(Received 19 May 2026, Accepted 19 July 2026, Scheduled 25 August 2026)

Multi-Objective Cooperative Control of a Three-Level NPC Active Power Filter Based on Single-Agent DDPG

Zihang Xie, Jianxun Mo, Ziqiang Liu*, Shijie Fang, Zibo Wang, Wenbiao Li, and Mingrui Chen

School of Electrical Engineering and Automation, Jiangxi University of Science and Technology, Ganzhou 341000, Jiangxi, China

ABSTRACT: Conventional multiloop control of diodeclamped active power filters suffers from response conflicts and slow regulation due to strong coupling among DC voltage, neutralpoint potential, and reactive power. This paper proposes a singleagent deep deterministic policy gradient (DDPG) framework for multiobjective cooperative control. Unlike generic DRL applications, our design retains the current PI inner loop for hardware safety. At the same time, a single agent replaces the three outer loops (voltage, neutralpoint, and reactive power) to eliminate cascade coupling. A composite reward function based on the L2 norm — rather than linear weighting — quadratically penalizes large errors, forcing the agent to prioritize the most severe deviations and resolve multiobjective conflicts optimally. Simulations show that the proposed strategy reduces gridside current THD from 6.7% to 1.09%, cuts midpoint balancing and DCbus settling times by 50% and 66.67% respectively, achieves zero overshoot and zero steadystate error, and exhibits strong robustness under voltage sags and load surges. This work offers a customized, lowtuningeffort intelligent solution for power quality control in complex grids.

1. INTRODUCTION

In recent years, the rapid growth of renewable energy gridconnection systems, highfrequency power converters, and nonlinear industrial loads has significantly worsened harmonic pollution in distribution networks. This problem is particularly acute in highpowerdensity environments such as rail transport, data centers, and renewable energy charging stations, where the widespread use of rectifiers and PWM converters has caused a steady rise in current distortion [1–3]. According to the IEEE 519–2014 standard, the total harmonic distortion (THD) at the point of common coupling must generally remain below 5%. Exceeding this limit can lead to equipment overheating, relay protection malfunctions, and reduced system stability.

Active power filters (APFs) are key devices for power quality management. They have been widely adopted in industrial distribution systems and renewableenergy gridconnection scenarios because of their fast dynamic response, high compensation accuracy, and strong adaptability to complex loads [4–7].

Conventional diodeclamped threelevel active power filters (NPCAPFs) typically employ a multiloop control architecture that includes a DCvoltage outer loop, a current inner loop, a reactivepower control loop, and a neutralpoint balancing loop. In this structure, the DCvoltage loop maintains bus energy stability; the current inner loop handles harmonic compensation and current tracking; the reactivepower loop regulates the power factor; and the neutralpoint balancing loop suppresses voltage shifts across the DCside capacitors. Although this architecture is physically clear, the multiple control loops are strongly coupled dynamically. For instance, the reactive compensation current influences the DCbus energy distribution, while neutralpoint balancing alters the zerosequence component distribution, which in turn affects current tracking performance. As a result, under parameter variations or complex load conditions, conventional proportional-integral (PI) control cannot simultaneously ensure fast dynamic response, high steadystate accuracy, and sufficient robustness. Moreover, the parameters of multiple PI controllers are usually tuned manually, which makes engineering commissioning complicated and prevents global optimization.

To tackle the challenges of strong multivariable coupling and nonlinearity, many researchers have recently attempted to incorporate nonlinear control and heuristic algorithms into APF systems. Refs. [8–10] explored fuzzy control, neural networks, and slidingmode control for harmonic suppression. Although these methods improve the robustness of singleloop control to some extent, they are still constrained by the traditional "divideandconquer" philosophy and thus fail to eliminate dynamic crosscoupling fundamentally among multiple loops.

With the growing adoption of artificial intelligence, modelfree and highly adaptive deep reinforcement learning (DRL) has opened a new paradigm for global optimization of power electronic converters [11–13]. For example, Ref. [14] presents a dynamic optimization algorithm for gridconnected photovoltaic APFs based on partial reinforcement learning. It adjusts control parameters in real time through partialreinforcement principles and markedly improves harmonic suppression, power factor, and dynamic robustness. Ref. [15] proposes a multiagent DRL method for voltage/reactivepower control in distribution networks using droop control, which enhances voltage stability by minimizing activepower losses. Ref. [16] introduces an efficiencyoptimization design for threelevel

* Corresponding author: Ziqiang Liu (9120230071@jxust.edu.cn).

