \delta^2$. The critic is updated similarly to the DQN approach. AC-based methods leverage TD sampling, making them more sample-efficient and suitable for continuous action spaces, as encountered in ISAC scenarios. Enhanced versions include Asynchronous Advantage Actor Critic (A3C) Mnih et al. (2016), Soft Actor Critic (SAC) Haarnoja et al. (2018), and Proximal Policy Optimization (PPO) Wu et al. (2021), which utilizes importance sampling for improved stability and efficiency .

2) Deep Deterministic Policy Gradient (DDPG)

In environments with high-dimensional continuous action spaces, stochastic policies can be computationally expensive due to frequent sampling. The DDPG algorithm Sumiea et al. (2024) addresses this by combining the Actor Critic framework with deterministic policies. DDPG approximates the greedy Q-value operation in DQN as:

$$\max_a Q(s, a) \approx Q(s, \mu(s \mid \theta)), \tag{16}$$

where $\mu(s \mid \theta)$ is the deterministic policy generated by the actor network. The critic network estimates the corresponding Q-value to evaluate the actor's decision.

Like DQN, DDPG employs experience replay and uses target networks for both actor and critic. Thus, it consists of four NN: (1) actor, (2) actor target, (3) critic, and (4) critic target. As illustrated in Fig. 5, the actor network takes the current state as input and outputs an action, which is evaluated by the critic to produce a Q-value. The actor's objective is to maximize this Q-value, and its loss function is defined as:

$$J^{\text{Actor}}(\theta) = -\frac{1}{m}\sum_{i=1}^{m} Q(s_i, a_i, w), \tag{17}$$

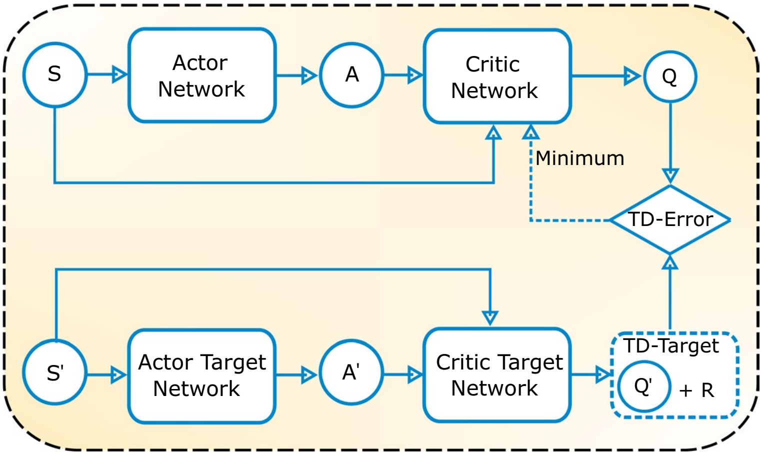


Fig. 5 Architecture of the DDPG algorithm based on actor critic framework with target networks

where *m* is the mini-batch size, and *w* denotes the critic parameters. The critic is updated using a similar method to Equation (8) in DQN.

DDPG also introduces a soft update mechanism for the target networks to improve stability:

$$w' \leftarrow \tau w + (1 - \tau) w', \tag{18}$$

where $\tau \ll 1$ is a small update rate that ensures gradual updates. This slow adjustment of target network parameters helps smooth learning dynamics and enhances convergence Sheraz et al. (2020).

3.6 Multiagent RL

Most practical scenarios in ISAC systems include various distributed entities, e.g., base stations, UAVs, vehicles, or RIS elements, which have to cooperate in achieving certain joint sensing and communication goals. Single-agent RL, however, cannot handle decentralized situations of this kind because of partial observability, inter-agent interference, and the requirement for coordinated decision-making. So, multi-agent reinforcement learning (MARL) is a decentralized and scalable framework that helps intelligent agents navigate complex couplings. MARL is a major research area that provides a decentralized and scalable framework for addressing complex problems involving multiple intelligent entities Buşoniu et al. (2010). Unlike single-agent RL, MARL uses inter-agent communication Tan (1993), learning-by-teaching Clouse (1996), and imitation learning Price and Boutilier (2003) to have a faster policy convergence and thus improve the overall system performance. Moreover, the deployment of parallel execution frameworks Kok et al. (2005) enhances exploration efficiency, with the consequence that several agents can interact with the environment at the same time.

The decision-making framework in MARL is usually represented as a partially observable MDP (POMDP) in which each agent only gets a local observation of the environment.

Each agent chooses an action based solely on its observation, and the joint action is the combination of all agents' actions. This joint action then interacts with the environment, producing a shared global reward that helps to improve policy learning.

A key algorithm in this domain is Multiagent Deep Deterministic Policy Gradient (MADDPG), which takes the single-agent DDPG and extends it to a multi-agent scenario. At training time, a centralized critic makes use of the global state to optimize individual actors. While at execution time, each actor can only access its local observation; hence, the centralized training with decentralized execution concept is maintained. Apart from MADDPG, there are also a number of value decomposition-based approaches that have gained much traction recently. Such methods combine the individual value functions of each agent to get the global action-value estimate. One example is the Value-Decomposition Network (VDN) Sunehag et al. (2017) that represents the joint action-value function as the sum of individual value functions. The QMix algorithm Rashid et al. (2020) uses a mixing NN to combine agents' value functions monotonically into the global Q-value.

Nevertheless, the advantages of MARL come with major challenges. The major limitation is probably the joint action space that is growing exponentially with the number of agents, which usually is referred to as the *curse of dimensionality*. Such complexity, in most cases, results in the convergence being slower than in single-agent settings. Additionally, an equally important problem is related to establishing effective learning objectives within the context of a multi-agent environment Buşoniu et al. (2010). Moreover, a single agent's best policy can also be affected by the changes in other agents' strategies Powers and Shoham (2004), thus, project coordination and cooperation among agents become necessary for the attainment of a global optimum. A comparative analysis of the discussed RL and DRL algorithms, including MARL techniques, is summarized in Table 5.

4 Recent advances in ISAC-enabled 6 G networks using RL/DRL

4.1 ISAC in RIS-assisted systems

RISs are positioning themselves as a keystone of 6 G networks, offering the dual ability of enhancing communication and sensing by manipulating wireless propagation. By phase-shifting and reflecting incoming signals, RIS minimizes signal attenuation, improves beamforming, and facilitates energy-efficiency (Khalid et al. 2025a, b, 2024, 2023; Chien et al. 2024).

Fig. 6 depicts a typical ISAC architecture empowered by RIS with RL/DRL algorithms. It features a base station, RIS, user devices, targets, and agents that learn to optimize beamforming and phase shifts.

Early works integrating RL/DRL into RIS-assisted ISAC systems mainly target energy efficiency (EE) in complex channels. For example, Ma et al. (2025) proposes a 3D geometry-driven propagation model for rate-splitting multiple access (RSMA) RIS-ISAC networks, capturing urban impairments such as Doppler shifts and double-Rayleigh fading. An EE maximization problem is formulated under communication and sensing QoS constraints, solved using a PPO DRL strategy to refine transceiver beamforming, RIS phases, and sensing. The actions consider beamformers and phases, states include SINRs, and

Table 5 Comparison of RL/DRL Algorithms

Algorithm	Variants	Key Characteristics	Advantages	Limitations
VI	Policy Iteration (PI)	Utilizes dynamic programming methods	Fast convergence through efficient iterations	Dependent on transition matrices; inefficient in large-scale problems
Q-Learning	SARSA, Dyna-Q	Maintains Q-values in a lookup table for each action	Flexible and effective via temporal-difference sampling	Poor scalability in continuous domains
DQN	Double DQN	Replaces Q-table with a neural network	Rapid convergence and ease of use	Risk of Q-value overestimation
DDQN	–	Uses two networks to decouple action evaluation and selection	More stable convergence than DQN	Still sensitive to hyperparameters; limited in highly non-stationary settings
REINFORCE	PPO	Based on stochastic policy optimization	Enables direct learning of optimal policies	High variance and low sample efficiency
AC	A2C, A3C	Combines value and policy learning components	Supports continuous control; improves stability	Can suffer from instability without careful tuning
DDPG	MADDPG	Handles continuous actions via deterministic policies	Stable learning and good performance in smooth control tasks	Limited exploration under deterministic policies; can get stuck in local optima
VDN	QMIX	Sums individual Q-values to approximate a global Q	Accelerates learning in cooperative tasks	Limited expressiveness for complex joint action values
QMIX	–	Combines individual Q-values through a mixing network	Improved joint Q estimation accuracy	Typically designed for monotonic value factorization assumptions

rewards penalize power and echo SNR violations. Simulations indicate higher EE over space-division multiple access benchmarks and greedy and random search baselines.

Wang et al. (2025) optimizes transmit beamforming and RIS phase in passive RIS-ISAC, balancing the sum data rate while minimizing the sensing CRB. An MDP model is developed with active/passive beamformers in actions, channel in the state space, and rewards quantified as penalties for power, rate, and CRB losses. Comparing soft actor critic (SAC), twin delayed deep deterministic policy gradient (TD3), and deep deterministic policy gradient (DDPG), the SAC agent demonstrates superior exploration efficiency and stability, yielding the highest achievable sum rate and the lowest CRB, underscoring the DRL's capability to manage high-dimensional optimization landscapes. Yang and Luo (2024) addressed secure beamforming for RIS-aided ISAC under the stringent CRB constraint, where the target could be an eavesdropper. By formulating a joint BS beamforming and RIS phase optimization-based secrecy rate maximization problem, they suggested a DDPG-based algorithm with a modified reward function to impose CRB constraints. Simulation results verified the secrecy sensing trade-off with better convergence and robustness compared to conventional optimization methods.

Shifting to specialized sensing, Jukuri et al. (2025) proposed novel Meta-IoT sensors that serve dual functions by leveraging frequency-response and reflected orthogonal fre-

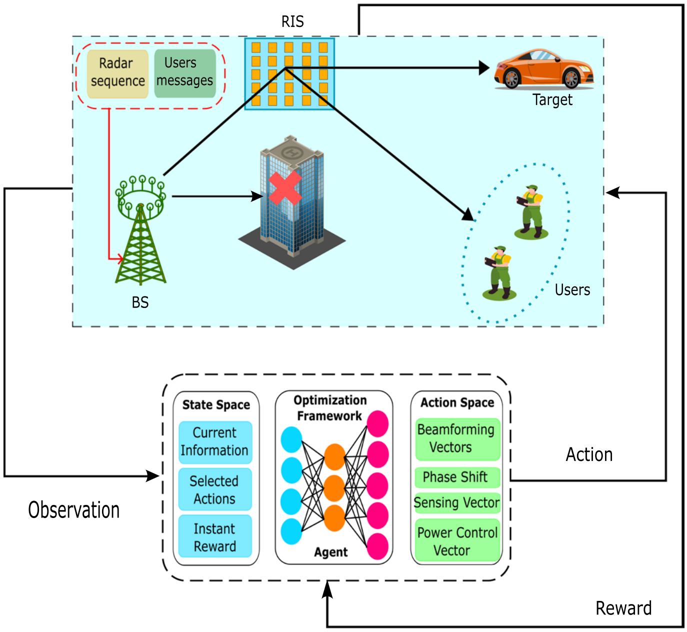


Fig. 6 RIS-assisted ISAC using RL/DRL algorithms

quency division multiple access-based reflections, without dedicated hardware. To handle non-convex optimization, they formulated hierarchical problems that were addressed using Gaussian process regression (GPR), which included sensing constraints while maximizing data rates. By combining GPR with multi-agent DRL (MADRL), the architecture reached a throughput of 33 Mbps, showed faster convergence with larger antenna arrays, and achieved a good compromise between performance (DQN) and fairness (Meta-PPO), outperforming both random and uniform policies. Chen et al. (2024) used the same MADRL framework to mmWave-THz RIS-ISAC contexts, enhancing phase and BS beamforming to balance bandwidth, data rate, and sensing resolution. The modular framework included a variety of channel conditions, which made UAV trajectory optimization in air-ground systems much better and allowed for ultra-low latency edge computing.

When deployed in vehicular networks, DRL adapts to integrated challenges. Ren et al. (2025) developed a DRL mechanism for RIS-ISAC in the Internet of Vehicles (IoV). The architecture instructs roadside units to transmit ISAC signals to approaching vehicles while contending with path loss and obstructions, thereby delegating high-complexity tasks to edge servers. The problem is formalized as an MDP aimed at maximizing the sum rates under sensing SINR, power, and latency thresholds. A dual-critic DDPG, enhanced by batch sampling and MSE loss to reduce Q-value overestimation, achieves a 79% increase in sum-rate compared to passive RIS, outperforming single-critic DDPG, alternating optimisation, and random-selection baselines in packed environments. After that, Saikia et al. (2024) came out with a hybrid DRL framework for RIS-aided cooperative ISAC in cloud radio access networks (C-RAN). The objective is to reduce the squared position error bound (SPEB) by optimising transmit beamforming, RIS phases, and subcarrier assignment, while adhering to power and SINR constraints. Using simultaneous localisation and communication (SLAC), the MDP-based HDRL coupled continuous action DDPG with discrete DQN. It used an actor-critic architecture and ensemble estimates to make sample-independence and operational stability stronger. Experimental assessments demonstrate that the SPEB exhibits an inverse relationship with RIS units, BS power, and antennas, hence validating HDRL's situational awareness within the urban smart-city framework.

Taking the problem to aerial platforms, Xie et al. (2025) tackled multi-objective problems in UAV-deployed RIS-ISAC setups using dual-function radar-communication (DFRC) schemes. The work formulated a communication-sensing-energy efficiency multi-objective problem (CSEMOP) that maximizes user sum rates and sensing accuracy, and minimizes UAV flight energy simultaneously, by adjusting BS beamforming, RIS phases, and flight paths. A multi-agent DDPG (MADDPG) is paired with diffusion models that refine actions, noise perturbation for enhanced exploration, and prioritized experience replay. Trials across different transmit powers, UAV flight altitudes, and antenna configurations revealed better convergence and overall quality than standard DDPG, TD3, and various baselines, validating the advantage of generative models in balancing multiple objectives in mobile RIS deployments. Cho et al. (2024) advanced a RIS-ISAC configuration on aerial platforms to improve battlefield situation awareness, minimizing the CRB in target location estimates by coordinating beamforming, RIS phases, and UAV paths within SINR thresholds. Using a DDPG architecture that is handled as a POMDP, the system state is made up of ARIS location and target estimates, the action is the UAV's speed, and the rewards are the opposite of the CRB penalty. The design also included near-field steering processing (NSP) in the designs for the receive antennas to reduce self-jamming and clutter returns. The trials showed that the sensing output was clearer and more stable than with fixed RIS phase and no-NSP settings. A comparable endeavour in Chen et al. (2024) concentrated on RIS-ISAC scenarios, where the design emphasised a max-min user rate, concurrently optimising transmit, RIS, and UAV trajectories while adhering to power, sensor SNR, and flight trajectory constraints.

