future internet

MDPI

Article

Reconfigurable-Intelligent-Surface-Enhanced Dynamic Resource Allocation for the Social Internet of Electric Vehicle Charging Networks with Causal-Structure-Based Reinforcement Learning

Yuzhu Zhang *,† and Hao Xu *,†

Department of Electrical & Biomedical Engineering, University of Nevada, Reno, NV 89557, USA
* Correspondence: yuzhuz@nevada.unr.edu (Y.Z.); haoxu@unr.edu (H.X.)
† These authors contributed equally to this work.

Abstract: Charging stations and electric vehicle (EV) charging networks signify a significant advancement in technology as a frontier application of the Social Internet of Things (SIoT), presenting both challenges and opportunities for current 6G wireless networks. One primary challenge in this integration is limited wireless network resources, particularly when serving a large number of users within distributed EV charging networks in the SIoT. Factors such as congestion during EV travel, varying EV user preferences, and uncertainties in decision-making regarding charging station resources significantly impact system operation and network resource allocation. To address these challenges, this paper develops a novel framework harnessing the potential of emerging technologies, specifically reconfigurable intelligent surfaces (RISs) and causal-structure-enhanced asynchronous advantage actor–critic (A3C) reinforcement learning techniques. This framework aims to optimize resource allocation, thereby enhancing communication support within EV charging networks. Through the integration of RIS technology, which enables control over electromagnetic waves, and the application of causal reinforcement learning algorithms, the framework dynamically adjusts resource allocation strategies to accommodate evolving conditions in EV charging networks. An essential aspect of this framework is its ability to simultaneously meet real-world social requirements, such as ensuring efficient utilization of network resources. Numerical simulation results validate the effectiveness and adaptability of this approach in improving wireless network efficiency and enhancing user experience within the SIoT context. Through these simulations, it becomes evident that the developed framework offers promising solutions to the challenges posed by integrating the SIoT with EV charging networks.

Keywords: reconfigurable intelligent surfaces; EV charging networks; SIoT; causal structure; A3C; reinforcement learning; RIS phase shift; energy efficiency

Citation: Zhang, Y.; Xu, H. Reconfigurable-Intelligent-Surface-Enhanced Dynamic Resource Allocation for the Social Internet of Electric Vehicle Charging Networks with Causal-Structure-Based Reinforcement Learning. *Future Internet* **2024**, *16*, 165. https://doi.org/10.3390/fi16050165

Academic Editors: Domenico Ursino, Francesco Cauteruccio and Luca Virgili

Received: 9 April 2024
Revised: 30 April 2024
Accepted: 3 May 2024
Published: 11 May 2024

1. Introduction

The implementation of a charging infrastructure and networks for electric vehicles (EVs) encounters numerous challenges [1], particularly when serving distributed EV charging networks with limited wireless network resources [2,3]. Factors such as congestion during EV travel, diverse preferences among EV users, and uncertainties in decision-making regarding charging station (CS) resources profoundly impact system operation and network resource allocation. Meanwhile, with the development of smart grids, electric vehicle charging networks have also experienced rapid growth [4]. Also, current and upcoming wireless networks, including 5G/6G and their subsequent versions, are expected to provide significantly improved data rates, reduced latency, and expanded network coverage compared to previous versions [5,6]. This progress in wireless networks stems from new design principles, enabling them to support a large number of connected devices simultaneously. Making sure things connect well and can share information easily is really important, especially for the growing number of Social Internet of Things (SIoT) apps [7,8]. They need smooth interactions to continue getting better.

In smart electric vehicle charging networks, the communication system adjusts multiple features to achieve desirable communication outcomes. For example, factors such as communication power allocation at charging stations, the presence of active or passive relays [9], the communication channel quality, and the presence of obstacles may affect the overall communication conditions [10]. As actions are taken to optimize the communication network, the system's communication state changes accordingly. In [11], this is represented using Markov Decision Processes (MDPs), with the communication quality serving as the reward factor. In this scenario, the numbers of interventions and states both exponentially increase.

To address these challenges, this paper introduces a novel framework leveraging emerging technologies, specifically reconfigurable intelligent surfaces (RISs) and causal-structure-based reinforcement learning techniques.

1.1. Background

Reconfigurable intelligent surfaces (RISs) are a revolutionary technology in the field of wireless communication and signal propagation [12]. The structure of RISs typically includes a dielectric surface panel, which is a subgroup of periodic structures [13] composed of repeating minimal geometric shapes called unit cells. Each unit cell contains conductive printed patches, also known as scatterers, with sizes that are a small fraction of the operating frequency wavelength. The macroscopic effect of these scatterers defines a specific impedance surface [14], which, when controlled, can manipulate reflected waves from the dielectric surface panel. Each scatterer or cluster of scatterers can be adjusted to reconstruct electromagnetic waves with desired characteristics across the entire surface. By intelligently controlling the phase, magnitude, and polarization of reflected or transmitted waves, RISs can enhance wireless communication performance, improve signal coverage, and reduce interference in various wireless systems.

Electric vehicles (EVs) are considered an emerging strategy to reduce the dependence on oil and provide opportunities to reduce carbon emissions [15]. The main elements of an EV system include EVs, charging stations equipped with charging points, and associated communication systems. Managing EV charging and optimizing their interaction with the power grid relies on appropriate communication infrastructure between EVs, charging stations, and the power grid. EV charging communication networks play a crucial role in promoting the widespread adoption of electric vehicles by providing reliable communication connections for EV owners, thereby contributing to the transition to sustainable transportation.

In [16], social networks are utilized for searching internet resources, routing traffic, or selecting effective content distribution strategies. The Internet of Things (IoT) [17] integrates a vast array of technologies, envisioning various things or objects interacting and cooperating with each other through a series of communication protocols to achieve common goals. The convergence of the Internet of Things and social networks into the Social Internet of Things (SIoT) [18] is anticipated to have many desirable impacts on the future world. The SIoT aims to enhance the functionality, usability, and effectiveness of IoT systems by leveraging social relationships. By enabling IoT devices to collaborate, share information, and interact based on social context, the SIoT seeks to create more intelligent and adaptable IoT environments capable of addressing diverse user needs and preferences.

Meanwhile, causal reinforcement learning (CRL) [19,20] is a branch of reinforcement learning (RL) that incorporates causal reasoning into the decision-making process. Causal structures in machine learning refer to the graphical representation of causal relationships among variables in a given system. These structures capture the cause–effect relationships between different variables, enabling the identification of causal factors and the prediction of system behavior. Understanding causal structures is essential for making informed decisions, conducting causal inference, and designing effective machine learning models that can accurately capture and leverage causal relationships. In the complex communication environment considered in this paper, actions may not directly lead to observed

outcomes. Instead, they may influence outcomes through intermediate variables, allowing CRL algorithms to make wiser decisions.

Furthermore, asynchronous advantage actor–critic (A3C) [21] is a type of reinforcement learning algorithm that combines the advantages of both policy-based and value-based methods. A3C uses asynchronous training to update multiple agents concurrently, allowing for more efficient exploration of the action space and faster convergence to optimal policies. By incorporating an actor–critic architecture, A3C can learn both action policies and value functions simultaneously, leading to more stable and effective learning in complex environments.

The framework proposed in this paper aims to optimize resource allocation, thereby enhancing SIoT support within EV charging networks. By integrating RIS technology for electromagnetic wave control and applying causal RL algorithms, the framework dynamically adjusts resource allocation strategies to adapt to changing conditions in EV charging networks.

1.2. Limitations of RISs and CRL

However, there are limitations of RIS technology as well as causal reinforcement learning.

Firstly, regarding RIS technology, its practical application may encounter some limitations. For example, the deployment of RISs may require a substantial amount of hardware equipment and a complex installation process, which could increase system costs and deployment difficulties. Additionally, the performance of RISs may be influenced by environmental conditions, such as building structures or weather conditions, which could affect the effectiveness of the RIS and consequently degrade the communication quality and reliability.

As for causal reinforcement learning, its limitations primarily manifest in the model complexity and training time. Causal reinforcement learning may necessitate a large amount of data for training, and in complex environments, it may require significant time to converge to optimal solutions. Furthermore, the design and optimization of causal reinforcement learning algorithms may require specialized knowledge and expertise, potentially limiting their application in practical systems.

Moreover, challenges may arise in the interaction and integration of both technologies in practical applications. For instance, effectively integrating RIS technology and causal reinforcement learning algorithms into smart electric vehicle charging communication networks to achieve synergistic effects would require further research and optimization.

In summary, despite the potential advantages of RIS technology and causal reinforcement learning in smart electric vehicle charging communication networks, they also face practical limitations that need to be considered and addressed in real-world applications.