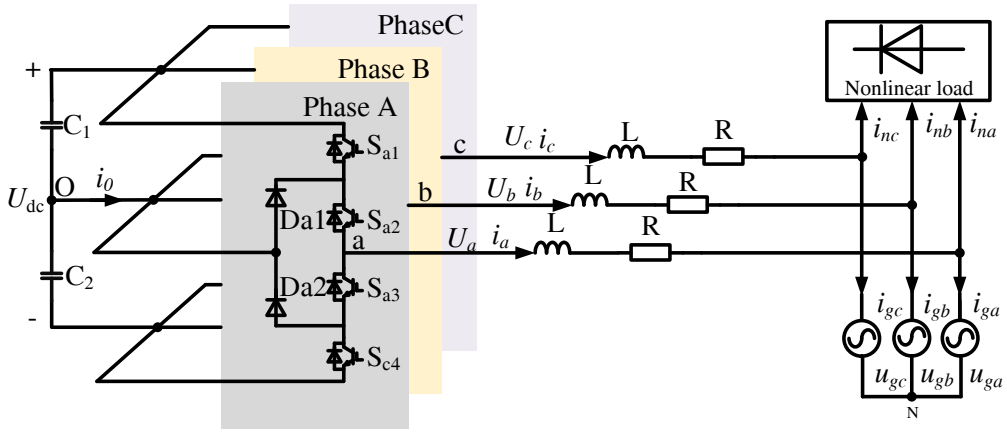

FIGURE 1. Three-level NPC active power filter topology.

active neutral-point clamped (ANPC) inverters based on the DDPG algorithm. By minimizing total power losses, it intelligently optimizes switching frequencies and hardware parameters, thereby improving the converter's adaptability to operating conditions and its overall efficiency. Ref. [17] proposes a modelfree control method for threelevel NPC converters based on deep reinforcement learning. By intelligently selecting switching states, it achieves capacitor voltage balancing and current tracking, and significantly improves robustness and dynamic performance under disturbances such as sudden operatingcondition changes and parameter drift. Ref. [18] uses a neural-network observer to compensate for parameter mismatches in three-level NPC converters, but it does not resolve the dynamic coupling among APF outer loops sharing the same current channel. This work introduces a single-agent DDPG to achieve unified optimization of the voltage, neutral-point, and reactive-power loops. However, most of these approaches only provide local improvements to traditional control architectures and do not fundamentally solve the multiloop coupling or the difficulty of parameter tuning.

To overcome these limitations, this paper proposes a multiloop cooperative control strategy based on a singleagent DDPG. Building on the existing innerloop PI control, the proposed strategy uses one agent to simultaneously replace the DCbus voltage loop, reactivepower loop, and neutralvoltage balancing loop, thus achieving unified optimization of the three objectives. We first establish a mathematical model of the three-phase NPCAPF and formulate the multiloop control problem as a Markov decision process. Then, we design the state space, action space, and reward function for multiobjective cooperative control to enable joint optimization. Finally, we verify the method through MATLAB/Simulink simulations and compare its steadystate performance, dynamic response, and robustness with those of the traditional PI scheme. The results show that our method effectively resolves multiloop coupling and significantly improves dynamic response speed, compensation accuracy, and system robustness.

2. NPC-APF TRADITIONAL MULTI-LOOP CONTROL ARCHITECTURE

The main circuit topology of a diode-clamped three-phase, three-level APF is shown in Fig. 1, where U_x ($x = a, b, c$) represents the three-phase grid voltage; R and L are the equivalent resistance and filtering inductance of the output circuit, respectively; i_x is the compensation current; U_{dc1} and U_{dc2} are the terminal voltages of the upper and lower split capacitors on the DC side, respectively; and the sum of the two, $U_{dc} = U_{dc1} + U_{dc2}$, is the APF's total DC bus voltage.

Based on the topology shown in Fig. 1 and disregarding the voltage drops across the switching devices and line losses, a mathematical model of the main circuit was established using Kirchhoff's voltage and current laws:

$$\begin{cases} \dfrac{di_a}{dt} = \dfrac{U_{ga} - Ri_a - U_a}{L} \\ \dfrac{di_b}{dt} = \dfrac{U_{gb} - Ri_b - U_b}{L} \\ \dfrac{di_c}{dt} = \dfrac{U_{gc} - Ri_c - U_c}{L} \end{cases} \tag{1}$$

where U_{gx} ($x = a, b, c$) represents the three-phase voltage of the power grid; i_x represents the APF compensation current; and U_x represents the three-phase output voltage on the AC side of the APF.