Moving beyond traditional RIS frameworks, Ye et al. (2025) examined intelligent omnisurface (IOS)-ISAC configurations to enhance sensing signal-to-noise ratio (SNR) while maintaining user throughput, achieved by heuristic optimisation of IOS coefficients without explicit channel state information (CSI). The DRL DeepOSC architecture used echo state networks (ESNs) with output nodes to turn previously recorded IOS configurations into state vectors, coefficients into actions, and SNR into rewards. The system sped up the learning process by adding agents that had already been trained on similar tasks. This was done by combining staged policy reuse (SPR) and staged policy distillation (SPD) modules. Outcome evaluations showed that sensor SNR and user satisfaction were better than both DoubleRIS and majorization-minimization frameworks. The SPR/SPD, on the other hand, cut down on the number of time slots needed by 55 58% and kept performance steady across different IOS designs, user habits, power levels, and situations that changed over time. In active RIS settings, Zhang et al. (2024) examined full-duplex ISAC systems that manage active RIS to support non-orthogonal multiple access (NOMA) devices, aiming to optimise the long-term average radar echo SINR through synergistic regulation of power and RIS coefficients. The system was set up as an MDP, with CSI as the states, RIS amplitudes and phases as the actions, and awards for high radar SINR and punishments for violations. SAC did better than DDPG and PPO algorithms. Simulations validated the advantages of active RIS, demonstrating a greater sensing SINR compared to passive RIS, with gains amplifying as elements increase. In RSMA-ISAC systems, Rahimi et al. (2025) talks about SAC-QA, a DL framework for joint BS precoding and beyond-diagonal RIS (BD-RIS) beamforming. The system tackles nonconvex optimisation with noisy CSI by merging action quantisation and low-rank sparse BD-RIS approximation. Simulations demonstrate enhanced convergence, optimal sum rate, and radar SINR in comparison to traditional SAC, DDPG, and standard RIS benchmarks. Adhikary et al. (2024) introduced an Age of Sensing (AoS)-holographic ISAC framework for HMIMO-capable 6 G networks. The framework ensures timely and robust holographic beamforming by using a variational autoencoder (VAE) to update the user's location and a DDPG-based DRL method to allocate power. Simulation results demonstrated enhancements in SINR and attainable rates relative to baseline methodologies, highlighting the potential of AoS-based DRL techniques in ISAC.

This evolution extends to simultaneously transmitting and reflecting RIS (STAR-RIS), where DRL addresses coupled optimizations. Zhu et al. (2023) examined STAR-RIS-aided ISAC secure communications and achieved secrecy rate maximization with echo SNR and rate constraints. Employing DDPG and SAC, they jointly optimized BS beamforming and STAR-RIS coefficients and proved that STAR-RIS significantly outperforms conventional RIS and double-spliced RIS benchmarks in secrecy performance. Zhang et al. (2024) focused on dual-functional constant modulus waveform design in STAR-RIS-ISAC, minimizing weighted multi-user interference (MUI) energy and waveform discrepancies by optimizing transmit waveforms and STAR-RIS coefficients under SINR and power constraints, considering independent and coupled phase-shifts. For the independent case, an ADMM algorithm provided closed-form solutions; for the coupled case, TD3 solved the MDP with channels/symbols as states, waveforms/coefficients as actions, and an objective function with penalties as rewards. Simulations showed superior convergence, BER, and rate over conventional RIS, with enhanced trade-offs as elements or power increased. Complementing this, Qing et al. (2025) introduced an attention-based decoupling space (AT-DS) DRL for STAR-RIS UAV air-ground ISAC, optimizing UAV trajectory, power, and STAR-RIS beamforming to maximize user SINR and CRB under energy, speed, and phase/amplitude constraints. Modeled as an MDP with UAV position/velocity/energy and channels as states, discrete velocities/power allocation/beamforming as actions, and normalized ISAC metrics with bonuses/penalties as rewards, AT-DS employed SAC with dual actors, one attention-fused for dynamic users, another for trajectory, outperforming DDPG, TD3, and SAC baselines in convergence, rewards, and scalability. Further enhancing STAR-RIS applications, Kamal et al. (2025) proposed DRL for secure multi-user ISAC in 6 G, maximizing secrecy rate via joint BS beamforming and STAR-RIS coefficients optimization under SINR, echo SNR, and power limits in Rician/Rayleigh channels with eavesdroppers. As an MDP with CSI as states, beamforming/phases as actions, and secrecy rate minus penalties as rewards, SAC outperformed DDPG in convergence, reward, secrecy rate, and echo SNR due to entropy regularization, surpassing traditional RIS in energy efficiency and security with increasing users/power/fading. Similarly, Kurma et al. (2024) developed MADDPG with dynamic delay alignment (DDA) modulation for resource management in active STAR-RIS THz ISAC, enabling dual-function BS to serve vehicular units and detect targets in time-selective fading. Formulated as a multi-agent MDP with CSI as states and beamformers as actions to maximize sum rate under sensing SNR, interference, power, and hardware constraints, simulations highlighted convergence and sum rate advantages over DDPG, PPO, M-PPO, passive STAR-RIS, and conventional RIS.

Discussion: Table 6 summarizes RIS ISAC systems with DRL/RL algorithms in urban RSMA-RIS to THz active-STAR-RIS. It highlights algorithms, such as PPO, SAC, TD3, and MADDPG, that achieve dynamic multi-objective trade-offs and generative AI (GenAI) flexibility. Key objectives, EE maximization, sum-rate balancing, and CRB minimization, are achieved through beamforming, phase shifts, and trajectories. Metrics such as convergence and scalability authenticate DRL's success in complex environments. This work provides RL/DRL solutions for RIS ISAC multi-user complexity, energy/spectrum-efficient secure communication in active/passive/STAR-RIS scenarios. Mobility, fading, and constraints are addressed with MADDPG/PPO, enabling new channel designs and multi-objective optimization. Challenges are real-time CSI, high-dimensional space, and hardware constraints for future research into multi-RIS frameworks, 6 G-THz architectures, and hybrid DRL/classical solutions for enhanced usability and reliability. The reviewed papers reflect a trajectory from EE optimizations of initial RIS setups to complex, mobile, and secure ISAC systems. Initially, the work was concentrated on passive RIS that could solve urban problems, then there was an extension to GenAI and hybrid DRL for aerial/vehicular mobility contexts. There has been a consistent focus on MDP modeling of the combined optimizations, along with improved convergence, throughput, and sensing accuracy when compared to the reference methods. Real-time processing speed and hardware limitations point out the potential for hybrid AI-traditional methodologies in future 6 G applications.

4.2 ISAC in UAV/satellite systems

UAVs and satellites have played an essential role in the development of ISAC technologies because they inherently possess mobility, can be deployed flexibly, have line-of-sight (LoS) channels, and are connected everywhere. ISAC and ISCC frameworks are becoming increasingly popular in the literature, for which DRL techniques are leveraged to optimize UAV trajectory planning and resource allocation jointly. Dynamic, environment-aware decision-making that is tailored to UAV mobility and heterogeneous mission goals is replacing static network planning, as figured out in Fig. 7.

Several research works utilize DRL for simultaneous communication and sensing optimization in both single and multi-UAV cases. In Tang et al. (2023), a PER-DDQN-based system handles target detection, data transmission scheduling, and UAV trajectory control, with faster convergence and better policy learning. The developed approach demonstrates a strong capability of adapting to different UAV flight modes and provides accurate sensing even when the data are highly variable. A different multi-objective DRL approach determines flight paths, power control, and data freshness for the acquisition of data by multi-UAV coordination Liu et al. (2024). The authors exhibited considerably decreased latency in data collection cycles and better scalability as the number of UAVs increased. The PBIA scheme in Hou et al. (2025) enables over-the-air ISCC by embedding DRL to optimize the service assignment, time-slot allocation, and power transmission in UAV swarms. This design features asynchronous user scheduling and exhibits a high level of generalization to unseen task distributions. An A3C-based decentralized DRL method for coordinating cooperative UAVs to carry out ISCC functions, which achieves more than 80% energy efficiency improvement over conventional methods, is described in Hou et al. (2025). Decentralized learning enables a system to be highly resilient in a non-stationary multi-agent environment, which is the defining characteristic of aerial networks.

Table 6 Summary of RL/DRL Approaches in RIS-Assisted ISAC Systems

Ref	System Model	RL/DRL Algorithm	Key Features	Optimization Variables	Objective	Main Results
Ma et al. (2025)	RSMA RIS-ISAC networks	PPO	3D geometry-driven propagation model, urban impairments	Transceiver beamforming, RIS phases, sensing	EE maximization under QoS constraints	Higher EE over benchmarks
Wang et al. (2025)	Passive RIS-ISAC	SAC, TD3, DDPG	Balances sum data rate and sensing CRB	Active/passive beamformers, RIS phase	Maximize sum data rate, minimize CRB	Highest sum rate, lowest CRB with SAC
Yang and Luo (2024)	RIS-aided secure ISAC	DDPG	Secrecy rate maximization with CRB constraint	BS beamforming, RIS phase	Maximize secrecy rate	Better convergence, robustness
Jukuri et al. (2025)	Meta-IoT sensors in RIS-ISAC	MADRL (GPR, DQN, Meta-PPO)	Dual-function sensors, hierarchical optimization	Phase shifts, beamforming	Maximize data rates with sensing constraints	Higher throughput, accelerated convergence
Chen et al. (2024)	mmWave–THz RIS-ISAC	MADRL	Trade-offs in bandwidth, data rate, sensing resolution	Phase shifts, BS beamforming	Optimize UAV trajectory, low latency	Uplift in air–ground systems
Ren et al. (2025)	RIS-ISAC in IoV	Dual-critic DDPG	Edge servers for high-complexity tasks	Beamforming, phases	Maximize sum rates under constraints	Sum-rate increment over passive RIS
Saikia et al. (2024)	RIS-aided cooperative ISAC in C-RAN	HDRL (DDPG + DQN)	SLAC, subcarrier assignment	Transmit beamforming, RIS phases, subcarriers	Minimize SPEB	SPEB scales inversely with RIS units and power
Xie et al. (2025)	UAV-deployed RIS-ISAC DFRC	MADDPG	CSEMOP	BS beamforming, RIS phases, UAV paths	Maximize rates/accuracy, minimize energy	Better convergence than DDPG/TD3
Cho et al. (2024)	RIS-ISAC on aerial platforms	DDPG	Battlefield awareness, NSP	Beamforming, RIS phases, UAV paths	Minimize CRB	Clearer sensing output
Chen et al. (2024)	RIS-ISAC	(Not specified; implied DRL)	Max–min user rate	Transmit, RIS, UAV trajectory	Maximize min user rate	(Not detailed)
Ye et al. (2025)	IOS-ISAC	DeepOSC (ESNs, SPR/SPD)	No explicit CSI, heuristic optimization	IOS coefficients	Boost sensing SNR, ensure throughput	Enhanced SNR, reduced time slots
Zhang et al. (2024)	Full-duplex active RIS-ISAC NOMA	SAC	Long-run average radar echo SINR	Power, RIS coefficients	Maximize radar SINR	Higher sensing SINR than passive RIS
Rahimi et al. (2025)	BD-RIS in RSMA-ISAC	SAC-QA	Action quantization, low-rank approximation	BS precoding, BD-RIS beamforming	Improve sum rate, radar SINR	Improved convergence, worst-case sum rate

Table 6 (continued)

Ref	System Model	RL/DRL Algorithm	Key Features	Optimization Variables	Objective	Main Results
Adhikary et al. (2024)	AoS-holographic ISAC HMIMO	DDPG with VAE	User location update, power allocation	Power allocation, holographic beamforming	Guarantee timely updates	SINR gains, rate improvements
Zhu et al. (2023)	STAR-RIS-aided secure ISAC	DDPG, SAC	Secrecy rate with constraints	BS beamforming, STAR-RIS coefficients	Maximize secrecy rate	Outperforms conventional RIS
Zhang et al. (2024)	STAR-RIS-ISAC	TD3 (coupled), ADMM (independent)	Constant-modulus waveform, MUI minimization	Transmit waveforms, STAR-RIS coefficients	Minimize MUI and discrepancies	Superior BER, rate
Qing et al. (2025)	STAR-RIS UAV air–ground ISAC	AT-DS SAC	Attention-based decoupling	UAV trajectory, power, STAR-RIS beamforming	Maximize SINR, minimize CRB	Better convergence than DDPG/TD3
Kamal et al. (2025)	Secure multi-user STAR-RIS ISAC	SAC	Eavesdroppers in Rician/Rayleigh channels	BS beamforming, STAR-RIS coefficients	Maximize secrecy rate	Higher reward and secrecy rate than DDPG
Kurma et al. (2024)	Active STAR-RIS THz ISAC	MADDPG with DDA	Resource management in fading	Beamformers	Maximize sum rate	Sum-rate advantages over baselines

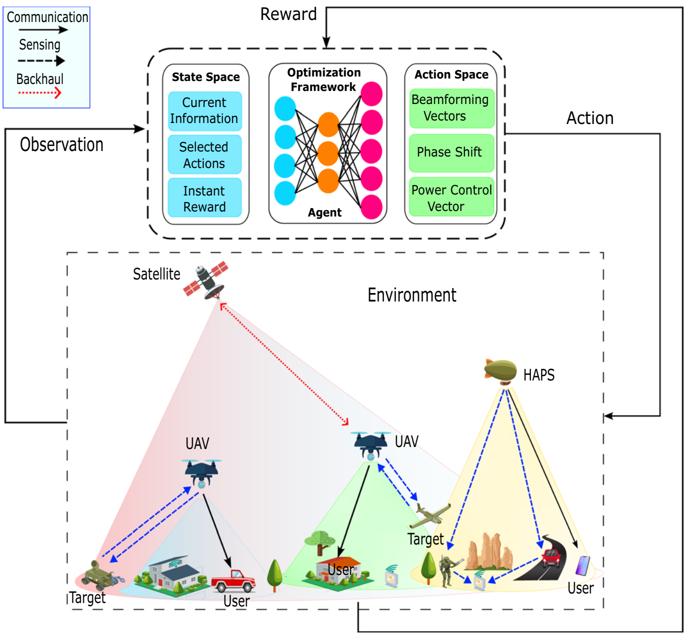


Fig. 7 UAV and satellite-based ISAC using RL/DRL algorithms

The authors in Chen et al. (2025) use TD3 algorithm in a digital twin to perform UAV autonomous tracking and obstacle avoidance in real-time. The hybrid digital-twin implementation enables closed-loop control and faster convergence in dynamic and cluttered terrains. Centralized and decentralized SAC-based solutions are explored in Qin et al. (2023) for managing UAV trajectories, user assignments, and power levels in multi-UAV ISAC settings. The methods demonstrate tradeoffs between the model complexity and the communication overhead, as well as the robustness of the distributed UAV clusters. In addition, Liu et al. (2024) introduces a flight-hover-communicate flight strategy based on DRL to enhance IoT connectivity as well as to save energy. The UAVs learn to switch between movement and transmission phases intelligently, reducing flight-induced energy waste. To address unknown IoT deployments, Liu et al. (2025) presents a dueling DDQN method to improve estimation precision and reduce flight overhead in UAV-ISAC networks. This model could easily handle sparsely deployed scenarios, leading to a better coverage ratio and sensing reliability in resource-constrained UAV settings.

Research work been shifting towards joint optimization of beamforming and trajectory. The authors in Gao et al. (2024) formulated a 3D UAV deployment problem with mobile users and solved it using a hybrid-reward multi-agent PPO (HR-MAPPO) framework. The solution was a combination of beamforming and trajectory control that outperformed traditional baselines. They also proposed a hierarchical user association strategy through K-means clustering that increases the network throughput and spatial reuse. Moreover, the problem of integration of RIS with UAV-ISAC in Moon et al. (2024) is divided into two parts and solved by DRL methods to perform RIS configuration and UAV beamforming optimization under joint throughput and sensing SNR constraints. Besides enhancing link robustness, this dual optimization also provides more freedom for urban and indoor sensing scenarios. In the context of agriculture, Saikia et al. (2025) combines the usage of DDPG for trajectory and beam allocation with that of a GAN model for increasing the realism and trustworthiness of the acquired sensing data. The use of DRL together with generative models may be a great tool for the development of self-adaptive UAVs capable of data-intensive tasks, such as crop monitoring or soil assessment.