1.3. Related Studies

In a significant study [22], deep reinforcement learning (DRL) was utilized to dynamically configure phase shifts in reconfigurable intelligent surfaces (RISs), leading to enhancements in signal coverage, a reduced interference, and an improved spectral efficiency. Moreover, in one study [23], an exploration was conducted to leverage deep Q-networks (DQNs) for enhancing RIS-supported massive multi-input multi-output (MIMO) setups. This proposition centered on an adaptable control strategy, dynamically adjusting the phase shifts and beamforming weights associated with the RIS. This adaptation resulted in notable enhancements in the system's capacity, coverage, and energy efficiency. In [24], the author tackles resource allocation hurdles in vehicular communications. This is achieved by employing a multi-agent deep deterministic policy gradient (DDPG) method, where vehicle-to-vehicle (V2V) communications serve as agents utilizing non-orthogonal multiple access (NOMA) [25] technology for spectrum sharing. By approaching the problem as a decentralized discrete-time and finite-state Markov Decision Process (DFMDP) and implementing the DDPG method, the suggested approach optimizes the sum-rate of V2I

communications. It also guarantees the latency and reliability requirements are met for safety-critical V2V transmissions amidst a dynamic vehicular setting.

In recent years, traditional social networking has evolved into more intricate social internetworking, extending beyond human users to objects. Ref. [26] has explored the Social Internet of Things (SIoT) and Multiple IoT (MIoT) paradigms, with the SIoT focusing on technological challenges of interacting IoT devices, while the MIoT delves into data-driven and semantics-based aspects of smart object interactions. This paper investigates this concept of the scope in multi-IoT scenarios, proposing formalizations and applications, followed by experiments evaluating its effectiveness compared to existing parameters like the diffusion degree and influence degree. In [27], the author proposes a symbiotic radio (SR) system that supports both Internet of Things (IoT) and cellular networks, allowing multiple users to receive information from the base station while multi-IoT devices backscatter their data via the same signal. Leveraging robust design methods, the system minimizes the transmit power under cellular outage probability and multi-IoT transmission rate constraints, addressing channel uncertainty and demonstrating effectiveness through simulation results. Ref. [28] proposes an energy- and trust-aware opportunistic routing approach for the cognitive radio Social Internet of Things (CR-SIoT), leveraging network coding and game-theoretic allocation of trusted channels to enhance the network performance, as validated by extensive simulation results. To improve the social edge service (SES) in the Social Internet of Things (SIoT), Ref. [29] proposed a hybrid graph deep learning (HAD) approach that employs an adaptive trust weight (ATW) model and a quotient user-centric coeval learning (QUCL) mechanism, achieving an improved communication and computation performance and enhancing SES reliability.

Ref. [30] presents structural causal modeling (SCM) as a method for ecologists to discern cause-and-effect relationships from observational data, overcoming biases common in traditional statistical analyses like confounding. Utilizing directed acyclic graphs (DAGs) and graphical rules like the backdoor and frontdoor criteria, SCM systematically estimates causal effects between variables of interest in ecological studies, showing promise for advancing causal inference without the need for randomized experiments.

However, there are few works that apply the causal structure in the field of communication. The application of causal reinforcement learning in RIS-assisted SIoT communication systems is a promising research direction. RISs can be used to adjust the transmission characteristics of signals to adapt to different communication environments and requirements. Causal reinforcement learning can utilize historical data and environmental feedback to provide intelligent decision support for an RIS, enabling it to adjust its operating mode and parameter settings according to real-time demands and network conditions. By learning based on causal relationships, the system can better understand the impact of RISs and make decisions based on these causal relationships, thereby improving the performance of the communication system.

1.4. Our Contribution

Compared to a previous work [31], this paper differs mainly in the following aspects:
Focus and Background:

This paper focuses on dynamic resource allocation in RIS-assisted electric vehicle charging communication networks under cellular networks, with base stations as the core, especially addressing the wireless communication environment within electric vehicle charging networks. In contrast, the second summary emphasizes dynamic resource allocation in RIS-assisted mobile ad hoc networks (MANETs), particularly in addressing time-varying and uncertain wireless communication environments within multi-mobile ad hoc wireless networks.

Optimization Methods:

This paper proposes an asynchronous advantage actor–critic (A3C) algorithm based on causal factors to optimize communication network resource allocation control. It learns feature representations from incomplete communication environment states to accelerate the training speed, understand the causal relationships in the environment, and transfer training results to similar communication environments. On the other hand, the second summary introduces an inner–outer joint online optimization algorithm for RIS-assisted MANETs. It utilizes the D-UCB algorithm for RISs and spectrum selection in the outer network and employs the TD3 algorithm to gain decentralized insights into RIS phase shifts and power allocation strategies in the inner network.

Algorithmic Structure:

The CF-A3C algorithm in this paper first acquires causal factors, uses them as state acquaintances of the A3C network, and updates the global network using experiences collected by multiple worker threads, eliminating the need for a replay buffer and promoting efficient exploration in resource allocation tasks. Conversely, the TD3 algorithm in the second summary adopts an actor–critic structure with three target networks and two hidden layer streams in each neural network to segregate state-value and action-value distribution functions, accelerating convergence speed and enhancing the learning efficiency.

This paper presents a novel framework aimed at addressing the challenges associated with integrating the Social Internet of Things (SIoT) with connected electric vehicle (EV) charging networks. This framework harnesses emerging technologies, including reconfigurable intelligent surfaces (RISs), causal structures, and A3C-based reinforcement learning techniques, to optimize resource allocation and enhance SIoT support within EV charging networks.

By integrating RIS technology, which enables control over electromagnetic waves, and applying causal RL algorithms, the framework dynamically adjusts resource allocation strategies to accommodate the evolving conditions in distributed EV charging networks. Importantly, this framework aims to simultaneously meet real-world social requirements, such as fulfilling EV user charging needs, while ensuring efficient utilization of network resources, thereby enhancing communication performance.

The primary achievements elucidated within this manuscript are as follows:

- Establishment of a model to represent the fluctuating and uncertain wireless communication setting for managing dynamic resource allocation in RIS-assisted electric vehicle charging communication networks. The model depicts the dynamic resource allocation system operating within an electric vehicle charging network.
- Design of a causal inference model capable of reasoning about and addressing causal relationships in the electric vehicle charging communication network by acquiring effective representation distributions.
- Proposal of a causal-factor-based asynchronous advantage actor–critic (A3C) algorithm based on the designed causal factor model for optimizing communication network resource allocation control. The feature representations are derived from learning the incomplete communication environment states. This method introduces a novel approach to training actor–critic networks, known as A3C, by directly updating global networks using experiences collected by multiple worker threads. By eliminating the need for a replay buffer, the method streamlines training and promotes efficient exploration in resource allocation tasks. This advancement accelerates learning while enhancing overall performance within the A3C framework.

The advantages of our work over existing contributions are presented in Table 1.

The results of experiments conducted in different environments demonstrate that the CF-A3C algorithm is highly competitive with state-of-the-art resource optimization algorithms across multiple evaluation metrics.

Table 1. Comparison with existing contributions.

Aspect	Existing Contribution	Advancements in Provided Algorithms
Encoder	Simple encoder architectures	Utilization of a deep encoder with the self-attention mechanism
Captures non-linearity	Linear or shallow neural networks	Deep neural networks capturing complex relationships
Self-attention mechanism	No or only basic attention mechanisms	Integration of self-attention for capturing dependencies
Causal factor extraction	Manual feature engineering approaches	Automatic extraction of causal factors from state vectors
Resource allocation	Fixed or rule-based allocation methods	Dynamic allocation optimization using the CF-A3C algorithm
Scalability	Limited scalability for large systems	Scalable solutions suitable for complex network scenarios
Real-world applicability	Limited applicability in practical settings	Practical solutions with a demonstrated real-world impact

2. System and Channel Model

2.1. Scenario Overview

In an electric vehicle charging communication network, there are two parts: uplink and downlink communication between the base station and vehicles and uplink and downlink communication between the base station and charging stations.

User demands are input information in the network, and this algorithm aims to optimize the channel capacity between base stations and vehicles while maximizing energy efficiency. By dynamically adjusting communication resource allocation strategies to meet the changing needs of different users, the algorithm achieves more flexible and efficient resource utilization by monitoring and analyzing changes in user demands and considering the real-time status of network resources. Furthermore, the application of RIS technology makes the utilization of network resources more flexible and efficient by adjusting the direction and intensity of electromagnetic wave transmission, enhancing signal coverage and the transmission accuracy and thus improving the network resource utilization efficiency while meeting users' charging service needs. Additionally, the algorithm optimizes resource allocation strategies to avoid resource waste and overuse by integrating techniques such as causal reinforcement learning, maximizing network resource utilization while meeting user demands and thereby enhancing the overall network efficiency and performance. In summary, the A3C resource allocation control algorithm, assisted by an RIS, effectively balances user demands (electric vehicle charging) with network resource utilization, thereby improving the performance and efficiency of smart electric vehicle charging communication networks. This paper focuses on the downlink communication process between the base station and vehicles. In this network, using RIS-assisted communication can optimize the channel, enhance communication reliability, and improve performance. The communication process in electric vehicle charging involves uplink and downlink transmissions between vehicles and base stations, charging stations, and base stations with reflective intelligent surface (RIS) assistance. We introduce the communication between vehicles and the base station in the following.