By applying the Clarke and Park transformations to convert the mathematical model from the three-phase stationary coordinate system to the rotating coordinate system, a simplified system model can be obtained as follows:

$$\begin{cases} L\dfrac{d}{dt}i_d = d_d U_{dc} - Ri_d + L\omega i_q - U_{gd} \\ L\dfrac{d}{dt}i_q = d_q U_{dc} - Ri_q + L\omega i_d - U_{gq} \end{cases} \tag{2}$$

where ω denotes the angular frequency of the grid; d_d and d_q correspond to the average values of the duty cycles of the d- and q-axes, respectively; i_d and i_q are the d- and q-axis components

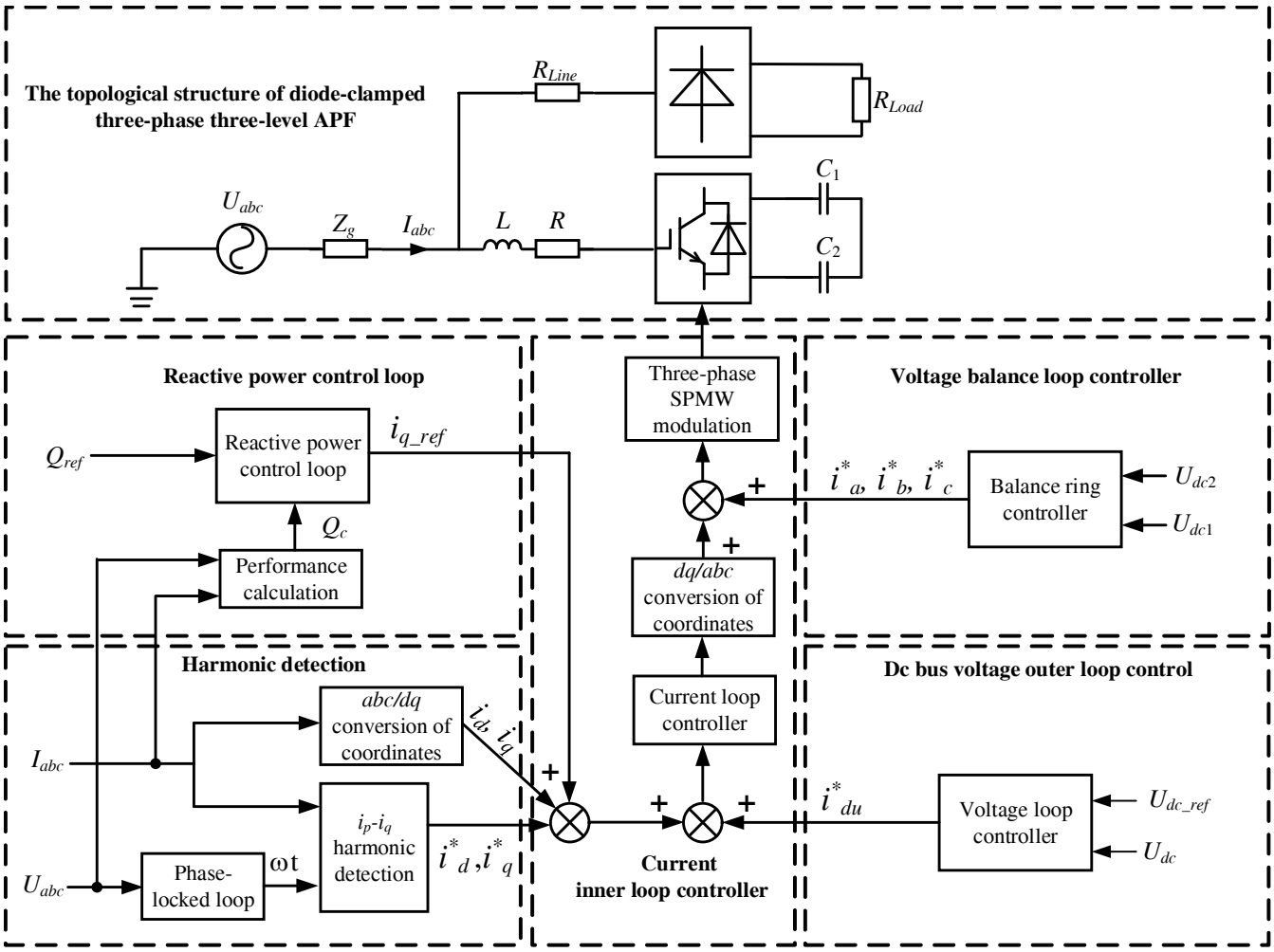

FIGURE 2. Traditional multi-loop PI control structure for NPC-APF.