Security-centric ISAC methods continue to be a main topic of discussion. The SURE-ISAC approach in Chen et al. (2024) makes use of PPO to determine the best paths for drones, their power of transmission, and scheduling to maintain secure communications in adversarial environments. The reward function integrates secrecy rate and adversarial avoidance, making it an excellent choice for the UAVs whose flights are operated in contested airspaces. Liu et al. (2024) depicts a DRL-based UAV emergency response system, which utilizes transfer learning to be able to react rapidly to a damaged situation. Basically, transfer learning from a pre-trained policy allows the agent to handle unforeseen situations fast after a disaster. By means of DRL, Paul et al. (2025) has come up with methods of suppressing jamming by reliable power and trajectory control in a hostile environment. Altogether, these works show that DRL can help increase the security of UAV-ISAC to withstand both active and passive forms of attacks.

From a multiple access standpoint, Amhaz et al. (2024) studies UAV-assisted cooperative NOMA-ISAC systems and uses DDPG to co-optimize beamforming and UAV positioning under QoS and CRB requirements. The setup improves the performance of the edge user and ensures sensing compliance with CRB by dynamically changing the UAV's roles between relay and base station. Zhu et al. (2024) presents a hybrid RL-based resource allocation strategy in NOMA-UAV networks and attains effective bandwidth and altitude control while limiting interference. This hybrid policy learning method tunes continuously with diverse user requirements and different mobility patterns, hence, it guarantees fairness in throughput and sensing fidelity.

The integration of ISAC with edge computing and cloud intelligence was the focus of the study conducted in Qi et al. (2024), where an actor-critic RL model was used to improve beamforming in LEO satellite ISAC networks challenged by outdated CSI. Their approach shows that it can be highly robust against channel dynamics and uses prediction to reduce retransmission delays to a minimum. Going further, Li et al. (2025) introduces a meta-DRL scheme for fast satellite-UAV edge computing task offloading. Meta-learning accelerates UAV adaptation to new IoT tasks; hence, the architecture is suitable for emergency and pop-up network deployments. Also, Ye et al. (2025) leverages federated PPO for UAV-ISAC edge networks to mitigate privacy issues and address non-IID sensing distributions. This approach promotes collaborative learning without revealing local data, which is essential in healthcare and surveillance.

In Wang et al. (2022), the authors extended digital twin and federated learning paradigms and present a DTFL-based DRL orchestrator for robust UAV-ISAC under hardware impairments. Such an architecture can perform adaptive decision fusion and coordinate in real time. Through local DTFL feedback, the system continuously modifies the UAV policies, thus offering resilience in situations of partial visibility and sensor disappointments. Xie et al. (2024) is based on MARL to lead a UAV swarm in an energy-efficient way while avoiding collisions in vehicular ISAC environments. The MARL agents collectively balance inter-UAV spacing, coverage efficiency, and safety in traffic-aware navigation tasks. Addressing ISAC training in cloud-centric or data-offloading scenarios is an emerging research frontier in which UAV intelligence is partially handed over to the edge-cloud continuum.

All the above contributions strongly demonstrate how DRL can be used to improve UAV-assisted ISAC/ISCC systems by optimizing flight paths, managing beams, ensuring secure communication, integrating RIS, and enabling edge intelligence. 6 G infrastructures in the future will most probably be designed around multi-agent coordination, federated secure learning, and real-world ISAC deployment. Moreover, by combining generative learning, proactive sensing models, and digital-twin-based control systems, new applications in disaster relief, autonomous urban mobility, and other areas can be unlocked.

Conversely, ISAC can be used for UAV detection, tracking, and anti-drone systems, a critical application in the civilian and defense sectors. In AlKhonaini et al. (2024), a hierarchical RL (HRL) framework uses RF signals for multi-level UAV classification (presence, model, flight mode) via REINFORCE with entropy regularization, achieving 99.7% detection accuracy and enhancing anti-drone security by repurposing communication RF for sensing without extra hardware. Liu et al. (2025) presents a DL-based ISAC system to track UAVs in multipath scenarios, detecting anomalies in the first Fresnel zone with bidirectional Long Short-Term Memory for trajectory prediction and smoothed 2D MUSIC for high angular resolution. For covert UAVs, agents can adjust trajectories and power to evade ISAC surveillance, showing the role of RL in the counter-drone game. These works integrate ISAC with RL/DRL for robust UAV surveillance and tackle observability and multipath interference in real-world deployments.

Discussion: Table 7 provides an exhaustive overview of recent work in DRL-based UAV-assisted ISAC and ISCC systems in different application scenarios. The table shows

Table 7 Summary of DRL-based UAV-assisted ISAC and ISCC literature

Ref.	System Model	Sensing Type and Metrics	Algorithm	UAV Use	Target Metric	DRL Method
Tang et al. (2023)	UAV-assisted ISAC	Target detection; utility	PER-DDQN	Single UAV	Joint utility maximization	DDQN
Liu et al. (2024)	Multi-UAV ISAC	Freshness; data collection	Pareto-front DRL	Multi UAV	Throughput and freshness	DDPG-based
Hou et al. (2025)	UAV swarm ISCC	Service success rate	PBIA (custom DRL)	Swarm UAVs	Energy efficiency	Actor–Critic
Chen et al. (2025)	AAV with ISAC	Tracking; SNR	TD3 + digital twin	Single AAV	Collision-free tracking	TD3
Liu et al. (2024)	Post-disaster ISAC	QoS; mission completion	Transfer DRL	Single UAV	Robust convergence	DDPG + Transfer
Qi et al. (2024)	LEO Satellite ISAC	CSI resilience	Actor–Critic	UAV Relay	Beamforming accuracy	Actor–Critic
Qin et al. (2023)	Multi-UAV ISAC	PSL and CRB	MASAC	Multi UAV	Spectral fairness	Soft Actor–Critic
Hou et al. (2025)	Distributed ISCC	Service rate	MCAI (A3C)	Multi UAV	Energy efficiency	A3C
Xie et al. (2024)	Vehicular ISAC	Sensing fairness	MARL	UAV Swarm	Energy-aware fairness	Multi-agent RL
Moon et al. (2024)	RIS-assisted ISAC	SNR, throughput	Joint Beam-RIS DRL	Single UAV	Joint SE/SNR gain	DDPG-style
Zhu et al. (2024)	NOMA UAV ISAC	Interference, CRB	Altitude-Bandwidth DRL	Multi UAV	Interference suppression	Hybrid RL
Paul et al. (2025)	Secure UAV ISAC	Secrecy rate	DRL + adversarial modeling	Single UAV	Transmission secrecy	DDPG-style
Saikia et al. (2025)	Smart agriculture ISAC	Data quality (GAN)	DDPG + GAN	Single UAV	Realistic sensing	DDPG
Ye et al. (2025)	Edge UAV-ISAC	Sensing accuracy	FedPPO	Multi UAV	Privacy-preserving coordination	PPO
Li et al. (2025)	Sat-UAV Edge ISAC	Task latency	Meta-DRL	UAV + Sat	Rapid task adaptation	DDPG + Meta
Liu et al. (2024)	IoT UAV ISAC	Coverage; energy	Flight-hover DRL	Single UAV	Max connectivity	Custom DRL
Liu et al. (2025)	IoT UAV ISAC	Estimation rate	Dueling DDQN	Single UAV	Flight + sensing gain	Dueling DDQN
Gao et al. (2024)	Roaming user ISAC	Beam gain; rate	HR-MAPPO	Multi UAV	Long-term SE	Multi-agent PPO
Chen et al. (2024)	Secure UAV ISAC	Secrecy; distance	SURE-ISAC (PPO)	Single UAV	Secrecy rate	PPO
Amhaz et al. (2024)	C-NOMA UAV ISAC	CRB + rate	DDPG	UAV Relay	QoS + sensing	DDPG
Wang et al. (2022)	DTFL UAV-ISAC	Impairments; sensing	DTFL Orchestrator	Multi UAV	FL convergence	DDPG-style

how different DRL variants, such as DDPG, PPO, A3C, and actor-critic, have been combined with UAV to solve problems related to path planning, sensing reliability, and communication performance. A range of system models, from single UAVs and swarms to satellite-UAV cooperation, shows that DRL is equally capable in both centralized and distributed architectures. Sensing metrics include target detection, beamforming gain, CRB, and secrecy rate, which together reflect the multifaceted nature of ISAC design objectives. Besides that, the majority of the articles rely on multi-agent settings and hybrid learning schemes for the agents to timely adjust their resource allocation to uncertain and limited CSI. The integration of the latest paradigms, such as RIS, digital twins, federated learning, and meta-learning, is also indicative of the trend towards smart, context-aware UAV-ISAC deployments. Overall, this table shows a huge potential of DRL to integrate sensing, communication, and control in future UAV-assisted 6 G networks.

4.3 ISAC in vehicular systems

Several works contributed to enhancing ISAC systems for V2X networks using learning-based methodologies to enhance energy efficiency and communication-sensing performance, as depicted in Fig. 8. Shang et al. (2025) put forward an energy-efficient ISAC system that used V2X dynamics as an MDP, and a DRL algorithm based on Actor-Critic, combined with Spiking NN was used to solve the V2X dynamics. The authors' method highlights the need for pilot signaling and significantly reduces the energy consumption of NN while still optimizing beamforming and power allocation, which led to higher sensing accuracy and communication rates. Similarly, Li et al. (2025) addressed vehicle mobility and multi-source perception data fusion for autonomous driving. They integrated ISAC

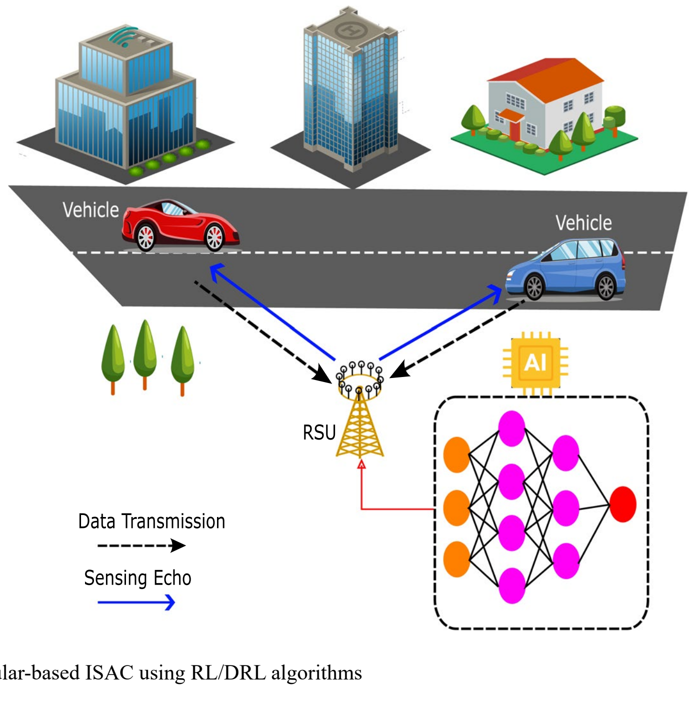


Fig. 8 Vehicular-based ISAC using RL/DRL algorithms

with vehicular edge computing (VEC) and formulated a DRL-based resource allocation scheme using an improved dueling twin delayed deep deterministic policy gradient (ID-TD3) algorithm. The method minimized task completion delay and energy consumption under dynamic channel conditions. Lai et al. (2025) investigated interference mitigation in multi-beam V2X communications by proposing IMSBA, a joint beam direction, power, and spectrum allocation algorithm. By combining multi-agent proximal policy optimization (MAPPO) and Stackelberg game theory, IMSBA improved V2V energy efficiency by 92.5% and enhanced overall communication and sensing performance in spectrum-limited millimeter-wave IoT scenarios.

Several papers have presented advances in ISAC systems that use RIS and DL. In Long et al. (2024), a RIS-assisted 6 G V2X system was proposed for vehicle positioning at cm-level accuracy combined with vehicle-to-vehicle communication capacity. The researchers introduced a phase modulation optimization in RIS units by a flexible deep deterministic policy gradient (FL-DDPG) algorithm integrated with an ϵ, greedy strategy, which resulted in a significant decrease in positioning error (up to 89%) and nearly tripled the achievable communication rates over classical methods. On the other hand, the paper Liu et al. (2024) dealt with the downlink V2I ISAC networks' joint beamforming and power allocation design. The authors solved the problem of non-convex Cramér -Rao lower bound (CRLB) sensing constraints by DRL and hence managed to improve the sum-rate communication performance while keeping the sensing accuracy. Furthermore, Yang et al. (2024) developed an integrated sensing, communication, and computation (ISCC) paradigm for vehicular networks, wherein over-the-air federated learning was utilized. The study made a joint beamforming and power allocation problem and used a hybrid RL method to solve it. It combined semidefinite relaxation and deep deterministic policy gradient algorithms. The results demonstrated that their approach outperformed others not only in terms of convergence but also in data rates in the presence of time-varying channel conditions.

ISCC has been further explored in the contexts of IoT and vehicular systems. A work in Ren et al. (2024) reveals the technological potentials of ISCC-enabled IoT systems, integrating ISAC with the mobile edge computing (MEC), over-the-air computation (AirComp), and federated edge learning (FEEL). The study formulated a unified performance optimization problem, which was solved by DRL for MEC-based ISCC systems, supported by experimental data. Additionally, the authors have reflected on the main research issues and the possible future development of IoT ISCC. Lin et al. (2024) in the case of Internet of Vehicles (IoV) conducted a study on joint user selection and bandwidth allocation for the ISAC-assisted IoV networks to optimize the sensing (by minimizing the miss detection ratio) while maintaining communication quality. They have proposed a user selection algorithm based on clustering and a DRL method that utilizes proximal policy optimization (PPO) for bandwidth allocation optimization, the results of simulations being consistent with the improved performance of both methods against the benchmarks.

Efficient resource and power allocation in V2X ISAC networks under dynamic environments was addressed in several studies. The work in Indulekha and Venkatesh (2024) presented a DRL-based dynamic allocation scheme that optimized modulation and coding scheme (MCS), subcarrier spacing (SCS), and transmission power, balancing sensing accuracy with communication reliability in 5 G V2X networks. Their strategy outperformed the conventional sensing-based semi-persistent scheduling (SB-SPS) method with notable gains in packet reception ratio (PRR) and Cramér-Rao lower bound (CRLB). Similarly, Shen et al. (2025) proposed a DRL-based approach for selecting sensing and communication functions in ISAC-enabled vehicular networks, addressing partial loss of environmental states through matrix completion and alternating direction method of multipliers (ADMM) with deep unfolding. This approach improved performance under uncertain and noisy conditions. Moreover, Liu et al. (2025) developed a joint channel, power, and bandwidth allocation framework using deep deterministic policy gradient and deep Q-network algorithms. Their DRL-based scheme enhanced spectral efficiency and communication rates by dynamically adapting to vehicle geographic positions and bandwidth availability.