Uplink Communication Process

Vehicles Send Data to yjr Base Station: Vehicles transmit uplink data to the base station through antennas based on their needs. These data may include the current status of the vehicle, charging demands, vehicle location, etc.

Downlink Communication Process

The Base Station Sends Data to Vehicles: The base station transmits downlink data to vehicles through antennas. These data may include the status information of charging stations, charging plans, traffic information, etc. If RIS-assisted communication is available, the base station can improve the channel quality and enhance the strength and reliability of signals reaching the vehicles through RISs.

RIS Processing of Downlink Signals: An RIS receives downlink signals sent by the base station and reflects the signals towards the direction of the vehicles based on pre-designed reflection coefficients and phase adjustments to enhance signal reception.

Vehicles Receive Signals: Vehicles receive downlink signals from both the base station and the RIS and utilize the received information to perform corresponding operations, such as adjusting charging behavior, updating charging plans, etc.

Information Interaction:

During the communication process in both uplink and downlink transmissions, there is interaction and processing of information among the base station, vehicles, and RISs. The base station is responsible for scheduling charging stations, processing information uploaded by vehicles, issuing charging plans, etc. Vehicles are responsible for uploading their own status, location, and other information, as well as receiving charging plans issued by the base station. RISs serve as an intermediary node, responsible for optimizing the channel, enhancing communication reliability and performance, and reflecting signals from the base station to vehicles or from vehicles to the base station. Through the above communication process and information interaction, the electric vehicle charging communication network needs to achieve efficient communication between charging stations and vehicles, providing support and optimization for the charging behavior of electric vehicles.

2.2. System Model

Exploring the RIS-enhanced electric vehicle charging network downlink procedure depicted in Figure 1, we observe base stations (BSs) acting as transmitters, equipped with N antennas, and one RIS composed of M element units for support, alongside L single-antenna electric vehicle users (VUs). The communication landscape is challenging, characterized by traffic congestion, buildings, and various obstacles, leading to blocked direct signal links from BSs to electric vehicle users. Consequently, a two-hop communication system is established, necessitating a BS to relay signals through an RIS to reach the users. For user k, the received signal at time t is presented as:

$$y_k(t) = (\mathbf{h}_{BV} + \mathbf{h}_{RV,k}(t)^H \Phi_k(t) \mathbf{H}_{BR,k}(t))\mathbf{x}(t) + n_k(t), \quad (1)$$

In this scenario, the transmitted signal on the k-th subcarrier is represented as $\mathbf{x}(t) \in \mathbb{C}^{M\times 1}$, the received signal is denoted by $y_k(t)$, and the additive white noise is represented as $n_k(t)$, following a normal distribution $\mathcal{CN}(0, \sigma_k^2)$. At time t, the line-of-sight channel gain is presented as $\mathbf{h}_B V$, and the non-line-of-sight channel gain is presented as the channel gain matrices from the base station to the RIS relay and from the RIS relay to the vehicle as $\mathbf{H}_{BR,k}(t) \in \mathbb{C}^{N\times M}$ and $\mathbf{h}_{RV,k}(t) \in \mathbb{C}^{1\times N}$, respectively. As Figure 1 shows, the direct link is blocked by buildings or other obstacles, so the communication between the base station and vehicle users is through the non-line-of-sight channel. Additionally, for user k at time t, the RIS comprises $M \times M$ reflecting elements, represented by the diagonal matrix $\mathbf{\Phi}_k(t)$, indicating their corresponding phases. Specifically, it is defined as $\mathbf{\Phi}_k(t) = \text{diag}[e^{j\theta_{1,k}(t)}, e^{j\theta_{2,k}(t)}, \ldots, e^{j\theta_{M,k}(t)}] \in \mathbb{C}^{M\times M}$. Considering the transmit power $p_k(t)$ from the base station to user k, the transmitted data $s_k(t)$, and the beamforming vector $\mathbf{q}_k(t)$

at the base station antennas, the term $x(t)$ is expressed as $\sum_{k=1}^{K} \sqrt{p_k(t)}\mathbf{q}_k(t)s_k(t)$, which is the transmitted signal $\mathbf{x}(t)$ at time t. The following constraints are applied to the transmit power at the base station:

$$E[|\mathbf{x}(\mathbf{t})|^2] = tr(\mathbf{P}(t)\mathbf{W}^H(t)\mathbf{W}(t)) \leq P_{max}, \tag{2}$$

where $\mathbf{P}(t) = \text{diag}[\mathbf{p}_1(t), \ldots, \mathbf{p}_L(t)] \in \mathbb{C}^{K\times K}$, $\mathbf{W}(t)$ is represented as $\mathbf{W}(t) = [\mathbf{w}_1(t), \ldots, \mathbf{w}_K(t)] \in \mathbb{C}^{M\times K}$, and $P_{\max}$ represents the maximum transmit power.

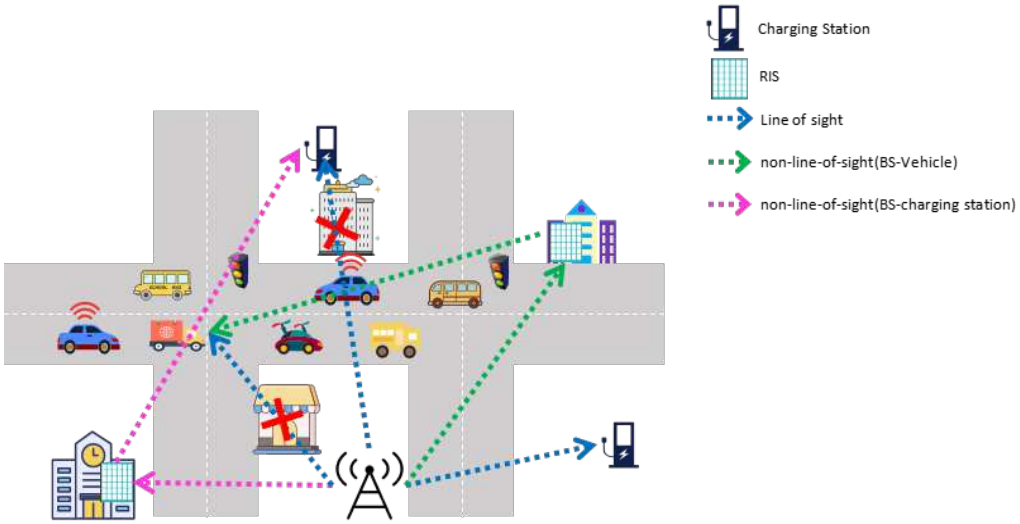


Figure 1. RIS-assisted wireless electric vehicle charging network.

2.3. RIS-Assisted Wireless Channel

We need to model two types of dynamic wireless channels in the system: one is the channel from the base station to the RIS relay, denoted as $\mathbf{H}_{BR}(t)$, and the other is the channel from the RIS relay to individual vehicle users (VUs), denoted as $\mathbf{h}_{RV,k}(t)$. The base station to the RIS channel model and the RIS to vehicle user channel model can be shown as:

Base station to RIS relay channel model:

$$\mathbf{H}_{BR}(t) = \sqrt{\beta_{BR}(t)} \times \mathbf{a}(\phi_R, \theta_R, t) \times \mathbf{a}^H(\phi_{BS}, \theta_{BS}, t) \tag{3}$$

Here, $\sqrt{\beta_{BR}(t)}$ represents the time-varying channel gain from the base station to the RIS relay. For the transmission data process from the BS to the RIS relay, the presentation of the array response vectors for multi-RIS units is denoted as $\mathbf{a}(\phi_{BS}, \theta_{BS}, t)$ and $\mathbf{a}(\phi_R, \theta_R, t)$. Specifically, $\mathbf{a}(\phi_{BS}, \theta_{BS}, t) = [a_1(\phi_{BS}, \theta_{BS}, t), \ldots, a_N(\phi_{BS}, \theta_{BS}, t)]^T \in \mathbb{C}^{N\times 1}$ and $\mathbf{a}(\phi_{RIS}, \theta_R, t) = [a_1(\phi_R, \theta_R, t), \ldots, a_M(\phi_R, \theta_R, t)]^T \in \mathbb{C}^{M\times 1}$. Next, the wireless channel model from the RIS relay to the user equipment (VU_k) is described as follows:

$$\mathbf{h}_{RV,k}(t) = \sqrt{\beta_{RV,k}(t)} \times \mathbf{a}^H(\phi_{RV,k}, \theta_{RV,k}, t) \tag{4}$$

Here, $\sqrt{\beta_{RV,k}(t)}$ characterizes the time-varying channel gain from the RIS relay to vehicle user k at time t ($k \in [1, \ldots, K]$), and $\mathbf{a}(\phi_{RV,k}, \theta_{RV,k}, t)$ represents the multi-antenna array response vector from the RIS relay to vehicle user k, defined as $\mathbf{a}(\phi_{RV,k}, \theta_{RV,k}, t) = [a_1(\phi_{RV,k}, \theta_{RV,k}, t), \ldots, a_M(\phi_{RV,k}, \theta_{RV,k}, t)]^T \in \mathbb{C}^{M\times 1}$.

In the context of the non-line-of-sight (NLOS) situation of the communication systems, the time-varying signal-to-interference-plus-noise ratio (SINR) for user k (where $k \in (1, \ldots, K)$) can be obtained as follows:

$$\gamma_k(t) = \frac{p_k(t)|(\mathbf{h}_{RV,k}^H(t)\Phi_k(t)\mathbf{H}_{BR,k}(t))\mathbf{q}_k(t)|^2}{\sum_{j\neq k}^K p_j(t)|\mathbf{h}_{RV,k}^H(t)\Phi_k(t)\mathbf{H}_{BR,k}(t))\mathbf{q}_j(t)|^2 + \sigma_k^2}, \tag{5}$$

Moreover, the spectral efficiency (SE) of the real-time system, measured in bps/Hz, can be expressed as:

$$\mathcal{R}(t) = \sum_{k=1}^{K} log_2(1 + \gamma_k(t)), \tag{6}$$

2.4. Causal Factors in RL Structure

2.4.1. Causal Graph

A directed acyclic graph (DAG) [32] is a finite graph $G = (V, E)$ consisting of a set of vertices V and a set of directed edges E, where each edge $e \in E$ is an ordered pair (u,v) indicating a direct connection from vertex u to vertex v. A DAG does not contain any directed cycles, meaning there is no sequence of edges that starts and ends at the same vertex by following the direction of the edges. Assigning a value to a particular variable X is denoted as an action or intervention. Let Pa_X denote the parent nodes of variable X; if variable X undergoes an intervention, according to the backdoor criterion, all edges from Pa_X to X are eliminated.

2.4.2. Structural Causal Model

The wireless environment exhibits ubiquitous causality, leading to causal changes in the wireless channel over time. Obtaining the causality of the time-varying wireless channel enables efficient modeling even with limited channel measurements. The key to representing wireless channel causality lies in developing suitable structural causal models (SCMs) [33]. In this paper, we denote the SCM as M, which is a tuple $< \mathbf{U}, \mathbf{V}, \mathcal{F}, P(u) >$. $\mathbf{V} = [V_1, \ldots V_n]$ represents a set of endogenous variables, i.e., variables influenced by other variables in the study. $\mathbf{U} = [U_1, \ldots, U_m]$ represents a set of exogenous variables, i.e., variables in the study not influenced by other variables. A set of structural functions determining V is defined as $\mathcal{F} = [f_1, \ldots, f_n]$. $P(u)$ represents the distribution over $\mathbf{U}$.

Next, we formalize the multi-RIS assisted wireless system in the causal domain as Figure 2 shows.

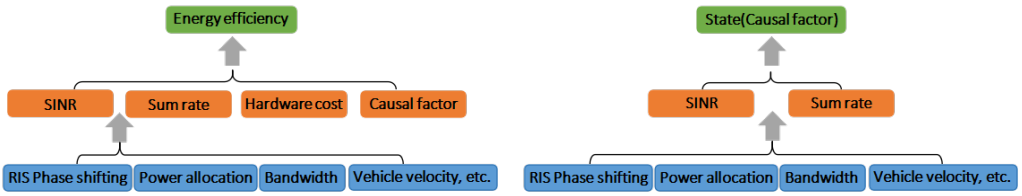


Figure 2. The causal graphs for both reward (**left**) and state (**right**) depict the situation of the state s_t in the RIS-assisted wireless electric vehicle charging network. In this graphical representation, all elements are classified into three tiers: the top layer corresponds to the outcome variables, represented by the green text box; the bottom layer corresponds to exogenous variables $\mathbf{U}$, consisting of manipulable variables, depicted in blue text boxes (X); and the variables that have a direct impact on the outcome variables (Z^R and S^R) are marked in orange, positioned between the green and blue layers. These variables are related to the internal factors denoted by $\mathbf{V}$. The interactions between these tiers are depicted by gray arrows, illustrating the intricate causal relationships facilitated by the structural functions $\mathcal{F}$. At each time step t, controllers adjust the blue nodes, influencing subsequent observations of values in the orange and green nodes.

2.5. Analysis of Reinforcement Learning and the SCM

Structural causal models (SCMs) play a crucial role in causal reinforcement learning (CRL) by providing a formal framework for representing the causal mechanisms underlying the environment. SCMs encode how variables in the environment interact with each other to produce observed outcomes, allowing agents to reason about causal relationships and make informed decisions, as in Figure 3. The following is a detailed explanation of the role of SCMs in CRL.

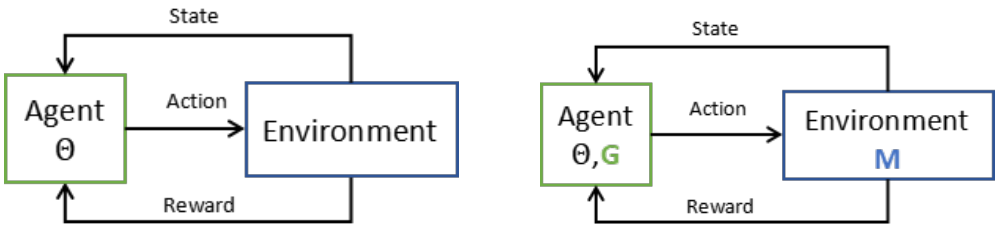


Figure 3. In the reinforcement learning structure (**left**), Θ denotes parameters concerning the environment. The agent receives feedback in the form of rewards. The agent's utility is defined by the reward function and it acts so as to maximize the expected rewards. In the causal reinforcement learning structure (**right**), G denotes the causal diagram, and M denotes the structural causal model. The environment and the agent will be linked through the combination of an SCM and a corresponding causal graph.

Representation of Causal Mechanisms: SCMs define the structural relationships between variables in the environment, including actions, states, and rewards. They specify how changes in one variable affect other variables, capturing the causal mechanisms that govern the dynamics of the environment. For example, an SCM might describe how taking certain actions influences the subsequent states of the environment and the resulting rewards received by the agent.

Causal Graph Construction: SCMs provide the foundation for constructing causal graphs, which represent the causal relationships between variables in the environment. Each node in the causal graph corresponds to a variable, and directed edges indicate causal influences between variables. SCMs specify the structure of the causal graph by defining the parents of each variable, reflecting the direct causal dependencies between variables.

Counterfactual Reasoning: SCMs enable counterfactual reasoning, allowing agents to reason about alternative scenarios and assess the causal effects of different actions. By manipulating the structural equations in an SCM, agents can simulate hypothetical interventions and predict how the environment would have behaved under different conditions. This allows agents to evaluate the causal consequences of their actions and make decisions that maximize expected rewards.

Policy Evaluation: SCMs facilitate the evaluation of policies by estimating the expected rewards associated with different action sequences. By simulating the causal mechanisms specified in an SCM, agents can compute the expected cumulative reward obtained by following a particular policy in a given environment. This allows agents to compare the effectiveness of different policies and select the one that maximizes long-term rewards.

Causal Inference: SCMs support causal inference by providing a formal framework for estimating causal effects from observational data or interventions. Agents can use techniques such as do-calculus or structural equation modeling to infer causal relationships from observed data and learn the structural parameters of the environment. This allows agents to build accurate causal models of the environment and make better decisions based on causal understanding.

In summary, SCMs play a central role in CRL by formalizing the causal relationships between variables in the environment, guiding decision-making through counterfactual reasoning and policy evaluation and facilitating causal inference from observational data. By leveraging SCMs, agents can acquire a deeper understanding of the causal structure

of their environment and make more informed and effective decisions in complex and uncertain scenarios.

3. Problem Formulation

Developing an efficient resource allocation algorithm presents a considerable challenge, primarily attributed to the mobility of users and the inherent uncertainty of the wireless channel. However, by integrating causality into the framework, we can potentially alleviate these challenges, as causality provides a means to better capture and understand uncertainties within the system.