The security of ISAC broadcasts in V2X networks was examined by Yao et al. (2025), who analysed a hybrid active-passive RIS-enhanced system. They suggested a safe way to send data that uses RSU beamforming, spectrum reuse, and RIS reflection to get the most out of vehicle-to-infrastructure (V2I) links while still keeping the quality of vehicle-to-vehicle (V2V) communication and radar sensing capabilities. A hierarchical twin delayed deep deterministic policy gradient (HTD3) method split the optimisation into two parts: one for allocating spectrum using a deep Q-network and the other for designing beamforming using TD3. Numerical findings showed that this strategy worked better at keeping secrets and finding radar than other ways. In a different area, Fan et al. (2024) looked at time-division ISAC (TD-ISAC) in self-driving cars and suggested a way for communication to help with sensing. The scheme effectively lowered the chances of missing a detection and the power of interference by modelling multi-vehicle signal interference as a POMDP and using a decentralised DRL algorithm for dynamic resource allocation. This was better than the baselines that didn't have help with communication or RL.

Also, beam alignment and tracking are very important for V2V networks that use millimetre waves. Zhao et al. (2024) developed an ISAC beam tracking algorithm utilising DRL to mitigate the significant latency associated with conventional tracking techniques. The algorithm combined data from sensors (such location and speed), the status of the buffer pool, and the distance between vehicles into the DRL policy to improve the accuracy of beam tracking and the amount of power used, as assessed by packet age and power consumption. The simulation results revealed that their IBTD algorithm used less power and had less delay than baseline approaches. This illustrated the benefits of merging sensing and communication tasks through RL.

Discussion: Table 8 shows recent RL/DRL-based approaches for improving ISAC in vehicular networks. The works cover different system assumptions and learning strategies but pursue the same goal: improving sensing and communication performance under different conditions. RIS and mmWave appear frequently, mainly to improve sensing resolution and to provide reliable communication links. Many studies rely on DRL variants such as DDPG, A2C, and DQN to jointly tune beamforming, power allocation, and RIS parameters, and they report gains in detection/localization accuracy and throughput. The assumed CSI (perfect versus partial) plays a central role in shaping both the learning design and the resulting system complexity. In highly dynamic multi-vehicle scenarios, multi-agent and actor-critic structures are increasingly used to improve coordination and scalability. Beyond physical-layer optimization, a smaller set of works incorporates edge AI and sensing-computation co-design to strengthen perception performance and reduce decision latency. Despite the reported improvements, several practical issues are still open. Partial observability is common in realistic deployments, convergence speed becomes critical under real-time constraints, and robustness can degrade quickly in high-mobility scenarios. These

Table 8 Summary of ISAC advances in vehicular networks using RL/DRL algorithms

Ref.	System Model	Sensing Type	Sensing Metric	CSI/SO	RL/DRL Algorithm	Opt. Variables	Objective
Shang et al. (2025)	RIS-assisted ISAC edge network	Comm., sensing fusion	SNR, detection delay	Imperfect	PPO, Cascade Net	Task offload, beamforming, power	Maximize energy efficiency, sensing
Li et al. (2025)	ISAC, vehicular edge computing	Sensing-aided fusion at RSU	Task completion delay, energy	Imperfect	Improved Dueling-TD3	Resource allocation	Minimize delay, average energy
Lai et al. (2025)	Multi-beam I2V, V2V ISAC	V2V beam sensing	Energy efficiency, SINR	Partial	MAPPO, Stackelberg game	Beam, power, frequency allocation	Maximize V2V efficiency, reduce interference
Long et al. (2024)	RIS-assisted 6 G V2X	Joint communication, positioning	3D FIM (position), achievable rate	Imperfect	Flexible-DDPG (FL-DDPG)	RIS phase shift, power allocation	Minimize position error, improve communication
Liu et al. (2024)	ISAC-assisted V2I downlink	Angle	CRLB (angle, distance)	Partial	DRL (DDPG-like)	Beamforming, power allocation	Maximize sum-rate, meet CRLB
Yang et al. (2024)	ISCC: ISAC, over-the-air FL	Sensing, FL, communication	Beamforming accuracy, FL delay	Partial	Hybrid RL (DDPG, SDR, Gaussian randomization)	Beamforming, power allocation	Maximize rate, ensure FL and sensing
Ren et al. (2024)	ISCC: ISAC, MEC, AirComp, FEEL	Varies (MEC)	Delay, accuracy	Varies	DRL (DQN, actor–critic)	Resource allocation	Joint delay–accuracy trade-off
Lin et al. (2024)	ISAC-IoV in SUMO environment	Vehicle detection	Miss detection ratio	Dynamic	PPO (DRL)	Bandwidth allocation	Minimize miss detection, balance communication
Indulekha and Venkatesh (2024)	5 G V2X with ISAC	Radar	CRLB, CBR, PRR	Dynamic	DRL (DDPG-based)	MCS, SCS, transmit power	Maximize communication, sensing performance
Shen et al. (2025)	ISAC-enabled AVs with partial environment state	SC function selection	Detection accuracy	Partial (POMDP)	DRL, ADMM, deep unfolding	SC function selection	Maximize sensing under partial observability
Liu et al. (2025)	ISAC in spectrum-constrained IoV	Radar sensing	Bandwidth use, reliability	Dynamic	DDPG, DQN	Channel, power, bandwidth allocation	Maximize communication reliability
Yao et al. (2025)	Hybrid RIS-assisted V2X ISAC	Radar, secure communication	Secrecy rate	Partial	HTD3 (hierarchical TD3, DQN)	Beamforming, spectrum reuse	Maximize secrecy rate, enable sensing

Table 8 (continued)

Ref.	System Model	Sensing Type	Sensing Metric	CSI/SO	RL/DRL Algorithm	Opt. Variables	Objective
Fan et al. (2024)	TD-ISAC in AVs	Radar (TD-ISAC)	Miss detection, interference	POMDP	DRL (decentralized)	Time-frequency resources	Reduce miss detection, interference
Zhao et al. (2024)	ISAC for mmWave V2V beam tracking	Motion tracking	Packet age, power use	Partial	DRL (IBTD)	Power, beam alignment, resource allocation	Minimize latency, improve tracking accuracy

Ref. - Reference, CSI - Channel State Information, Opt. - Optimization, CRLB - Cramér–Rao Lower Bound, FL - Federated Learning, SO - State Observability

points suggest that future ISAC-enabled vehicular designs will likely need more adaptive (and possibly hybrid) learning structures.

4.4 ISAC in other 6 G technologies and cross-layer architectures

Recent research has put forward various DRL-driven frameworks to address ISAC difficulties in contexts beyond UAVs. A twin-delayed deep deterministic policy gradient (TD3) approach is created in Yang et al. (2025) to help with resource allocation in RIS-assisted ISAC networks. The agent designs sensing beamforming and communication activities at the same time, taking into account power and QoS limits. The findings show that both throughput and target detection accuracy have improved. The authors incorporate a channel-uncertainty-aware training technique to more accurately represent real-world RIS deployments, where inadequate channel knowledge can diminish learnt policies; in such scenarios, the suggested method demonstrates superiority over baseline heuristic solutions.

Similarly, Mamaghani et al. (2025) introduces a federated DRL framework for spectrum access and sensing coordination, where multiple base stations learn collaboratively without exchanging raw data. This design is intended to preserve privacy while still benefiting from distributed experience. The reward construction accounts for spectrum utilization, sensing reliability, and inference delay, and the evaluation suggests that the learned policy adapts to fast-changing traffic patterns while reducing training overhead in heterogeneous networks. In Zhang et al. (2025), a hybrid DRL optimization approach is presented for allocating radar and communication resources in vehicle-mounted ISAC units. Using proximal policy optimization (PPO), the policy is trained to improve sensing accuracy while maintaining communication performance in time-varying vehicular conditions. Heuristic scheduling is combined with the DRL policy to speed up learning and make policy updates feasible in real time, and the reported results remain favorable under high-speed mobility and multipath fading. Finally, Farzanullah et al. (2025) proposes a multi-agent DRL (MADRL) framework for joint task offloading and beamforming in edge-assisted ISAC to reduce both latency and energy consumption. To improve scalability in vehicular edge networks, the agents exchange information through a topology-aware coordination mechanism, and the strongest gains are observed in dense urban deployments with frequent handovers and pronounced multi-user interference.

In Nikbakht et al. (2024), an actor-critic method is developed to learn time-critical packet scheduling and waveform adaptation for real-time mmWave ISAC. Simulation results under realistic channel models indicate faster learning and stronger interference suppression. The agent's training also injects expert knowledge via imitation learning, which helps stabilize and refine the policy in the early training phase. In addition, the learned controller can switch its emphasis between communication-oriented and sensing-oriented allocations as system demands change. To improve adaptability across operating conditions, Chen et al. (2025) proposes a meta-learning DRL approach that adjusts the sensing-communication trade-off when network conditions vary. The stated advantages are especially noticeable in transfer contexts, when the policy has to work with variable node densities and mobility regimes. This is in contrast to normal DRL baselines. The meta-update uses a first-order approximation, which lets the policy change quickly with only a little retraining. In smart city situations, Hung et al. (2025) uses a PPO-based architecture to regulate both the paths of vehicles and the alignment of ISAC beams at the same time. The model adapts to changes in traffic while lowering latency and sensing errors. It also has a predictive control feature that makes it more sensitive to impediments and mobility events. The findings endorse the application of multi-task DRL formulations in highly dynamic urban implementations.

In addition to terrestrial environments, Ouyang et al. (2024) examines DRL-based beam selection and resource scheduling for satellite-terrestrial ISAC, aiming to enhance coverage and sensing accuracy within Doppler and latency limitations. The agent learns beam planning strategies that take into account geometry by adding orbital characteristics to the observation space. This makes the agent more resilient when links degrade or when the signal-to-noise ratio is low.

From a cloud-edge perspective, Wang et al. (2024) presents a federated actor-critic architecture for ISAC-aware task migration in dispersed networks, targeting low-latency replies and dynamic load balancing. There is a knowledge-distillation step that compresses shared models and cuts down on synchronisation costs. Experiments show that this leads to faster convergence and lower migration costs than non-federated baselines. To lower the cost of exploration and make things more stable, Smeenk et al. (2024) combines DRL with digital-twin modelling. This uses a virtual copy of the wireless environment that is always being updated to help with proactive sensing and strong beam design. The results indicate that the connection between the DRL and the digital twin becomes increasingly robust against hardware imperfections and channel variations as the twin is continuously refined with actual data. For multi-user mmWave ISAC, Zhai et al. (2023) creates a PPO-based controller that changes the beam directions and transmit power based on how users move and how the channel changes, which makes the system work more efficiently and more reliably when there is blockage or severe path loss. Finally, Shen et al. (2024) presents a DDPG-driven framework for latency-aware task scheduling and beamforming in vehicular ISAC utilising a hybrid RSU-UAV deployment. The utility formulation clearly takes into account job deadlines, network congestion, and mobility prediction. The results indicate that offloading delays are shorter while still allowing for real-time sensing and communication.

The work in Xie et al. (2025) studies federated multi-agent DRL (F-MADRL) for privacy-preserving optimization in smart-factory ISAC settings. The agents learn sensing-communication trade-offs collaboratively while keeping local data on-site and maintaining inference accuracy. The reported results indicate scalability to large industrial deployments and improved resilience against adversarial behavior. In Li et al. (2024), an energy-aware

ISAC design is considered, where a SAC agent jointly controls power allocation and waveform parameters. Battery limitations and mission-critical sensing requirements are explicitly modeled for mobile-robot networks, and the evaluation shows longer operating time without sacrificing sensing reliability. Finally, Mason and Pegoraro (2025) proposes a cloud-native DRL framework for satellite ISAC supported by a digital twin. The agent uses real-time satellite telemetry in combination with a virtual training environment to learn how to proactively tune transmission and sensing parameters. The findings indicate that there was a higher energy efficiency without compromising on the continuous coverage over LEO constellations.

On the whole, these papers demonstrate that DRL is not limited to only aerial platforms in ISAC scenarios but can significantly increase the complexity of the environment, network size, and control objectives. The increasing popularity of multi-agent learning, federated training, and digital twin-assisted optimization points to a transition to systems that can make coordinated distributed decisions, keep data local, and minimize the cost of online exploration. These directions naturally extend ISAC research into vehicular, satellite, and edge-cloud deployments, where mobility, delay, and resource coupling are more pronounced. Looking ahead, tighter coupling with cross-layer optimization and emerging foundation-model tools (including large language models) may provide a practical path to jointly orchestrate sensing, communication, and computation under mission-critical constraints.

Discussion: Table 9 provides 15 recent works that apply DRL to ISAC across different deployment settings. The studies span RIS-assisted architectures, vehicular scenarios, satellite-terrestrial links, and smart-factory environments, indicating that DRL can be adapted to heterogeneous channel conditions, network structures, and hardware constraints. Across these contributions, actor-critic families such as TD3, PPO, DDPG, and SAC are used most often, largely because they handle continuous actions and high-dimensional control typical of ISAC (e.g., beamforming, power control, waveform selection, and scheduling).

Several studies extend these baseline DRL algorithms with different system-level designs, most notably federated learning (Mamaghani et al. 2025; Wang et al. 2024; Xie et al. 2025) to enable collaborative training without exposing local data, and digital-twin Smeenk et al. (2024); Mason and Pegoraro (2025) for real-time emulation. Meta-learning Chen et al. (2025) and federated multi-agent schemes Xie et al. (2025) further point to a shift toward solutions that scale with network size, transfer more reliably across operating conditions, and remain robust in the presence of adversarial or uncertain environments. In terms of outcomes, the table reports recurring improvements in throughput, sensing accuracy, energy efficiency, and latency, which support the use of DRL for real-time ISAC adaptation. For example, Yang et al. (2025); Ouyang et al. (2024) associate throughput and coverage gains with geometry-aware beam control, whereas Li et al. (2024); Mason and Pegoraro (2025) focus on energy-aware configurations in mobile and satellite platforms. Many of these gains were directly related to the domain-specific additions, such as mobility prediction Shen et al. (2024), knowledge distillation Wang et al. (2024), and beam-aware policy structures Zhai et al. (2023), which serve to lessen the learning burden and increase stability under mobility and blockage.