Initially, we conduct an analysis of power consumption throughout the resource allocation process and formulate the optimization problem with a focus on enhancing energy efficiency. Subsequently, leveraging the causal structure within reinforcement learning (RL), we reframe the problem into a causal Markov Decision Process (MDP). This approach enables us to incorporate causal relationships among variables, facilitating more informed decision-making in dynamic environments.

To address the resource allocation optimization dynamically, we propose an actor–critic reinforcement learning algorithm tailored to the causal MDP framework. By iteratively refining the policy through actor–critic updates, our algorithm aims to learn optimal resource allocation strategies that balance energy efficiency and performance.

Furthermore, we delve into the intricacies of our proposed algorithm, providing detailed explanations of its components, such as actor and critic networks, reward functions, and exploration strategies. Additionally, we discuss the training procedure and potential extensions or enhancements to our approach.

3.1. Power Consumption

The total power dissipated in the system, encompassing K users, comprises various components, including the base station transmit power (p_t), hardware static power at the base station (P_{BS}), power consumed by the RIS relay (P_M), and power consumption at the user equipment (P_{VU}). With these components considered, the total power operating on the RIS-assisted wireless network can be defined as follows

$$\mathcal{P}_{total}(t) = \sum_{k=1}^{K}(\xi p_t(t) + P_{VU}(t)) + P_{BS}(t) + P_R(t), \tag{7}$$

where $\xi \cong \nu$ with ν to evaluate the ability to effectively convert input electrical power into output radio frequency (RF) power by the power amplifier.

Considering (7) as the denominator of the energy efficiency (EE) function, then the EE performance $\eta_{EE} \cong (B \cdot \mathcal{R})/\mathcal{P}_{total}$, with B presenting the bandwidth, can be obtained using (6) and (7) as

$$\eta_{EE}(t) = \frac{B\sum_{k=1}^{K} log_2(1+\gamma_k(t))}{sum_{k=1}^{K}(\xi p_t(t) + P_{VU}(t)) + P_{BS}(t) + P_R(t)}, \tag{8}$$

3.2. Optimal Problem Formulation

As depicted in Figure 2, the goal is to maximize the energy efficiency $\eta_{EE}(t)$ by jointly optimizing the transmit power $\mathbf{P} = [p_1(t), p_2(t), \ldots, p_K(t)]$ from the BS, RIS selection and the phase shift matrix $\mathbf{\Phi} = [\boldsymbol{\phi_1(t), \phi_2(t), \ldots, \phi_M(t)}]$ from the RIS.

Markov Decision Process (MDP) formulation encompasses essential components including the state, action, transition probability function, reward, and environment. These components are elucidated as follows:

- State space: Consider S as the state space, which comprises the following constituents: (i) the channel gains for communication links: h_k^t and g_k^t, (ii) the velocity and position of intelligent vehicle agents v_k, p_k, (iii) actions involving the configuration of phase

shifting for the RIS components and power distribution of VU_k implemented at time $t-1$, and (iv) the energy efficiency at time $t-1$. Hence, S encompasses:

$$s^{(t)} = \{\{h_k^t, g_k^t\}_{k\in K}, v_k, p_k, a^{(t-1)}, \{\eta_{EE,k}^t\}_{k\in K}\} \tag{9}$$

- Action space: Symbolized as A, the action space encompasses the array of actions available to the agent. It includes the manipulation of phase shifting for individual RIS components and the adjustment of transmission power at the base station. The action $a^{(t)}$ is expressed as:

$$a^{(t)} = \{\mathbf{\Theta}, \{\mathbf{W}_k\}_{k\in K}\} \tag{10}$$

- Transition Probability Function (P): This characterizes the likelihood of transitioning between states and a particular action. Formally, it is represented as $P(s'|s,a)$, where s' denotes the subsequent state, s denotes the current state, and a signifies the action.
- Reward function: The agent is provided with an immediate reward r_i^t, representing the energy efficiency as defined in Equation (8).

$$r_k^t = \eta_{EE,k}^t \tag{11}$$

- Value Function: This indicates the anticipated cumulative reward, originating from a specific state according to a predetermined policy π. Employing V_t^π to symbolize the value function at time step t under policy π, it signifies the anticipated total of rewards beginning from $s_t = s$ and extending to the conclusion of the episode:

$$V_t^\pi(\mathbf{P}, \mathbf{\Phi}, t) = E\left[\sum_{\tau=t}^{T_F} R(s_\tau, \pi(s_\tau, \tau)|s_t = s)\right] \tag{12}$$

- The *Q-value function* denotes the anticipated return, commencing from a designated state $s_t = s$, executing a particular action $a_t = a$, and subsequently adhering to policy π. It is articulated as follows:

$$Q_t^\pi(s,a) = R(s,a) + E\left[\sum_{\tau=t+1}^{T_F} R(s_\tau, \pi(s_\tau, \tau))|s_t = s, a_t = a\right] \tag{13}$$

In line with the foundational principles of optimal control theory [34], the optimal value function, along with the optimal policies for optimal resource allocation modulation, can be derived as follows:

$$V^*(\mathbf{\Phi}, \mathbf{P}, t) = \max_{\mathbf{u}_\Phi, \mathbf{u}_P} V(\mathbf{\Phi}, \mathbf{P}, t) \tag{14}$$

Here, $\mathbf{u}_P \in \mathbb{C}^{N\times N}$ and $\boldsymbol{u}_\Phi \in \mathbb{C}^{M\times M}$ represent the resource allocation control policies. Specifically, $\mathbf{u}_P$ pertains to the transmit power control policy, while $\boldsymbol{u}_\Phi$ corresponds to the RIS phase shift control policy.

Additionally, adhering to Bellman's principle of optimality [35], the dynamic representation of the finite horizon optimal cost function unfolds as follows:

$$V^*(\mathbf{\Phi}, \mathbf{P}, t) = \max_{\mathbf{u}_\Phi, \mathbf{u}_P} \{r(\mathbf{\Phi}, \mathbf{P}, t)\} + V^*(\mathbf{\Phi}, \mathbf{P}, t+1) \tag{15}$$

3.3. Causal MDP Formulation

In this section, we introduce causal factors that influence the state variables within the framework. Actions in causal MDPs are depicted as interventions. At each state s, we construct the reward graph $G_R(s)$. The variables representing rewards are R and states are denoted by S. The agent can modify variables X_I, but it is not allowed to intervene on the parent variables of R (represented as $Z^R = Pa_R$) or the parent variables of S (represented

as $Z^S = Pa_S$). As described in Section 3.3, an intervention applied to an action of size m is symbolized as $do(X_{\text{sub}} = x)$, where $|X_{\text{sub}}| = m$ and $x = [x_1, \ldots, x_m]$.

At each state s, the causal graph $G_R(s)$ includes the variables $X^R = [X_I, Z_R, R]$. It is important to note that the identity of variables in the causal graphs remains consistent across states, although the underlying distributions may vary. Detailed explanations of these notations are provided in Figure 2. Utilizing this causal knowledge, we define the transition probability function as $P(z' \mid z, a, s)$, and the reformulation of the reward function becomes:

$$R(s,a) = \sum_{z \in Z} R(s, Z = z) P(Z = z | s, a) \tag{16}$$

Consider the function called $R(s, Z = z)$, which tells us the expected reward when we have a certain state and parent pair. Now, we will introduce another function called the Q-value function, denoted as $q_h^\pi : S \times Z \to R$. This function is all about giving us the total rewards we can expect under a policy π, starting from a given state $s_t = s$ and a parent state $z_t = z$ and continuing until the episode ends:

$$q_t^\pi(s,z) = R(s,z) + E\left[\sum_{\tau=t+1}^{T_F} R(s_\tau, \pi(s_\tau, \tau)) | s_t = s, z_t = z\right] \tag{17}$$

According to the Bayesian Law of total probability formula, $Q_t^\pi(s,a)$ can be expressed as the sum over all possible states z in Z of the conditional probability of Z being z given the current state s and action a, denoted by $P(Z = z|s,a)$, multiplied by the expected reward $q_t^\pi(s,z)$. Expressing $Q_t^\pi(s,a)$ as $\sum_{z \in Z} P(Z = z \mid s,a) q_t^\pi(s,z)$ is a consequence of considering an MDP enhanced with dynamic causal graphs G^R and G^S. This formulation can be captured by a tuple $M_C = \langle S, A, P, R, T, G^R, G^S \rangle$.

4. Causal MDP-Based RIS-Assisted Resource Allocation Optimization with Online Reinforcement Learning

An ensemble DNN algorithm is provided to address the problem of the probability distribution of the causal factor Z required in Equation (17). Then, based on the obtained causal information, to tackle Equation (15), incorporating intervened actions from (10), we employ the asynchronous advantage actor–critic(A3C) algorithm to optimize the entire network output. The structure of our proposed algorithm is illustrated in Figure 4.

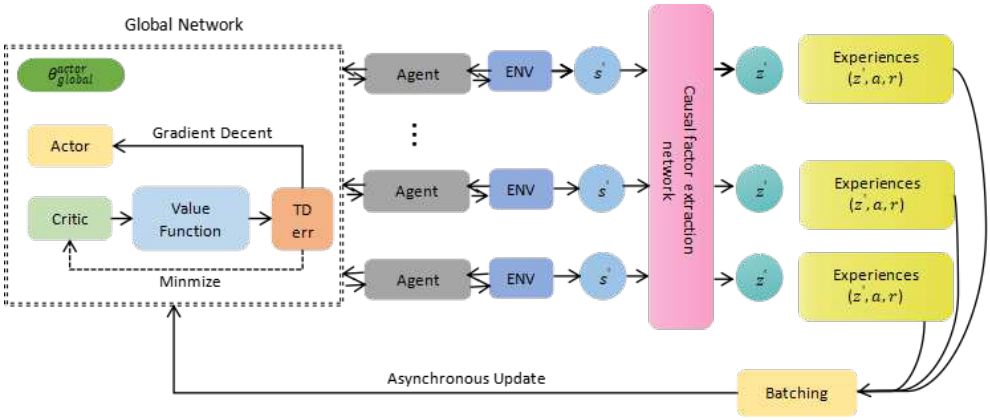


Figure 4. A3C network structure.

4.1. Causal Factor Encoder and Decoder with a Deep Neural Network

There are L electric vehicle (EV) users in the charging network and L corresponding subcarriers in the network, where each subcarrier's state information is represented as the concatenation of $\mathbf{s}_i$ described in Equation (9) and a described in Equation (10), $[s_t^i, a_t] \in \mathbb{R}^d$, where $i = 1, 2, \ldots, L$, and d is the dimensionality of the state features. The input sequence is represented as $\mathbf{S} = [[\mathbf{s}_t^1, a_t], [\mathbf{s}_t^2, a_t], \ldots, [\mathbf{s}_t^L, a_t]] \in \mathbb{R}^{d \times L}$. The corresponding feature

representation $Z_t^i \in \mathbb{R}^{d\times 1}$ can be regarded as the global environmental information for subcarrier i, s_j^t is subcarrier j's state and α_{ij} is a weight for s_t^j:

$$z_t^i = \text{Softmax}\left(h_t^j\right) = \frac{\exp\left(\text{relu}\left[\alpha^T\left[h_t^i||h_t^j\right]\right]\right)}{\sum_{v\in N_i}\left(\text{relu}\left[\alpha^T\left[h_t^i||h_t^v\right]\right]\right)}, \tag{18}$$

where $\alpha \in \mathbb{R}^{d\times 1}$ is the attention vector.

For the state set of the subcarrier i's neighborhood L_i, i.e., $s_t^{L^i} = \{s_j^t\}_{j\in Ni}$, we employ fully connected (FC) layers (i.e., a shared weight matrix $W_i \in \mathbb{R}^{d\times 1}$, where d is set to 256 in the simulation section) to transform the input state s_j^t of each subcarrier j and then obtain the embedding $z_j^t \in \mathbb{R}^{d\times 1}$. Thus, there will be a total of $|L_i|$ embeddings, i.e., $\{z_j^t\}_{j\in L_i}$. Finally, an attention mechanism layer aggregates all these embeddings to obtain $Z_t^{L_i} \in \mathbb{R}^{d\times 1}$, which can be considered as the current global embedding of subcarrier i. The network architecture is shown in Figure 5. The complementation is provided in Algorithm 1 below.

Algorithm 1 Causal factor extraction based on a self-attention mechanism

1: **Input:** State set of each subcarrier s_j^t, $\forall j$; subcarrier neighborhood sets L_i, $\forall i$; weight matrix W_i; attention vector l
2: **Position Encoding:**
3: Initialize the position encoding for each state vector to consider the sequence order.
4: **Encoder:**
5: **for** each time step **do**
6: Concatenate the state vectors of all subcarriers into a matrix X.
7: **for** encoder layer **do**
8: Pass the output through a feed-forward neural network layer to capture non-linear relationships.
9: Compute the self-attention scores for each subcarrier's state vector.
10: Apply the self-attention mechanism to obtain the weighted sum of each user's state representation.
11: **end for**
12: Obtain the causal factor for each subcarrier from the final encoder layer output.
13: **end for**
14: **return** Causal factor for each subcarrier.

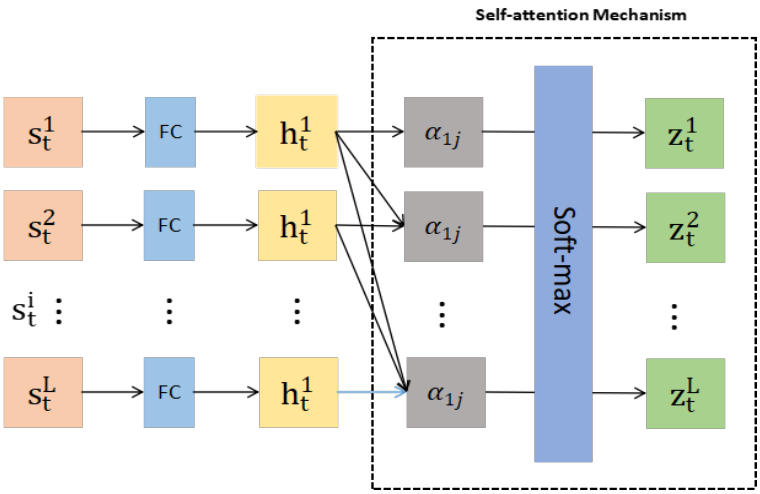


Figure 5. Causal factor extraction network architecture.

After the last hidden layer, the extracted causal factor representation for the system can be utilized for subsequent tasks, which is the resource allocation optimization of the communication network. The algorithm is provided in the following.

4.2. An Algorithm Utilizing Causal Factor-A3C for RIS Phase Shifting and Power Allocation

The overall concept is shown in Figure 6. The causal factor extraction network learns from the SCM of the environment and trains on the agent networks.

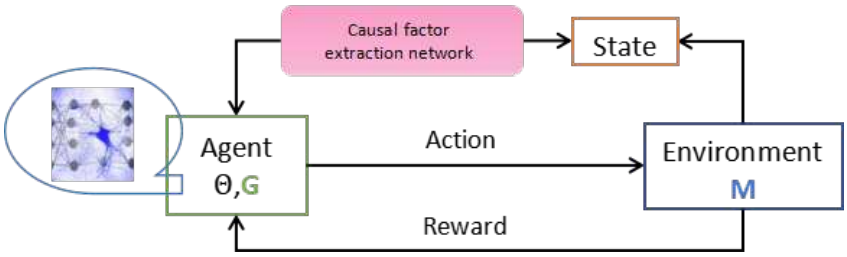


Figure 6. Overall concept.

A3C facilitates parallel training by allowing multiple worker threads to interact with their copies of the environment and update the global network asynchronously.

The A3C algorithm utilizes an asynchronous architecture, enabling parallel training across multiple worker threads. Each worker thread interacts with its instance of the environment and updates the global actor and critic networks asynchronously.

At the core of the A3C method lies the central actor network, responsible for determining which actions to take based on the current situation represented by z. Guided by its parameters θ_{actor}, this network is fine-tuned to select the best possible action given the state, denoted as $a = \pi(z|\theta_{\text{actor}})$. Meanwhile, the critic network evaluates the value of chosen actions by estimating the Q-value function $Q(z, a|\theta_{\text{critic}})$, where θ_{critic} represents the critic network's parameters. These parameters are adjusted during training to capture the overall long-term rewards represented by $q(\cdot)$.

In our A3C approach (Algorithm 2), we directly update the global actor and critic networks based on experiences collected by multiple worker threads. Here is how the A3C network would be regenerated without the replay buffer.