Finally, multi-agent learning coupled with cloud-edge coordination signifies a move toward more decentralized intelligence in future 6 G networks. The research Farzanullah et al. (2025); Xie et al. (2025) suggests that distributed policy learning combined with limited controlled synchronization can perform well in dense time-varying scenarios, which

Table 9 Summary of ISAC advances in other 6 G technologies using RL/DRL algorithms

Reference	System Model	Algorithm Used	Key Feature	Performance Metrics
Yang et al. (2025)	RIS-assisted ISAC	TD3	Channel-uncertainty-aware training	↑ Throughput, ↑ detection accuracy (vs. heuristic)
Mamaghani et al. (2025)	Federated ISAC with base stations	Federated DRL	Privacy-preserving policy coordination	↑ Privacy, ↓ training overhead
Zhang et al. (2025)	Vehicular ISAC	Hybrid DRL, PPO	Heuristic scheduling, DRL refinement	↑ Sensing accuracy, ↓ convergence time under mobility
Farzanullah et al. (2025)	Edge-assisted ISAC	MADRL	Topology-aware coordination	↓ Latency, ↓ energy cost in dense networks
Nikbakht et al. (2024)	mmWave ISAC	Actor–Critic	Imitation-learning-based waveform adaptation	↑ Convergence speed, ↓ interference, adaptive switching
Chen et al. (2025)	General ISAC with mobility	Meta-DRL	Fast adaptation via meta-policy	↑ Transferability, ↓ retraining time
Hung et al. (2025)	Smart city vehicle ISAC	PPO	Predictive mobility-aware control	↓ Latency, ↓ sensing error, responsive control
Ouyang et al. (2024)	Satellite-terrestrial ISAC	DRL	Geometry-aware beam planning	↑ Coverage, ↑ sensing under Doppler and delay
Wang et al. (2024)	Cloud-edge distributed ISAC	Federated Actor–Critic	Knowledge distillation for compression	↓ Latency, ↑ load balance, ↓ migration cost
Smeenk et al. (2024)	Digital twin ISAC	DRL	Real-time model updates	↑ Robustness to hardware impairments, ↓ exploration cost
Zhai et al. (2023)	mmWave ISAC with mobility	PPO	Beam-aware control	↑ Spectral efficiency, ↑ robustness in blockage
Shen et al. (2024)	Vehicular ISAC with RSU-UAV	DDPG	Latency-aware task scheduling	↓ Offloading delay, ↑ safety in traffic
Xie et al. (2025)	Smart factory ISAC	F-MADRL	Privacy-preserving multi-agent learning	↑ Privacy, ↑ inference accuracy, ↓ adversarial risk
Li et al. (2024)	Mobile robot ISAC	SAC	Energy-efficient waveform design	↑ Energy efficiency, ↑ operational time
Mason and Pegoraro (2025)	Satellite ISAC with digital twin	DRL	Cloud-native training and telemetry	↑ Global coverage, ↑ energy efficiency, proactive control

is especially relevant to URLLC and mission-critical sensing. Overall, the table reveals a noticeable trend away from standalone DRL methods toward more realistic configurations that take into account the limitations of actual systems, privacy issues, and real deployment. It emphasizes the cross-layer learning and collaborative intelligence to maintain a good level of performance in everyday situations.

5 Research challenges, future directions, and lessons learned

The section reveals the key challenges of employing RL/DRL in the ISAC paradigm, discusses prospective research directions, and finally presents the main lessons that could work in a rapidly developing research field.

5.1 Current challenges

Despite the fact that RL/DRL techniques have achieved promising results in ISAC systems, various challenges still hinder their real-world deployment. These challenges are caused not only by the fundamental drawbacks of learning-based algorithms but also by the specific demands of ISAC, e.g., low delay, high reliability, and joint optimization of communication and sensing. In the following, key challenges are explained.

5.1.1 Computational complexity and latency

Training and inference of the RL/DRL methods is one of the biggest hurdles in their use for ISAC because of the high computational complexity. Usually, DRL algorithms involve the design and training of a large NN with millions of parameters. The training of the model requires extensive interactions with the environment to test different actions through trial and error which makes it computationally intensive and time-consuming. The training can be done offline with sufficient resources, but deploying these models in real-time ISAC systems is not feasible because of extremely low-latency requirements. For instance, a base station or ISAC-enabled radar system might have to make decisions within a sub-millisecond time frame, which is practically impossible when inference needs to be done from a large NN. The problem is further worsened by hardware constraints of mobile or edge devices, as such equipment usually does not have powerful GPUs or TPUs for fast computations. Therefore, the main issue is still how to drastically cut down the computational expense of the RL/DRL algorithms while maintaining their level of accuracy.

5.1.2 Exploration exploitation tradeoff

Another challenge arises from the basic exploration-exploitation tradeoff in RL. In ISAC systems, exploration means developing new policies that not only enhance communication but also sensing performance, while exploitation refers to using the current best policy to achieve the highest immediate performance. A major hurdle in establishing such a balance is in dynamic environments where channel conditions, user mobility, and sensing tasks change rapidly. Excessive exploration can degrade the system performance, resulting in unreliable communication or inaccurate sensing; on the other hand, greedy exploitation means missing out on the discovery of better long-term strategies. Furthermore, exploration incurs extra overhead since the system needs to use some of its resources to carry out suboptimal actions to gather information. The use of RL/DRL methods for efficient exploration without compromising the stability of ISAC performance remains a pressing issue.

5.1.3 Scalability to large-scale systems

Scalability is also a significant barrier to the widespread applicability of RL/DRL in ISAC. Next-generation wireless networks are characterized by a vast number of users, antennas, and sensing targets, and all of these elements need to be optimized together. Consequently, the state and action spaces become so large and are referred to as 'the curse of dimensionality'. Under these circumstances, the conventional RL methods either fail to converge or the training time becomes practically impossible. While DRL lessens the problem to some extent by using function approximation, the problem still remains, as the complexity of ISAC scenarios keeps increasing. For example, multi-antenna ISAC systems may require the optimization of beamforming vectors that are spread over hundreds of antennas, and multi-user networks require resource allocation among tens or hundreds of users. If there are no scalable algorithms, it means that RL/DRL methods cannot be successfully utilized in real-world ISAC deployments.

5.1.4 Generalization to unseen environments

Generalization is another major challenge. Policies trained in a specific environment will not serve the purpose when the setting is new and unseen. In ISAC, this issue is more severe because radio channels and sensing conditions can change significantly across time, frequency, and space. A DRL model, for instance, trained in a typical suburban environment will most likely fail to deliver in urban or high-mobility scenarios. ISAC applications often involve mission-critical tasks, where poor generalization can lead to significant performance degradation. Current DRL models tend to overfit the training data, which reduces their robustness when the operating environment deviates from the training assumptions. The absence of generalization points to the need for adaptive algorithms that could adapt to the variations and dynamics of ISAC environments.

5.1.5 Security and robustness issues

RL/DRL algorithms, by nature, are exposed to security risks as well as model uncertainties. For ISAC, adversarial perturbations, or jamming attacks, can mislead a trained model and result in wrong resource allocation, low sensing accuracy, or communication outages. The stochastic nature of wireless channels adds another layer of uncertainty, making it impossible for policies to consistently deliver a good performance. DRL robustness could also be compromised by hardware impairments, such as non-linear amplifiers or quantization errors. Therefore, it is a major challenge to ensure the security and robustness of RL/DRL algorithms against adversarial attacks, mismatches of models, and uncertainties in the environment of ISAC systems.

5.1.6 Data efficiency and sample complexity

Finally, RL/DRL algorithms usually are required to interact with the environment thousands of times before they can identify a promising policy. So, obtaining such training data in ISAC systems could be challenging. Real-world ISAC systems, unlike in simulation, take no advantage of granting excessive resources or time for exploration. Data efficiency is a major concern for safety-critical applications, such as vehicular networks or defense systems, where exploring poor policies may lead to catastrophic consequences. The huge data requirement forbids the use of RL/DRL in ISAC from a practical point of view and calls for the development of methods that can be adequately trained on limited data.

5.2 Future research directions

Some promising research directions can be envisaged to address the challenges mentioned above and to harness the potential of RL/DRL in ISAC systems fully. These research directions aim to increase scalability, enhance adaptability, and guarantee robustness without losing practical feasibility. A brief description of the main avenues for future work is provided below.

5.2.1 Lightweight and distributed RL/DRL

An essential research topic is designing efficient, lightweight RL/DRL algorithms by reducing the computational complexity and memory requirements. Model compression methods, such as pruning, quantization, and knowledge distillation, might be used to reduce the NN size without a notable loss in performance. Besides that, distributed learning setups, such as federated RL, may be investigated for ISAC scenarios in which several nodes jointly train models without centralizing the raw data. Not only does that cut down the communication burden, but it also enhances data privacy and scalability. It will be very important to come up with such algorithms that can be executed on edge devices having restricted hardware capabilities for the real-time ISAC deployment.

5.2.2 Multi-agent RL for ISAC

Because ISAC can include various cooperating nodes, such as base stations, radars, UAVs, and vehicles, MARL is really a great tool for these interacting partners. In MARL, each agent independently decides its actions from the limited local observations, but at the same time, the agents coordinate their actions to maximize the overall system utility. That is why this idea fits so well with ISAC cases, such as cooperative beamforming, joint target tracking, or spectrum sharing among heterogeneous nodes. Nonetheless, MARL comes with its own set of challenges, e.g., non-stationarity of the environment and agent-to-agent communication overhead. It is recommended for future studies to look into how to develop scalable MARL algorithms that can guarantee stable learning and efficient coordination in very large ISAC networks.

5.2.3 Transfer learning, meta-learning, and split learning

Transfer learning, meta-learning, and split learning can improve the generalization capability and practical deployment of RL/DRL policies in ISAC systems. Transfer learning enables a model trained in one scenario to be adapted to another related scenario, thereby reducing the time and data required for retraining. This is particularly useful in ISAC environments, where channel conditions, sensing requirements, and network configurations may vary significantly across deployments. Meta-learning, in contrast, aims to train models that can adapt rapidly to new tasks or environments using only a small amount of additional data. For example, a meta-trained policy can quickly adjust to changes in channel conditions, user mobility, or sensing demands without requiring training from scratch.

In addition, split learning offers a promising privacy-aware and computation-efficient framework for future ISAC systems, especially in satellite, edge-assisted, and distributed communication settings Sun et al. (2024). In split learning, model training is partitioned across different network entities, so that raw local data do not need to be fully shared with a central node. This characteristic makes split learning attractive for ISAC scenarios involving privacy-sensitive sensing data, limited onboard processing resources, or distributed intelligence across terrestrial and non-terrestrial platforms. Therefore, integrating transfer learning, meta-learning, and split learning with RL/DRL can significantly improve the flexibility, scalability, privacy preservation, and resilience of future ISAC systems.

5.2.4 Joint optimization frameworks

Additionally, upcoming studies should be oriented towards creating unified optimization paradigms that integrate the communication and sensing goals as a whole instead of stressing them as separate issues. The majority of the work done so far aims at one goal: either increasing communication throughput or enhancing sensing accuracy Khan et al. (2024, 2023). But ISAC, by its very nature, is about compromise. RL can be an effective solution to multi-objective optimization problems. Designing reward functions that include both sensing and communication metrics can be the means to policies that result in synergistic performance gains. In addition, the optimization of RL/DRL-based ISAC solutions can be extended to the consideration of control objectives such as energy efficiency, latency, and reliability, which will contribute to their feasibility and consistency with the requirements of the real world.

5.2.5 Secure and robust learning

One of the main goals of the research agenda is to make sure that RL/DRL algorithms that work in ISAC systems are strong and safe. Adversarial machine learning is when the opponent modifies the environment or the input data on purpose to confuse the learner. This makes the learning process a target for these kinds of attacks. In these situations, mission-critical ISAC deployments could have very bad effects. So, safe-RL approaches need to be put into place to make sure that the minimum performance is fulfilled, together with strong RL models that can keep working even when things go wrong. You may also conceive of agent uncertainty as a part of the model and utilise it to make decisions that will protect you against changes in the channel, hardware problems, and mistakes in the model. To put it another way, a truly safe ISAC will need to have adaptive algorithms that can work well even when the environment is hostile or unpredictable.

5.2.6 Task-centric ISAC for edge AI

Future RL/DRL-aided ISAC research should increasingly shift from optimizing link-level metrics (e.g., rate/SINR) to task-level objectives that directly reflect edge-AI performance, such as inference accuracy, end-to-end latency, energy consumption, and robustness under mobility and intermittency (He et al. 2024; Yao et al. 2025; Wen et al. 2024). In multi-device edge AI, this calls for joint decision-making over (i) sensing configuration (waveform/beam/pulse scheduling and sensing-update rates), (ii) communication resource allocation (bandwidth, power, access, and retransmissions), and (iii) computation orchestration (task offloading, split inference, model selection/compression, and scheduling across heterogeneous edge devices). RL/DRL can provide a principled framework for learning cross-layer policies under coupled constraints (freshness of sensed data, compute budgets, queueing delays, and reliability), while enabling collaboration among devices via cooperative perception, feature/gradient sharing, or semantic/task-oriented transmissions.

5.2.7 Integration with emerging technologies

Another exciting direction is the integration of RL/DRL for ISAC with emerging wireless technologies. For example, RIS and holographic surfaces can provide additional degrees of freedom in controlling the propagation environment, which RL/DRL can exploit for joint sensing and communication optimization Khan et al. (2025). Similarly, RL/DRL can help ISAC in NTNs, such as satellite and HAPS-assisted systems, that have changing topologies and long propagation delays Khan et al. (2022). Additionally, semantic communications, which transmit meaningful information instead of raw data, can be used with RL/DRL for ISAC to achieve efficient and intelligent resource allocation. Examining the relationship between RL/DRL and these emerging technologies offers numerous opportunities for future research.

5.2.8 Benchmarking and standardized environments

The need for standardized benchmarking environments and datasets for RL/DRL in ISAC is a useful but largely ignored area of research. Most works consider simulation setups specific to their own system models. Therefore, it is difficult to compare different methods properly. The development of open-source platforms and standardized assessment measures will help to reproduce results and make fair comparisons and also speed up progress in this field. In addition, large-scale datasets that show realistic ISAC scenarios will aid supervised pretraining of DRL models. They make the online training less complicated. The setting of guidelines and standards will assist close the gap between theoretical research and practical deployment.

5.3 Lessons learned

The thorough study of RL/DRL applications in ISAC might give useful information that can help both researchers and professionals. These lessons show what RL/DRL works well, what has to be done, and what the rules are for making future ISAC systems.

5.3.1 RL/DRL is powerful but not a universal solution

RL/DRL provides adaptive, model-free optimisation in dynamic ISAC contexts. But it can't be used as a one-size-fits-all answer. It still has to be compared to standard optimisation methods like semidefinite relaxation, iterative heuristics, and convex programming. Each of these strategies has its own pros and cons. When accurate system models are available, traditional algorithms work effectively for optimisation tasks that are low-dimensional and well-structured. They frequently have strict maths, quick convergence, and conclusions that are straightforward to understand. In beamforming design scenarios, such as static or slowly time-variant channels, closed-form methods can effectively meet power budget and CRB restrictions without necessitating extensive training data. This is not the case in high-dimensional, non-convex optimisation contexts that involve mobility, CSI uncertainty, and environmental uncertainties, where conventional methods frequently depend on poor approximations that may degrade performance. RL/DRL is also model-free, can optimise things online, and can handle more than one goal at a time without needing to be reformulated Xie et al. (2025). In numerous studied studies, RL/DRL algorithms have exhibited considerable performance enhancements compared to traditional optimisation methods, particularly in dynamic and high-dimensional contexts Cho et al. (2024); Chen et al. (2024). But this comes with a lot of computational complexity, a lot of time spent on training, and sample inefficiency during exploration. There are also no assurances of optimality or stability, which can lead to policies that aren't optimal or stable in safety-critical ISAC applications Ye et al. (2025).

A balanced perspective endorses hybrid strategies: traditional techniques for initialisation or straightforward subproblems, and RL/DRL for advanced adaptation in complicated and uncertain scenarios. This shows how useful RL/DRL could be as a complement to existing methods, especially in the complicated 6 G ISAC scenario. In cases where the system model is well understood and manageable, classical optimisation or heuristic methods may still be effective. So, RL/DRL should be seen of as tools that function well with other methodologies, not as ways to replace them.

5.3.2 Cross-disciplinary design is essential

For RL/DRL to perform successfully in ISAC, it needs to be applied in a method that combines communication theory, signal processing, and machine learning. Data-driven approaches often fail to perform effectively with wireless channels and sensor signals due to their disregard for certain physical-layer characteristics. RL/DRL algorithms can converge faster, be easier to understand, and work more reliably if they use domain expertise while designing state representations, reward functions, and learning architectures. This indicates that forthcoming ISAC research ought to concentrate on hybrid methodologies that integrate model-based optimisation with data-driven learning.