Algorithm 2 CF-A3C-based resource allocation optimization algorithm

Input: Global network parameters θ^{global}, number of threads N_{threads}
Initialize global network parameters θ^{global}
for thread = 1 to N_{threads} **in parallel do**
  Initialize thread-specific environment and network parameters $\theta^{\text{local}} \leftarrow \theta^{\text{global}}$
  Initialize episode counter $episodes = 0$
  while not done **do**
    Receive initial observation state s_1
    Put state s_t into the causal factor extraction network to obtain causal factor z_t
    for t = 1 to $T_{\max}$ **do**
      Select action $a_t = \pi(z_t|\theta^{\text{local}})$ using policy network
      Execute action a_t and observe reward r_t and new state s_{t+1}
      Put state s_{t+1} into the causal factor extraction network to obtain causal factor z_{t+1}
      Perform gradient ascent on global network parameters using the observed transition
      Synchronize local network parameters with global network parameters
    end for
    Increment episode counter $episodes$
  end while
end for

Global actor and critic networks are initialized with parameters θ^{actor}global and θ^{critic}global. Multiple worker threads are created, each with its instance of the environment and local actor and critic networks. Each worker thread interacts asynchronously with its environment. At each time step t, the worker receives the current state s_t from the environment and puts it into the causal factor extraction network to get the causal output z_t. They then select an action a_t according to the policy $\pi(z_t|\theta_{\text{actor}})$, execute the action a_t and observe the reward r_t and the next state s_{t+1}. Also, they obtain z_{t+1} through the same process. They store the experience tuple (z_t, a_t, r_t, z_{t+1}), calculate the advantage function using the sampled batch, update the actor parameters $\theta^{\text{actor}}_{\text{global}}$, and calculate the loss for the critic network using the sampled batch to update the critic parameters $\theta^{\text{critic}}_{\text{global}}$.

The gradients computed by each worker are applied asynchronously to the global actor and critic networks. The global networks are updated using the gradients from each worker, ensuring that all workers are trained on the most recent version of the networks.

This approach eliminates the need for a replay buffer, allowing for direct updates to the global networks based on fresh experiences collected by multiple workers. It promotes efficient exploration, accelerates training, and improves the overall performance by leveraging the A3C framework in optimizing resource allocation tasks.

5. Simulation

5.1. Simulation Setup

In this section, we reveal the results of our novel optimization technique using computer simulations, focusing on addressing the resource allocation problem within communication networks for electric vehicle charging, enhanced by RISs. The results underscore the efficacy of our approach in enhancing the aggregate transmission rate, thus highlighting its competitive edge in addressing this pertinent issue.

We consider an RIS-assisted electric vehicle charging network communication system, where a base station equipped with N = 4 antennas serves L = 4 vehicle users, and the auxiliary RIS is equipped with M = 8 × 8 reflector elements. The vehicles are uniformly distributed and move freely on a road with a width of 10 m and a length of 20 m at a speed of vs. = 18 km/h. The carrier frequency and bandwidth are set to fc = 5.9 GHz and B = 20 MHz, respectively. The channel matrices between the base station and RIS and between the base station and users, denoted as $\mathbf{H}_{BR}$ and $\mathbf{H}_{RR}$, respectively, follow dynamic Rayleigh distributions. The system parameters are summarized in Table 2.

Table 2. Simulation parameters.

Parameter	Value	Parameter	Value
BS antenna num. N	4	RIS element number	8×8
Number of electric vehicle users	4	Bandwidth B	20 MHz
BS transmission power	20 dBm	VU hardware cost power	10 dBm
RIS hardware power	10 dBm	Path loss/1 m	−30 dBm
Target SINR threshold	20 dBm	Power of noise	−80 dBm
Hidden layer size (actor)	256	Hidden layer size (critic)	256
learning rate (actor)	0.005	Learning rate (critic)	0.01
Discount factor	0.99	Optimizer	Adam
Max training steps	10,000	Training batch size	54

For comparison, we employ four benchmark schemes: (1) a phase control scheme based on deep deterministic policy gradient (DDPG) technology, (2) a phase control scheme

based on joint transmit beamforming and a phase shift design method, (3) a random phase shifting scheme, where the phase shifting of RIS reflector elements is randomly configured, and (4) a traditional scheme without RISs, utilizing only direct links between the base station and users.

5.2. Performance of the Online Causal-Factor-A3C-Based Algorithm for Optimal Resource Allocation

We use the Rayleigh distribution [36] to simulate fading conditions in wireless channels, especially in scenarios where there is no line-of-sight path. Additionally, we adopt the Monte Carlo method to simulate the dynamic behavior of channels in real-world scenarios, as it can account for various uncertainty factors and generate random samples that follow the expected distribution.

(1) Spectral Efficiency and Energy Efficiency with Optimal Resource Allocation vs. number of BS antennas and RIS units

The developed algorithm will optimize the control of RIS phase shifting and allocation of transmit power to stimulate the full potential of multi-RIS-assisted wireless networks. The performance of the developed reinforcement learning algorithm based on CF-A3C is illustrated below. Figure 7 compares the different learning process numbers of RIS units ($M = 2, 4, 8, 12$). As shown in Figure 8, increasing the number of RIS units can improve energy efficiency, and the proposed CF-A3C algorithm significantly outperforms other benchmark algorithms.

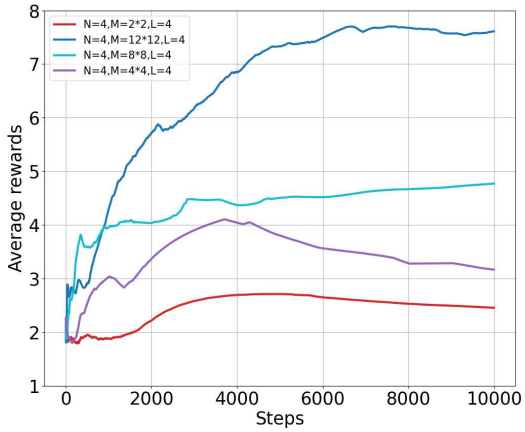


Figure 7. Average rewards vs. num. of RIS units with M = 2 × 2, 4 × 4, 8 × 8, 12 × 12.

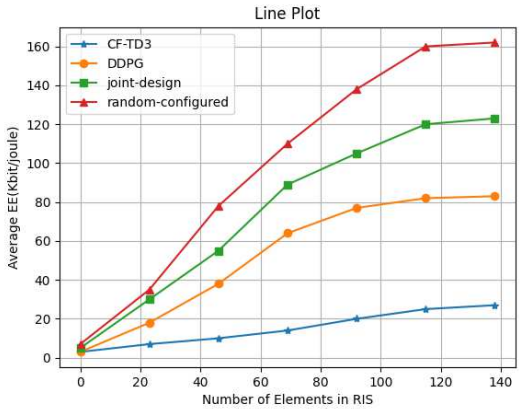
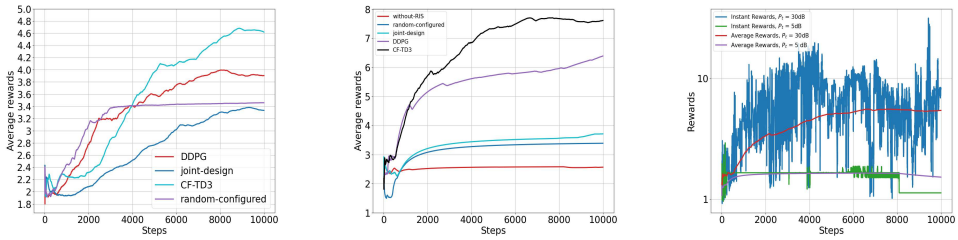


Figure 8. Average EE vs. num. of RISx with different methods.

(2) Evaluation of Online Learning Performance

Subsequently, an assessment was conducted to observe the learning dynamics of the average reward of the network across successive time steps. As depicted in Figure 9, there

is a noticeable upward trend in the average reward as time progresses. Moreover, the newly devised resource allocation algorithm, leveraging CF-A3C reinforcement learning, showcases its ability to approach the optimal solution efficiently within a set timeframe. This contrasts with the baseline algorithm, as our approach maintains convergence even amidst the variability of wireless channels.



(**a**) (**b**) (**c**)

Figure 9. The average rewards vs. time steps. (**a**) Average rewards vs. time steps. (**b**) Average rewards vs. time steps. (**c**) Average reward vs. time steps under different P_{max}.

6. Conclusions

This paper introduces a pioneering causal-factor-based A3C learning approach tailored for optimizing the communication system of RIS-assisted electric vehicle charging networks within finite time constraints. Distinguished from conventional methods, our algorithm leverages online causal learning to discern optimal RIS deployment and resource allocation strategies. Through dedicated causal factor extraction networks, it distils pertinent causal insights from the system state, enabling maximal exploitation of RIS potential. Subsequently, the A3C reinforcement learning algorithm, grounded on causal factors, adeptly learns the optimal transmit power and RIS phase adjustments, thereby bolstering wireless network performance metrics like energy efficiency, even amidst real-time uncertainties and constrained training data. Simulation-based comparisons with existing algorithms affirm the effectiveness of our proposed approach.

Author Contributions: Conceptualization, Y.Z. and H.X.; Methodology, Y.Z. and H.X.; Software, Y.Z.; Writing—original draft, Y.Z.; Supervision, H.X. All authors have read and agreed to the published version of the manuscript.