5.3.3 Industry perspectives on RL/DRL in ISAC

Though the surveyed literature is more academic in nature, industry is driving ISAC for 6 G, with RL/DRL convergence for real-time optimization. Huawei showed an ISAC-THz prototype operating at 140 GHz for millimeter resolution sensing (e.g., 3D imaging through obstacles) and 220 GHz for 240 Gbit/s communications, employing compressed sensing for signal processing. This translates theoretical possibilities into practical areas, such as smart cities, medical, and security applications, though no direct application of RL is mentioned; convergence with DRL could be explored in the next phases for waveform adjustment purposes Li et al. (2022). Ericsson has applied RL in 5 G RAN optimization with an improvement in throughput and a reduction in power consumption, and is extending this to 6 G ISAC for tasks such as object detection, environmental mapping, and massive mixed reality. Collaborative efforts, for example, Technology Innovation Institute (TII)-funded *"MARL for UAV Swarm ISAC,"*, bridge industry and academia for optimizing trajectories for improved sensing SNR in bistatic systems Atsu et al. (2025). Overall, industry efforts, such as Keysight's testbed for OFDM ISAC, focus on hardware implementation and standardization (3GPP Release 18+), while RL/DRL shows promise for addressing dynamic 6 G networks. Trials of RL-ISAC, on the other hand, are still limited.

5.3.4 Synergy between communication and sensing

A key insight from the literature is that the tradeoff between communication and sensing should not always be treated as a zero-sum game. With carefully designed RL/DRL policies, synergistic gains can be achieved, whereby advancements in one domain also enhance performance in the other. For example, improved sensing accuracy can provide accurate channel status information, which improves communication performance. On the other hand, high-quality communication links can support distributed sensing tasks in a more effective manner. Recognizing and exploiting these synergies can initiate new opportunities for joint optimization in ISAC.

5.3.5 Robustness and scalability are as important as performance

Most of the work is on performance measures, like how fast you can get a job done or how accurate your sensors are. But for real-world ISAC deployment, stability and scalability are just as important. For mission-critical applications, a DRL strategy that functions well but breaks when the environment changes is not useful. Similarly, algorithms that function effectively in small-scale systems may not perform optimally in environments with numerous users, antennas, or sensing targets. Researchers must strike a balance among performance, resilience, and scalability. This means that their solutions have to perform well in a wide range of settings and changing environments.

5.3.6 Bridging theory and practice remains a challenge

Finally, a critical point to remember is that most of the studies are talking about work at the level of concepts shown through simulations. Simulations are essential for algorithm validation, while typically depending on a framework of ideal assumptions, such as channel models, hardware limitations, or mobility patterns. Integrating theoretical research with practical application remains a challenge. More realistic simulation frameworks and experimental testbeds must be developed, and closer collaboration among standards bodies, industry, and academia is essential. For RL/DRL-based ISAC systems to work in real life, everyone needs to work together like this.

6 Conclusion

This survey provided an overview of RL/DRL as enabling tools for ISAC systems. It connected fundamental ISAC concepts, including joint waveform design, sensing tasks (estimation, detection, and recognition), and radar architectures, with representative RL/DRL algorithms. Particular attention was given to how these algorithms could operate in model-free settings and support optimization under imperfect or unavailable CSI. Key methods included value-based algorithms (e.g., DQN and DDQN), policy-gradient and actor–critic methods (e.g., A2C and DDPG), and multi-agent approaches (e.g., MADDPG and QMIX). Recent developments were classified according to major application scenarios, including RIS-assisted systems, UAV/satellite systems, vehicular networks, and emerging 6 G paradigms. The survey also presented new taxonomies, algorithmic comparisons, and performance-oriented summaries. Major challenges included computational latency, the exploration–exploitation trade-off, scalability, security, and generalization. Promising future directions included lightweight distributed RL, meta-learning, holographic MIMO, semantic communications, and task-oriented edge intelligence. Overall, the reviewed literature indicated that RL/DRL could effectively address uncertainty through domain-aware hybrid modeling, joint sensing–communication optimization, robustness-oriented design, and closer integration between theory and practice. The convergence of RL and DRL was expected to play an important role in enabling self-optimizing 6 G networks with reliable connectivity and pervasive sensing for smart cities and autonomous systems.

Author contributions Conceptualization: WUK, MA; Methodology: WUK, WK, MI, CSK, MA, STS; Visualization: WUK, MA; Supervision: WUK; Writing original draft: WUK, WK, MI, MA; Writing review & editing: WUK, WK, MI, CFK, MA, STS.

Funding No funding was received for this manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

Ade Krisna Respati M, Lee BM (2024) A survey on machine learning enhanced integrated sensing and communication systems: architectures, algorithms, and applications. IEEE Access 12:170 946-170 964

Adhikary A, Raha AD, Qiao Y, Munir MS, Halder M, Hong CS (Mar. 2024) "Age of sensing empowered holographic ISAC framework for nextg wireless networks: a VAE and DRL approach," TechRxiv

Ahmed M, Xu F, Lyu Y et al (2025) RIS-driven resource allocation strategies for diverse network environments: A comprehensive review. Trans Emerg Telecommun Technol 36(6):e70160

AlKhonaini A, Sheltami T, Mahmoud A, Imam M (2024) UAV detection using reinforcement learning. Sensors 24(6):1870

Allu R, Katwe M, Singh K, Shin H (Jul. 2025) "Energy efficiency optimization for integrated sensing and communications-aided full duplex MIMO system with imperfect CSI," IEEE Transactions on Vehicular Technology, pp. 1–14

Alsheikh MA, Hoang DT, Niyato D, Tan H-P, Lin S (2015) Markov decision processes with applications in wireless sensor networks: a survey. IEEE Communications Surveys & Tutorials 17(3):1239–1267

Amhaz A, Elhattab M, Sharafeddine S, Assi C (2024) UAV-assisted NOMA for enhancing ISAC: a deep reinforcement learning solution. IEEE Commun Lett 29(2):249–253

Atsu OM, Naoumi S, Bomfin R, Chafii M (2025) “Reinforcement learning for enhancing sensing estimation in bistatic ISAC systems with UAV swarms,”. [Online]. Available: https://arxiv.org/abs/2501.06454
Buşoniu L, Babuška R, De Schutter B (2010) “Multi-agent reinforcement learning: an overview,” Innovations in multi-agent systems and applications-1, pp. 183–221
Chen M, Shu F, Zhu M, Wu D, Yao Y, Zhang Q (2025) Reinforcement-learning-based AAV 3-D target tracking and digital-twin-assisted collision avoidance with integrated sensing and communication. IEEE Internet Things J 12(13):24 916-24 928
Chen Y, Yang H, Ou X, Jiang Y, Xiong Z (2025) Anti-jamming resource allocation for integrated sensing and communications based on game-guided reinforcement learning. IEEE Wireless Communications Letters 14(1):223–227
Chen X, Cao X, Xie L, He Y (2024) “DRL-based joint trajectory planning and beamforming optimization in aerial RIS-assisted ISAC system,” in. IEEE International Workshop on Radio Frequency and Antenna Technologies (iWRF&AT) 2024:510–515
Chen Z, Huang L, So HC, Jiang H, Zhang XY, Wang J (Dec. 2024) “Deep reinforcement learning over RIS-assisted integrated sensing and communication: Challenges and opportunities,” IEEE Veh. Technol. Mag
Chen Z, Yang P, Wei J, Duan R, Xiao Y, Wu G, Zammit S (2024) “Secure UAV-aided reinforcement learning for integrated sensing and communication,’ in. International Conference on Future Communications and Networks (FCN) 2024:1–6
Chien TV, Tu LT, Khalid W, Yu H, Chatzinotas S, Renzo MD (2024) RIS-assisted wireless communications: long-term versus short-term phase shift designs. IEEE Trans Commun 72(2):1175–1190
Cho H, Yoo S, Jung BC, Kang J (2024) “Enhancing battlefield awareness: an aerial ris-assisted isac system with deep reinforcement learning,” in MILCOM 2024 - 2024 IEEE Military Communications Conference (MILCOM), pp. 469–474
Chung J (2013) Playing atari with deep reinforcement learning. Comput Ence 21(351–362):2
Clouse JA (1996) “Learning from an automated training agent,” Adaptation and Learning in Multiagent Systems. Springer Verlag, Berlin, p. 195
Cui Q, You X, Wei N, Nan G, Zhang X, Zhang J, Lyu X, Ai M, Tao X, Feng Z et al (2025) Overview of AI and communication for 6G network: fundamentals, challenges, and future research opportunities. Sci China Inf Sci 68(7):171301
Ding R, Zhou F, Giordani M, Wu Q, Zorzi M (2026) “Toward 6G: spectrum management in low-altitude aerial information networks,” IEEE Network, pp. 1–9
Drampalou SF, Uzunidis D, Vetsos A, Miridakis NI, Karkazis P (Dec. 2024) “A user-centric perspective of 6G networks: a survey,” IEEE Access, 12:190 255–190 294
Du Z, Liu F, Li Y, Yuan W, Cui Y, Zhang Z, Masouros C, Ai B (2025) Toward ISAC-empowered vehicular networks: framework, advances, and opportunities. IEEE Wirel Commun 32(2):222–229
Fan Y, Fei Z, Huang J, Wang X (2024) Reinforcement learning-based resource allocation for multiple vehicles with communication-assisted sensing mechanism. Electronics 13(13):2442
Fanni SC, Febi M, Aghakhanyan G, Neri E (2023) “Natural language processing,” in Introduction to artificial intelligence. Springer, pp. 87–99
Farhadi A, Olfat A, Masouros C (Jan. 2025) “Joint beamforming and resource allocation for STAR-IRS-aided SCMA ISAC systems using meta deep reinforcement learning,” IEEE Transactions on Wireless Communications, pp. 1–1
Farzanullah M, Zhang H, Sediq AB, Afana A, Erol-Kantarci M (Mar. 2025) “Beam selection in ISAC using contextual bandit with multi-modal transformer and transfer learning,”
Ferrari E, Cavalcanti D, Frascolla V, Morato A, Vitturi S, Cenedese A (2025) “Wireless time-sensitive networking for real-time unmanned ground vehicle control and mapping,” in 2025 IEEE 30th International Conference on Emerging Technologies and Factory Automation (ETFA), pp. 1–4
Gao Q, Zhong R, Shin H, Liu Y (2024) MARL-based UAV trajectory and beamforming optimization for ISAC system. IEEE Internet Things J 11(24):40 492-40 505
Gong J, Gao D, Tian R, Liu X, Peng M (Feb. 2025) “Frame structure design and task scheduling in 5G NR for ISAC systems: a deep reinforcement learning approach,” IEEE Transactions on Vehicular Technology, pp. 1–15
González-Prelcic N, Furkan Keskin M, Kaltiokallio O, Valkama M, Dardari D, Shen X, Shen Y, Bayraktar M, Wymeersch H (2024) The integrated sensing and communication revolution for 6G: vision, techniques, and applications. Proc IEEE 112(7):676–723
Haarnoja T, Zhou A, Hartikainen K, Tucker G, Ha S, Tan J, Kumar V, Zhu H, Gupta A, Abbeel P et al (2018) “Soft actor-critic algorithms and applications,” arXiv preprint arXiv:1812.05905
He Y, Yu G, Cai Y, Luo H (2024) Integrated sensing, computation, and communication: system framework and performance optimization. IEEE Trans Wireless Commun 23(2):1114–1128

Hou P, Huang Y, Zhu H, Lu Z, Huang S-C, Yang Y, Chai H (2025) Distributed DRL-based integrated sensing, communication, and computation in cooperative UAV-enabled intelligent transportation systems. IEEE Internet Things J 12(5):5792–5806

Hou P, Zhu H, Lu Z, Huang S-C, Yang Y, Chai H (Sept. 2025) “Learning-based over-the-air integrated sensing, communication and computation in UAV swarm-enabled intelligent transportation systems,” IEEE Transactions on Green Communications and Networking, pp. 1

Hung HL, Huy NH, Luong NC, Pham Q-V, Niyato D, Hoa NT (Jun. 2025) “Beamforming design for physical security in movable antenna-aided ISAC systems: a reinforcement learning approach,” IEEE Transactions on Vehicular Technology, pp. 1–5

Indulekha KP, Venkatesh TG (2024) “Dynamic resource and power allocation strategy for ISAC-aided 5G V2X networks using reinforcement learning,” in IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS), pp. 1–6

Jiang W, Zhou Q et al (2024) Terahertz communications and sensing for 6G and beyond: a comprehensive review. IEEE Communications Surveys & Tutorials 26(4):2326–2381

Jiang P, Li M et al (Jul. 2025) “SLP-based dual-functional waveform design for ISAC systems: a deep learning approach,” IEEE Transactions on Vehicular Technology, pp. 1–15

Jiang Y, Li X, Zhu G, Li H, Deng J, Han K, Shen C, Shi Q, Zhang R (Feb. 2025) “Integrated sensing and communication for low altitude economy: opportunities and challenges,” IEEE Communications Magazine, pp. 1–7

Jukuri CS, Parihar AS, Singh K et al (Jun 2025) “Machine learning optimization in the internet of metamaterial things with ISAC,” TechRxiv preprint, posted Jun. 27, 2025

Kaleem Z, Khalid W, Ahmad A, Yu H, Almasoud AM, Yuen C (2024) Reinforcement learning for energy-efficient user association in UAV-assisted cellular networks. IEEE Trans Aerosp Electron Syst 60(2):2474–2481

Kalkan Y (2024) 20 years of MIMO radar. IEEE Aerosp Electron Syst Mag 39(3):28–35

Kamal MM, Abideen SZU, Al-Khasawneh MA, Alabrah A, Larik RSA, Marwat MI (2025) Optimizing secure multi-user ISAC systems with STAR-RIS: a deep reinforcement learning approach for 6G networks. IEEE Access 13:31 472-31 484

Khalid W, Kaleem Z, Ullah R, Chien TV, Noh S, Yu H (2023) Simultaneous transmitting and reflecting-reconfigurable intelligent surface in 6G: design guidelines and future perspectives. IEEE Network 37(5):173–181

Khalid W, Rehman MAU, Chien TV, Kaleem Z, Lee H, Yu H (2024) Reconfigurable intelligent surface for physical layer security in 6G-IoT: designs, issues, and advances. IEEE Internet Things J 11(2):3599–3613

Khalid W, Boulogeorgos AAA, Chien TV, Lee J, Lee H, Yu H (2025) Optimal operation of active RIS-aided wireless powered communications in IoT networks. IEEE Internet Things J 12(1):390–401

Khalid W, et al (2025) “Malicious reconfigurable intelligent surfaces: security threats in 6G networks,” IEEE Internet of Things Journal, 12(15):30063–30085

Khan WU, Ahmed M, Sheemar CK, Di Renzo M, Lagunas E, Mahmood A, Shah ST, Dobre OA, Querol J, Chatzinotas S (2025) “Survey on beyond diagonal RIS enabled 6G wireless networks: fundamentals, recent advances, and challenges,” arXiv preprint arXiv:2503.08423

Khan WU, Mahmood A, Jamshed MA, Lagunas E, Ahmed M, Chatzinotas S (2024) “Beyond diagonal RIS for 6G non-terrestrial networks: potentials and challenges,” IEEE Network

Khan WU, Jameel F, Ristaniemi T, Khan S, Sidhu GAS, Liu J (2019) Joint spectral and energy efficiency optimization for downlink noma networks. IEEE Transactions on Cognitive Communications and Networking 6(2):645–656

Khan WU, Liu J, Jameel F, Sharma V, Jäntti R, Han Z (2020) Spectral efficiency optimization for next generation NOMA-enabled IoT networks. IEEE Trans Veh Technol 69(12):15 284-15 297

Khan WU, Li X, Zeng M, Dobre OA (2021) Backscatter-enabled NOMA for future 6G systems: a new optimization framework under imperfect SIC. IEEE Commun Lett 25(5):1669–1672

Khan WU, Li X, Ihsan A, Khan MA, Menon VG, Ahmed M (2021) NOMA-enabled optimization framework for next-generation small-cell IoV networks under imperfect SIC decoding. IEEE Trans Intell Transp Syst 23(11):22 442-22 451

Khan WU, Jameel F, Li X, Bilal M, Tsiftsis TA (2021) Joint spectrum and energy optimization of NOMA-enabled small-cell networks with QoS guarantee. IEEE Trans Veh Technol 70(8):8337–8342

Khan WU, Javed MA, Nguyen TN, Khan S, Elhalawany BM (2021) Energy-efficient resource allocation for 6G backscatter-enabled NOMA IoV networks. IEEE Trans Intell Transp Syst 23(7):9775–9785

Khan WU, Jameel F, Kumar N, Jäntti R, Guizani M (2021) Backscatter-enabled efficient V2X communication with non-orthogonal multiple access. IEEE Trans Veh Technol 70(2):1724–1735

Khan WU, Nguyen TN, Jameel F, Jamshed MA, Pervaiz H, Javed MA, Jäntti R (2021) Learning-based resource allocation for backscatter-aided vehicular networks. IEEE Trans Intell Transp Syst 23(10):19 676-19 690

Khan WU, Lagunas E, Ali Z, Javed MA, Ahmed M, Chatzinotas S, Ottersten B, Popovski P (2022) Opportunities for physical layer security in uav communication enhanced with intelligent reflective surfaces. IEEE Wirel Commun 29(6):22–28

Khan WU, Jamshed MA, Lagunas E, Chatzinotas S, Li X, Ottersten B (2022) Energy efficiency optimization for backscatter enhanced NOMA cooperative V2X communications under imperfect CSI. IEEE Trans Intell Transp Syst 24(11):12 961-12 972

Khan WU, Ihsan A, Nguyen TN, Ali Z, Javed MA (2022) NOMA-enabled backscatter communications for green transportation in automotive-industry 5.0. IEEE Trans Industr Inf 18(11):7862–7874

Khan WU, Javed MA, Zeadally S, Lagunas E, Chatzinotas S (2023) Intelligent and secure radio environments for 6G vehicular aided hetnets: key opportunities and challenges. IEEE Communications Standards Magazine 7(3):32–39

Khan WU, Lagunas E, Mahmood A, Chatzinotas S, Ottersten B (2023) RIS-assisted energy-efficient LEO satellite communications with NOMA. IEEE Transactions on Green Communications and Networking 8(2):780–790

Khan WU, Lagunas E, Mahmood A, Ali Z, Asif M, Chatzinotas S, Ottersten B (2023) Integration of noma with reflecting intelligent surfaces: a multi-cell optimization with SIC decoding errors. IEEE Transactions on Green Communications and Networking 7(3):1554–1565

Khan WU, Ali Z, Lagunas E, Mahmood A, Asif M, Ihsan A, Chatzinotas S, Ottersten B, Dobre OA (2023) Rate splitting multiple access for next generation cognitive radio enabled leo satellite networks. IEEE Trans Wireless Commun 22(11):8423–8435

Khan WU, Mahmood A, Sheemar CK, Lagunas E, Chatzinotas S, Ottersten B (2024) Reconfigurable intelligent surfaces for 6G non-terrestrial networks: assisting connectivity from the sky. IEEE Internet of Things Magazine 7(1):34–39

Khan WU, Mahmood A, Bozorgchenani A, Jamshed MA, Ranjha A, Lagunas E, Pervaiz H, Chatzinotas S, Ottersten B, Popovski P (2024) Opportunities for intelligent reflecting surfaces in 6G empowered V2X communications. IEEE Internet of Things Magazine 7(6):72–79

Khan WU, Lagunas E, Mahmood A, Asif M, Ahmed M, Chatzinotas S (2025) Integration of beyond diagonal RIS and UAVs in 6G NTNs: enhancing aerial connectivity. IEEE Wirel Commun 32(3):56–63

Kok JR, Spaan MT, Vlassis N (2005) Non-communicative multi-robot coordination in dynamic environments. Robot Auton Syst 50(2–3):99–114

Kong L, Pang X et al (2025) Signal enhancement and suppression schemes for bi-static ISAC with IRS-mounted target. IEEE Trans Commun 73(5):3718–3732

Kurma S, Singh K, Mumtaz S, Tsiftsis TA, Li C-P (2024) Resource optimization in active-STAR-RIS-aided THz ISAC systems with DDA modulation: A machine-learning approach. IEEE Trans Wireless Commun 23(10):15 291-15 307

Lai J, Wang D, Zhao Y (2025) IMSBA: a novel integrated sensing and communication beam allocation based on multi-agent reinforcement learning for mmWave Internet of vehicles. Appl Sci 15(11):6069

Lei L, Tan Y, Zheng K, Liu S, Zhang K, Shen X (2020) Deep reinforcement learning for autonomous internet of things: model, applications and challenges. IEEE Communications Surveys & Tutorials 22(3):1722–1760

Leyva L, Castanheira D, Silva A, Gameiro A (2024) Hybrid beamforming design for communication-centric ISAC. IEEE Sens J 24(13):21 179-21 190

Li Z, Liu F, Yang W, Peng S, Zhou J (2022) A survey of convolutional neural networks: analysis, applications, and prospects. IEEE Transactions on Neural Networks and Learning Systems 33(12):6999–7019

Li S, Zhang S, Wang Z, Zhou Z, Wang X, Mumtaz S, Guizani M, Frascolla V (2023) Asynchronous FDRL-based low-latency computation offloading for integrated terrestrial and non-terrestrial power IoT. IEEE Network 37(5):33–41

Li Y, Khan F, Ahmed M et al (2025) RIS-based physical layer security for integrated sensing and communication: a comprehensive survey. IEEE Internet Things J 12(16):32444–32468

Li C, Chai L, Zhang Y, Yang M, Zhao R, Zhang Z, Li D, Wan S (2025) Deep reinforcement learning-based resource allocation with enhanced perception and low-latency for autonomous driving in ISAC-aided VEC. ACM Transactions on Design Automation of Electronic Systems 30(3):1–34

Li D, Du B, Bai Z (Feb. 2025) “Deep reinforcement learning-enabled trajectory and bandwidth allocation optimization for UAV-assisted integrated sensing and covert communication,” Drones, 9(3):

Li O, He J, Zeng K, Yu Z, Du X, Zhou Z, Liang Y, Wang G, Chen Y, Zhu P, Tong W, Lister D, Ibbetson L (2022) “Integrated sensing and communication in 6G: a prototype of high resolution multichannel thz sensing on portable device,” EURASIP Journal on Wireless Communications and Networking, 2022(106). [Online]. Available: https://doi.org/10.1186/s13638-022-02172-w

Lin F, Xu C, Yuan W (2024) “DRL-based user selection and bandwidth allocation in ISAC-assisted IoV networks,” in IEEE 24th International Conference on Communication Technology (ICCT), pp. 488–493

Li X, Pegoraro J, de la Oliva A, Sciancalepore V, Brenes J et al. (2025) “Multix: advancing 6G-RAN through multi-technology, multi-sensor fusion, multi-band and multi-static perception,” IEEE Wireless Communications, pp. 1–8

Liu F, Cui Y et al (2022) Integrated sensing and communications: toward dual-functional wireless networks for 6G and beyond. IEEE J Selected Areas in Commun 40(6):1728–1767

Liu A, Huang Z et al (2022) A survey on fundamental limits of integrated sensing and communication. IEEE Communications Surveys & Tutorials 24:994–1034

Liu X, Wu J, Zhao C, Liu Z (2025) Integrated sensing and communications for UAV assisted internet of things based on deep reinforcement learning. IEEE Trans Veh Technol 74(6):9604–9616

Liu C, Xia M, Zhao J, Li H, Gong Y (2025) Optimal resource allocation for integrated sensing and communications in internet of vehicles: a deep reinforcement learning approach. IEEE Trans Veh Technol 74(2):3028–3038

Liu H, Zhong L, Tellambura C, Li Y, Cheng W (2025) Waveform design for integrated sensing and communications with PAPR constraint. IEEE Wireless Communications Letters 14(6):1653–1657

Liu H, Alsusa E, Al-Dweik A (2025) “UAV tracking using channel-anomaly-based deep learning in ISAC systems,” arXiv preprint arXiv: 1036227/techrxiv.175492191.18649834/v1

Liu A, Jiang W, Huang S, Feng Z (Jun. 2025) “Multi-modal integrated sensing and communication in internet of things with large language models,” IEEE Internet of Things Magazine, pp. 1–9

Liu Y, Li X, He B, Gu M, Huangfu W (Nov. 2024) “UAV-enabled diverse data collection via integrated sensing and communication functions based on deep reinforcement learning,” Drones, 8(11)

Liu Y, Mao W, Li X, Huangfu W, Ji Y, Xiao Y (2024) “UAV-assisted integrated sensing and communication for emergency rescue activities based on transfer deep reinforcement learning,” in Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, ser. ACM MobiCom ’24. New York, NY, USA: Association for Computing Machinery, p. 2142–2147

Liu Y, Zhang S, Li X, Huang Y, Fang Y, Cao H (2024) “Deep reinforcement learning-based beamforming design in ISAC-assisted vehicular networks,” in IEEE Wireless Communications and Networking Conference (WCNC), pp. 1–6

Liu X, Zhong W, Wang X, Duan H, Fan Z, Jin H, Huang Y, Lin Z (May 2024) “Deep reinforcement learning-based 3D trajectory planning for cellular connected UAV,” Drones, 8(5)

Li B, Wang X, Ahn S, Park S-I, Wu Y (2024) “Successive resource allocation in multi-user ISAC system through deep reinforcement learning,” in ICC 2024 - IEEE International Conference on Communications, pp. 5577–5583

Long X, Zhao Y, Wu H, Xu C-Z (2024) Deep reinforcement learning for integrated sensing and communication in RIS-assisted 6G V2X system. IEEE Internet Things J 11(24):39 834-39 849

Lu S, Liu F, Li Y, Zhang K, Huang H, Zou J, Li X, Dong Y, Dong F, Zhu J, Xiong Y, Yuan W, Cui Y, Hanzo L (2024) Integrated sensing and communications: recent advances and ten open challenges. IEEE Internet Things J 11(11):19094–19120

Luo B, Yang Y, Wu H-N, Huang T (2020) Balancing value iteration and policy iteration for discrete-time control. IEEE Transactions on Systems Man and Cybernetics Systems 50(11):3948–3958

Luo X, Lin Q, Zhang R, Chen H-H, Wang X, Huang M (Apr. 2025) “ISAC A survey on its layered architecture, technologies, standardizations, prototypes and testbeds,” IEEE Communications Surveys & Tutorials, pp. 1

Luong NC, Huynh-The T, Vu T-H, Le DV, Nguyen HT, Hai ND, Nguyen G-V, Anh NDD, Niyato D, Kim DI, Pham Q-V (Jul. 2025) “Advanced learning algorithms for integrated sensing and communication (ISAC) systems in 6G and beyond: a comprehensive survey,” IEEE Communications Surveys & Tutorials, pp. 1–1

Luong NC, Lu X, Hoang DT, Niyato D, Kim DI (2021) Radio resource management in joint radar and communication: a comprehensive survey. IEEE Communications Surveys & Tutorials 23(2):780–814

Ma Z et al (Jun. 2025) “Deep reinforcement learning for energy efficiency maximization in RSMA-IRS-assisted ISAC system,” IEEE Trans. Veh. Technol

Magbool A, Kumar V, Wu Q, Renzo MD, Flanagan MF (2025) A survey on integrated sensing and communication with intelligent metasurfaces: trends, challenges, and opportunities. IEEE Open Journal of the Communications Society 6:7270–7318

Magbool A, Kumar V, Wu Q, Renzo MD, Flanagan MF (Aug. 2025) “A survey on integrated sensing and communication with intelligent metasurfaces: trends, challenges, and opportunities,” IEEE Open Journal of the Communications Society, pp. 1–1

Mamaghani MT, Zhou X, Yang N, Swindlehurst AL (2025) “Securing integrated sensing and communication against a mobile adversary: a stackelberg game with deep reinforcement learning,”

Mason F, Pegoraro J (2025) Using deep reinforcement learning to enhance channel sampling patterns in integrated sensing and communication. IEEE Wireless Communications Letters 14(3):821–825

Menghani G (2023) Efficient deep learning: a survey on making deep learning models smaller, faster, and better. ACM Comput Surv 55(12):1–37

Mnih V, Badia AP, Mirza M, Graves A, Lillicrap T, Harley T, Silver D, Kavukcuoglu K (Feb. 2016) "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PmLR, pp. 1928–1937

Mnih V, Kavukcuoglu et al (Feb. 2015) "Human-level control through deep reinforcement learning," Nature, 518(7540):529–533

Moon S, Lee C-G, Liu H, Hwang I (2024) Deep reinforcement learning-based sum rate maximization for RIS-assisted ISAC-UAV network. ICT Express 10(5):1174–1178

Nikbakht H, Wigger M, Shamai S, Poor HV (Dec. 2024) "A memory-based reinforcement learning approach to integrated sensing and communication,"

O'Donoghue B, Osband I, Munos R, Mnih V (2018) "The uncertainty bellman equation and exploration," in International conference on machine learning, pp. 3836–3845

Okati N, Noll Barreto A, Uzeda Garcia L, Wigard J (2026) Co-existence analysis of terrestrial and non-terrestrial networks in s-band using stochastic geometry. IEEE Trans Commun 74:4431–4445

Ouyang W, Liu Q, Mu J (2024) AI-Dulaimi A, Jing X, Liu Q, "Communication-efficient federated learning for large-scale multiagent systems in ISAC: data augmentation with reinforcement learning,." IEEE Syst J 18(4):1893–1904

Paul A, Singh K, Kaushik A, Li C-P, Dobre OA, Di Renzo M, Duong TQ (2025) Quantum-enhanced DRL optimization for DoA estimation and task offloading in ISAC systems. IEEE J Sel Areas Commun 43(1):364–381

Pei J, Frascolla V, Al-Dulaimi A, Liu W, Aldhyani THH, Bashir AK, Mumtaz S (2026) Distributed large models training optimization with real-time wireless channel feedback. IEEE J Sel Areas Commun 44:2231–2243

Powers R, Shoham Y (Dec. 2004) "New criteria and a new algorithm for learning in multi-agent systems," Advances in neural information processing systems, vol. 17

Price B, Boutilier C (2003) Accelerating reinforcement learning through implicit imitation. Journal of Artificial Intelligence Research 19:569–629

Qi X, He X, Li D, Chen X (2024) Sensing-assisted end-to-end beamforming based on reinforcement learning in LEO satellite communications, in. IEEE/CIC International Conference on Communications in China (ICCC Workshops) 2024:699–704

Qin Y, Zhang Z, Li X, Huangfu W, Zhang H (2023) Deep reinforcement learning based resource allocation and trajectory planning in integrated sensing and communications UAV network. IEEE Trans Wireless Commun 22(11):8158–8169

Qing L, Xiang W, Ling X, Xu W, Li X, Liu J (May 2025) "Joint optimization of STARS-assisted air-ground ISAC system using deep reinforcement learning," IEEE Trans. Veh. Technol