Funding: The support of the National Science Foundation (Grant No. 2128656) is gratefully acknowledged.

Data Availability Statement: Due to the involvement of our research data in another study, we will not provide details regarding where data supporting the reported results can be found.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dimitriadou, K.; Rigogiannis, N.; Fountoukidis, S.; Kotarela, F.; Kyritsis, A.; Papanikolaou, N. Current trends in electric vehicle charging infrastructure; opportunities and challenges in wireless charging integration. *Energies* **2023**, *16*, 2057. [CrossRef]
2. Umoren, I.A.; Shakir, M.Z.; Tabassum, H. Resource efficient vehicle-to-grid (V2G) communication systems for electric vehicle enabled microgrids. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 4171–4180. [CrossRef]
3. Gyawali, S.; Xu, S.; Qian, Y.; Hu, R.Q. Challenges and solutions for cellular based V2X communications. *IEEE Commun. Surv. Tutor.* **2020**, *23*, 222–255. [CrossRef]
4. Lu, J.; Li, X. The Benefits of Hydrogen Energy Transmission and Conversion Systems to the Renewable Power Grids: Day-ahead Unit Commitment. In Proceedings of the 2022 North American Power Symposium (NAPS), Salt Lake City, UT, USA, 9–11 October 2022; pp. 1–6. [CrossRef]
5. Răboacă, M.S.; Bizon, N.; Thounthong, P. Intelligent charging station in 5G environments: Challenges and perspectives. *Int. J. Energy Res.* **2021**, *45*, 16418–16435. [CrossRef]
6. Salahdine, F.; Han, T.; Zhang, N. 5G, 6G, and Beyond: Recent advances and future challenges. *Ann. Telecommun.* **2023**, *78*, 525–549. [CrossRef]

7. Verma, S.; Kawamoto, Y.; Fadlullah, Z.M.; Nishiyama, H.; Kato, N. A survey on network methodologies for real-time analytics of massive IoT data and open research issues. *IEEE Commun. Surv. Tutor.* **2017**, *19*, 1457–1477. [CrossRef]
8. Atzori, L.; Iera, A.; Morabito, G.; Nitti, M. The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization. *Comput. Netw.* **2012**, *56*, 3594–3608. [CrossRef]
9. Liu, W.; Ding, J.; Zheng, J.; Chen, X.; Chih-Lin, I. Relay-assisted technology in optical wireless communications: A survey. *IEEE Access* **2020**, *8*, 194384–194409. [CrossRef]
10. Mohr, J.J.; Sohi, R.S. Communication flows in distribution channels: Impact on assessments of communication quality and satisfaction. *J. Retail.* **1995**, *71*, 393–415. [CrossRef]
11. Yang, Y.; Dang, S.; He, Y.; Guizani, M. Markov decision-based pilot optimization for 5G V2X vehicular communications. *IEEE Internet Things J.* **2018**, *6*, 1090–1103. [CrossRef]
12. Araghi, A.; Khalily, M.; Safaei, M.; Bagheri, A.; Singh, V.; Wang, F.; Tafazolli, R. Reconfigurable Intelligent Surface (RIS) in the Sub-6 GHz Band: Design, Implementation, and Real-World Demonstration. *IEEE Access* **2022**, *10*, 2646–2655. [CrossRef]
13. Choi, J.H.; Itoh, T.; Chen, Z.; Liu, D.; Nakano, H.; Qing, X.; Zwick, T. Beam-scanning leaky-wave antennas. *Handb. Antenna Technol.* **2016**, *3*, 1698–1732.
14. Araghi, A.; Khalily, M.; Xiao, P.; Tafazolli, R. Holographic-based leaky-wave structures: Transformation of guided waves to leaky waves. *IEEE Microw. Mag.* **2021**, *22*, 49–63. [CrossRef]
15. Huang, K.; Kanaroglou, P.; Zhang, X. The design of electric vehicle charging network. *Transp. Res. Part D Transp. Environ.* **2016**, *49*, 1–17. [CrossRef]
16. Mei, A.; Morabito, G.; Santi, P.; Stefa, J. Social-aware stateless forwarding in pocket switched networks. In Proceedings of the 2011 Proceedings IEEE INFOCOM, Shanghai, China, 10–15 April 2011; pp. 251–255.
17. Madakam, S.; Ramaswamy, R.; Tripathi, S. Internet of Things (IoT): A literature review. *J. Comput. Commun.* **2015**, *3*, 164–173. [CrossRef]
18. Roopa, M.; Pattar, S.; Buyya, R.; Venugopal, K.R.; Iyengar, S.; Patnaik, L. Social Internet of Things (SIoT): Foundations, thrust areas, systematic review and future directions. *Comput. Commun.* **2019**, *139*, 32–57.
19. Gasse, M.; Grasset, D.; Gaudron, G.; Oudeyer, P.Y. Causal reinforcement learning using observational and interventional data. *arXiv* **2021**, arXiv:2106.14421.
20. Zhu, S.; Ng, I.; Chen, Z. Causal discovery with reinforcement learning. *arXiv* **2019**, arXiv:1906.04477.
21. Babaeizadeh, M.; Frosio, I.; Tyree, S.; Clemons, J.; Kautz, J. Reinforcement learning through asynchronous advantage actor-critic on a gpu. *arXiv* **2016**, arXiv:1611.06256.
22. Huang, C.; Mo, R.; Yuen, C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 1839–1850. [CrossRef]
23. Lee, G.; Jung, M.; Kasgari, A.T.Z.; Saad, W.; Bennis, M. Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces. In Proceedings of the ICC 2020–2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; pp. 1–6.
24. Xu, Y.H.; Yang, C.C.; Hua, M.; Zhou, W. Deep Deterministic Policy Gradient (DDPG)-Based Resource Allocation Scheme for NOMA Vehicular Communications. *IEEE Access* **2020**, *8*, 18797–18807. [CrossRef]
25. Saito, Y.; Kishiyama, Y.; Benjebbour, A.; Nakamura, T.; Li, A.; Higuchi, K. Non-orthogonal multiple access (NOMA) for cellular future radio access. In Proceedings of the 2013 IEEE 77th Vehicular Technology Conference (VTC Spring), Dresden, Germany, 2–5 June 2013; pp. 1–5.
26. Cauteruccio, F.; Cinelli, L.; Fortino, G.; Savaglio, C.; Terracina, G.; Ursino, D.; Virgili, L. An approach to compute the scope of a social object in a Multi-IoT scenario. *Pervasive Mob. Comput.* **2020**, *67*, 101223. [CrossRef]
27. Wang, J.; Liang, Y.C. Transmit Beamforming Design for Multiuser Multi-IoT-Device Symbiotic Radios. In Proceedings of the ICC 2023-IEEE International Conference on Communications, Rome, Italy, 28 May–1 June 2023; pp. 943–948. [CrossRef]
28. Wang, X.; Zhong, X.; Li, L.; Zhang, S.; Lu, R.; Yang, T. TOT: Trust aware opportunistic transmission in cognitive radio Social Internet of Things. *Comput. Commun.* **2020**, *162*, 1–11. [CrossRef]
29. Mekala, M.S.; Srivastava, G.; Park, J.H.; Jung, H.Y. An effective communication and computation model based on a hybridgraph-deeplearning approach for siot. *Digit. Commun. Netw.* **2022**, *8*, 900–910. [CrossRef]
30. Arif, S.; MacNeil, M.A. Applying the structural causal model framework for observational causal inference in ecology. *Ecol. Monogr.* **2023**, *93*, e1554. [CrossRef]
31. Zhang, Y.; Xu, H. Distributed Data-Driven Learning-Based Optimal Dynamic Resource Allocation for Multi-RIS-Assisted Multi-User Ad-Hoc Network. *Algorithms* **2024**, *17*, 45. [CrossRef]
32. Lipsky, A.M.; Greenland, S. Causal directed acyclic graphs. *JAMA* **2022**, *327*, 1083–1084. [CrossRef] [PubMed]
33. Mooij, J.M.; Janzing, D.; Schölkopf, B. From ordinary differential equations to structural causal models: The deterministic case. *arXiv* **2013**, arXiv:1304.7920.
34. Kirk, D.E. *Optimal Control Theory: An Introduction*; Courier Corporation: North Chelmsford, MA, USA, 2004.

35. Rosu, I. The Bellman Principle of Optimality. 2002. Available online: http://faculty.chicagogsb.edu/ioanid.rosu/research/notes/bellman.pdf (accessed on 2 May 2024).
36. Chvojka, P.; Zvanovec, S.; Haigh, P.A.; Ghassemlooy, Z. Channel characteristics of visible light communications within dynamic indoor environment. *J. Lightwave Technol.* **2015**, *33*, 1719–1725. [CrossRef]