Rahimi M, Sobhi-Givi S, Nouri M, Behroozi H, Uysal M (2025) "A deep reinforcement learning framework for joint BS and beyond diagonal RIS beamforming design in ISAC systems," in 2025 13th Iran Workshop on Communication and Information Theory (IWCIT)

Rashid T, Samvelyan M, De Witt CS, Farquhar G, Foerster J, Whiteson S (2020) Monotonic value function factorisation for deep multi-agent reinforcement learning. J Mach Learn Res 21(178):1–51

Ren R, Zhao J, Zhang Q, Wang D, Li J (2025) DRL beamforming in RIS-aided IoV for integrated-sensing-communication-computation. IEEE Internet Things J 12(14):28 201-28 213

Ren R, Zhao J, Zou D, Zhang Q, Wang D, Xu W (2026) Collaborative computation in integrated sensing, communication, and computation system for autonomous driving. IEEE Trans Intell Transp Syst 27(1):883–894

Ren R, Zhao J, Li Y, Zou D, Zhang Q, Gong Y (2024) "DRL-based performance optimization of IoT integrated sensing, communication, and computation systems," in 2024 3rd International Joint Conference on Information and Communication Engineering (JCICE), pp. 134–138

Saikia P, Singh K, Huang W-J, Duong TQ (2024) Hybrid deep reinforcement learning for enhancing localization and communication efficiency in RIS-aided cooperative ISAC systems. IEEE Internet Things J 11(18):29 494-29 510

Saikia P, Jee A, Singh K, Huang W-J, Boulogeorgos A-AA, Tsiftsis TA (Mar. 2025) "Hybrid-RIS empowered UAV-assisted ISAC systems: transfer learning-based DRL," IEEE Transactions on Communications, pp. 1

Shang C, Yu J, Hoang DT (Jul. 2025) "Energy-efficient and intelligent ISAC in V2X networks with spiking neural networks-driven DRL," arXiv preprint arXiv:2501.01038

Sheemar CK, Khan WU, Alexandropoulos G, Querol J, Chatzinotas S (Feb. 2025) "Holographic joint communications and sensing with cramer-rao bounds," arXiv preprint arXiv:2502.15248

Sheemar CK, Querol J et al (2025) "Joint communications, sensing, and positioning in 6G multi-functional satellite systems: survey and open challenges," arXiv preprint arXiv:2509.25937

Sheemar CK, Thiruvasagam P et al (Jan. 2025) “Joint communications and sensing for 6G satellite networks: use cases and challenges,” arXiv preprint arXiv:2501.05243

Shen X, Zheng H, Lin J, Feng X (2025) Joint deep reinforcement learning and unfolding for sensing and communication function selection in vehicular networks. IEEE Trans Veh Technol 74(3):4933–4945

Shen L, Li B, Zhu X (Apr. 2024) “Robust offloading for edge computing-assisted sensing and communication systems: a deep reinforcement learning approach,” Sensors, 24(8):

Sheraz M, Ahmed M, Hou X, Li Y, Jin D, Han Z, Jiang T (2020) Artificial intelligence for wireless caching: schemes, performance, and challenges. IEEE Communications Surveys & Tutorials 23(1):631–661

Smeenk C, Zhao Z, Schneider C, Robert J, Galdo GD (2024) “Optimizing radio resources for radar services in ISAC systems by deep reinforcement learning,” in IEEE 35th International Symposium on Personal. Indoor and Mobile Radio Communications (PIMRC) 2024:1–6

Song X, Fang Y, Wang F, Ren Z, Yu X, Zhang Y, Liu F, Xu J, Ng DWK, Zhang R et al (2025) An overview on IRS-enabled sensing and communications for 6G: architectures, fundamental limits, and joint beamforming designs. SCIENCE CHINA Inf Sci 68(5):150301

Strinati EC, Alexandropoulos GC, Amani N, Crozzoli M, Madhusudan G, Mekki S, Rivet F, Sciancalepore V, Sehier P, Stark M, Wymeersch H (2025) Toward distributed and intelligent integrated sensing and communications for 6G networks. IEEE Wirel Commun 32(1):60–67

Sumiea EH, Abdulkadir SJ, Alhussian HS, Al-Selwi SM, Alqushaibi A, Ragab MG, Fati SM (May 2024) “Deep deterministic policy gradient algorithm: a systematic review,” Heliyon, 10(9)

Sun J, Wu C, Mumtaz S, Tao J, Cao M, Wang M, Frascolla V (2024) An efficient privacy-aware split learning framework for satellite communications. IEEE J Sel Areas Commun 42(12):3355–3365

Sunehag P, Lever G et al (2017) “Value-decomposition networks for cooperative multi-agent learning,” arXiv preprint arXiv:1706.05296

Tan M (1993) “Multi-agent reinforcement learning: Independent vs. cooperative agents,” in Proceedings of the tenth international conference on machine learning, 330–337

Tang R, Chai R, Li P (2023) “Deep reinforcement learning-based sensing and communication scheduling algorithm for UAV-assisted target detection systems,” in 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), pp. 1–5

Tishchenko A, Khalily M, Shojaeifard A, Burton F, Björnson E, Renzo MD, Tafazolli R (Jan. 2025) “The emergence of multi-functional and hybrid reconfigurable intelligent surfaces for integrated sensing and communications -A survey,” IEEE Communications Surveys & Tutorials, pp. 1–1

van Erven T, Harremos P (2014) Renyi divergence and kullback-leibler divergence. IEEE Trans Inf Theory 60(7):3797–3820

Van Hasselt H, Guez A, Silver D (Feb. 2016) “Deep reinforcement learning with double Q-learning,” in Proceedings of the AAAI conference on artificial intelligence, 30(1)

Wang Z, Luo Z, Lei Y (2025) DRL-based RIS-assisted integrated sensing and communication: joint design of beamforming and reflection phase shift. Physical Communication 71:102669

Wang Z, Bapst V, Heess N, Mnih V, Munos R, Kavukcuoglu K, De Freitas N (Nov. 2016) “Sample efficient actor-critic with experience replay,” arXiv preprint arXiv:1611.01224

Wang M, Chen P, Cao Z, Chen Y (Jan. 2022) “Reinforcement learning-based UAVs resource allocation for integrated sensing and communication (ISAC) system,” Electronics, 11(3)

Wang C, Li G, Zhang H, Wong K-K, Li Z, Ng DWK, Chae C-B (Sept. 2024) “Fluid antenna system liberating multiuser MIMO for ISAC via deep reinforcement learning,” IEEE Transactions on Wireless Communications, 23(9):10 879–10 894

Wei Z, Qu H, Wang Y, Yuan X, Wu H, Du Y, Han K, Zhang N, Feng Z (2023) Integrated sensing and communication signals toward 5G-A and 6G: a survey. IEEE Internet Things J 10(13):11 068-11 092

Wei Z, Jia J, Niu Y, Wang L, Wu H, Yang H, Feng Z (2025) Integrated sensing and communication channel modeling: a survey. IEEE Internet Things J 12(12):18850–18864

Wei W, Yu X, Cui G (Jun. 2025) “Transmit beampattern synthesis for phased-MIMO radar via a unified design framework,” IEEE Transactions on Aerospace and Electronic Systems, pp. 1–12

Wen D, Liu P, Zhu G, Shi Y, Xu J, Eldar YC, Cui S (2024) Task-oriented sensing, computation, and communication integration for multi-device edge ai. IEEE Trans Wireless Commun 23(3):2486–2502

Wen D, Zhou Y, Li X, Shi Y, Huang K, Letaief KB (Dec. 2024) “A survey on integrated sensing, communication, and computation,” IEEE Communications Surveys & Tutorials, pp. 1–1

Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach Learn 8(3):229–256

Wu Z, Yu C, Ye D, Zhang J, Zhuo HH et al (2021) Coordinated proximal policy optimization. Adv Neural Inf Process Syst 34:26 437-26 448

Xie Z, Wang Z, Zhang Z, Wang J, Jiang Z, Han Z (2024) Distributed UAV swarm for device-free integrated sensing and communication relying on multi-agent reinforcement learning. IEEE Trans Veh Technol 73(12):19 925-19 930

Xie W, Sun G, Wang J, Du H, Kang J, Niyato D, Huang K, Leung VCM (Feb. 2025) "Multi-objective low-altitude IRS-assisted ISAC optimization via generative AI-enhanced deep reinforcement learning," arXiv preprint arXiv:2502.10687
Xie C, Xiu Y, Yang S, Miao Q, Chen L, Gao Y, Zhang Z (Jan. 2025) "Deep reinforcemnet learning for robust beamforming in integrated sensing, communication and power transmission systems," Sensors, 25(2):
Xu H, Sun J, Yuan Y, Yi W (2024) Beam pattern-aware resource allocation for phased array radar network under suppression jamming. IEEE Sens J 24(15):24 825-24 840
Xu X, Zhu F, Han S, Yu Z, Zhao H, Wang B, Zhang P (2024) Swin-Loc: transformer-based CSI fingerprinting indoor localization with MIMO ISAC system. IEEE Trans Veh Technol 73(8):11 664-11 679
Yang L, Wei Y, Feng Z, Zhang Q, Han Z (2024) Deep reinforcement learning-based resource allocation for integrated sensing, communication, and computation in vehicular network. IEEE Trans Wireless Commun 23(12):18 608-18 622
Yang Y, Gao Y, Ding Z, Wu J, Zhang S, Han F, Qiu X, Gao S, Wang Y-G (2024) Advancements in Q-learning meta-heuristic optimization algorithms: a survey. Wiley Interdisciplinary Reviews Data Mining and Knowledge Discovery 14(6):e1548
Yang S, Luo W (2024) "DRL-based secure beamforming design for RIS-assisted ISAC system: a CRB perspective," in. International Conference on Future Communications and Networks (FCN) 2024:1–6
Yang K, Xu J et al (Jan. 2025) "Sensing-centric energy-efficient hybrid beamforming design for terahertz integrated sensing and communications," IEEE Transactions on Vehicular Technology, pp. 1–6
Yang S, Yao J, Tang J, Wu T, Elkashlan M, Yuen C, Debbah M, Shin H, Valenti M (2025) "Towards intelligent antenna positioning: leveraging DRL for FAS-aided ISAC systems,"
Yao Y, Zhu Z, Miao P, Cheng X, Shu F, Wang J (2025) Optimizing hybrid RIS-aided ISAC systems in V2X networks: a deep reinforcement learning method for anti-eavesdropping techniques. IEEE Trans Veh Technol 74(6):9224–9239
Yao J, Xu W, Zhu G, Huang K, Cui S (2025) Energy-efficient edge inference in integrated sensing, communication, and computation networks. IEEE J Sel Areas Commun 43(10):3580–3595
Ye X, Mao Y, Yu X, Fu L (2025) Intelligent omni-surface-aided integrated sensing and communications based on deep reinforcement learning with knowledge transfer. IEEE Trans Wireless Commun 24(5):4344–4360
Ye X, Mao Y, Yu X, Sun S, Fu L, Xu J (Jan. 2025) "Integrated sensing and communications for low-altitude economy: a deep reinforcement learning approach,"
Yonel B, Choudhury N, Yazici B (2024) Phaseless multistatic synthetic aperture radar imaging. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 17:6134–6147
Yu Y, Si X, Hu C, Zhang J (2019) A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 31(7):1235–1270
Zhai W, Wang X, Greco MS, Gini F (2023) "Reinforcement learning based integrated sensing and communication for automotive MIMO radar," in 2023 IEEE Radar Conference (RadarConf23), pp. 1–6
Zhang J et al (2024) Joint design for STAR-RIS aided ISAC: decoupling or learning. IEEE Trans Wireless Commun 23(10):14 365-14 379
Zhang Q, Ji K et al (2024) Joint communication and sensing system performance evaluation and testbed: a communication-centric approach. IEEE Network 38(5):286–294
Zhang SMIJA, Rahman ML, Wu K, Huang SMIX, Guo FIYJ, Fellow ISC, Yuan FIJ, Chen S (2021) Enabling joint communication and radar sensing in mobile networks–a survey. IEEE Communications Surveys & Tutorials 24:306–345
Zhang JA, Rahman ML, Wu K, Huang X, Guo YJ, Chen S, Yuan J (2022) Enabling joint communication and radar sensing in mobile networks–A survey. IEEE Communications Surveys & Tutorials 24(1):306–345
Zhang Y, Masouros C, Xu T (May 2025) "Zero-power integrated sensing and communication in smart healthcare environments," *IEEE Transactions on Cognitive Communications and Networking*, pp. 1–1
Zhang J, Xu S, Li C, Huang Y, Yang L (Jun. 2025) "Efficient beam selection for ISAC in cell-free massive MIMO via digital twin-assisted deep reinforcement learning,"
Zhang B, Zheng K, Tong C, Yang K, Yan K (2024) "Deep reinforcement learning for active RIS-assisted full-duplex integrated sensing and communication systems," in 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), pp. 2680–2685
Zhao Y, Liu H, Liu X, Wang X, Gao Z, Huang L (2024) IBTD: a novel ISAC beam tracking based on deep reinforcement learning for mmWave V2V networks. IET Commun 18(19):1403–1416
Zhao J, Ren R, Zou D, Zhang Q, Xu W (2024) IoV-Oriented integrated sensing, computation, and communication: system design and resource allocation. IEEE Trans Veh Technol 73(11):16 283-16 294
Zhou W, Zhang R, Chen G, Wu W (2022) Integrated sensing and communication waveform design: a survey. IEEE Open Journal of the Communications Society 3:1930–1949
Zhu X, Liu J, Lu L, Zhang T, Qiu T, Wang C, Liu Y (2025) Enabling intelligent connectivity: a survey of secure ISAC in 6G networks. IEEE Commun Surv Tutor 27(2):748–781

Zhu Z, Gong M, Chu Z, Xiao P, Sun G, Mi D, He Z, Tong F (2023) “DRL-based STAR-RIS-assisted ISAC secure communications,” in. International Conference on Ubiquitous Communication (Ucom) 2023:127–132

Zhu Q, Liu R, Liu Q, Chen C (Jan. 2024) “Resource allocation for UAV swarm-assisted green ISAC networks via multi-agent RL,” IEEE Transactions on Green Communications and Networking, pp. 1

Authors and Affiliations

Wali Ullah Khan[1] · Waqas Khalid[2] · Muhammad Iqbal[3] · Chiew Foong Kwong[2] · Manzoor Ahmed[4] · Syed Tariq Shah[5]

✉ Waqas Khalid
Waqas.Khalid@nottingham.edu.cn

✉ Syed Tariq Shah
syed.shah@essex.ac.uk

Wali Ullah Khan
waliullahkhan30@gmail.com

Muhammad Iqbal
iqbal.marjan@gmail.com

Chiew Foong Kwong
Kwong@nottingham.edu.cn

Manzoor Ahmed
manzoor.achakzai@gmail.com

1 School of Computing, Gachon University, 1342 Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do 13120, South Korea

2 Department of Electrical and Electronic Engineering, and the Next Generation Internet of Everything Laboratory (NGIoE Lab), University of Nottingham Ningbo China, Ningbo 315100, China

3 Department of Communications Engineering, National Chung Cheng University, Chiayi, Taiwan

4 School of Computer and Information Science and also with the Institute for AI Industrial Technology Research, Hubei Engineering University, Xiaogan City 432000, China

5 School of Computer Science and Electronic Engineering, University of Essex, Colchester, UK