of the APF compensation current; u_{gd} and u_{gq} are the average values of the three-phase grid voltages in the d- and q-axes, respectively, over the switching cycle.

It can be seen from Equation (2) that there are cross-coupling terms $L\omega i_q$ and $L\omega i_d$ between the d-axis and q-axis currents. When the system operating conditions change, this coupling relationship leads to a deterioration in the dynamic performance of conventional PI control; in particular, current oscillations and dynamic overshoot are more likely to occur under conditions of sudden load change and grid disturbance.

2.1. Outer-Loop Control of DC Bus Voltage

As shown in Fig. 2, the core control objective of the outer loop of the DC bus voltage is to maintain the total DC-side voltage U_{dc} stable at the reference value U_{dc_ref}, thereby providing stable energy support for the inverter's power conversion. When compensating for harmonics and reactive power, the APF generates switching and line losses, requiring it to draw a small amount of active power from the grid to offset these losses. This active power is controlled by the DC voltage outer loop through the adjustment of the active current reference value, with harmonic detection employing the i_p-i_q method, which is based on instantaneous reactive power theory [19].

The DC bus voltage deviation is defined as:

$$\Delta u_{dc} = U_{dc\_ref} - U_{dc} \tag{3}$$

When the voltage deviation is fed into the PI controller, the output is the fundamental component i^*_{du} of the active current reference for the d-axis, and the control law is as follows:

$$i^*_{du} = K_{pv}\Delta u_{dc} + K_{iv}\int_0^t \Delta u_{dc}dt \tag{4}$$

Here, K_{pv} and K_{iv} represent the proportional and integral coefficients of the voltage loop, respectively.

2.2. Voltage Balancing Loop Controller

In the case of the NPC three-level topology, neutral point potential drift is one of the key issues affecting the system's stable operation. When an imbalance occurs to the voltages across the upper and lower DC-side capacitors, it causes a shift in the inverter output voltage vector, thereby introducing even-order harmonics and increasing voltage stress on the switching de-

vices. In severe cases, this may even lead to device damage owing to overvoltage.

To suppress neutral potential drift, this study regulates the charging and discharging processes of the upper and lower capacitors by injecting a zero-sequence current. The neutral voltage deviation is defined as

$$\Delta u_b = U_{dc1} - U_{dc2} \tag{5}$$

The voltage deviation is fed into the PI controller to generate a zero-sequence compensation current reference value i_0^* (such that $i_a^* = i_b^* = i_c^* = i_0^*$):

$$i_0^* = K_{pb}\Delta u_b + K_{ib}\int_0^t \Delta u_b dt \tag{6}$$

where K_{pb} and K_{ib} are the proportional and integral coefficients of the equilibrium loop, respectively.

As the zero-sequence component affects the distribution of the three-phase currents, there is strong coupling between neutral-point balancing control and current compensation. This is one of the key reasons why global optimisation is difficult to achieve with conventional multi-loop PI control.

2.3. Reactive Power Control Loop

The reactive power compensation capability is a key indicator for assessing an APF's power quality management performance. By dynamically adjusting the q-axis current component, the APF can improve the system power factor and reduce fluctuations in reactive power.

In the dq coordinate system, the instantaneous reactive power of the system can be expressed as:

$$Q_C = \frac{3}{2}(i_d U_q - i_q U_d) \tag{7}$$

where i_d and i_q are the d-axis and q-axis components of the grid-side current, and U_d and U_q are the d-axis and q-axis components of the grid-side voltage.

The grid reactive power error is defined as:

$$\Delta Q = Q_{ref} - Q_C \tag{8}$$

where Q_{ref} represents the desired reactive power reference value for the power grid.

When the reactive power error is fed into the PI controller, the output is the q-axis reference current i_{ref}^q, and the control law is as follows:

$$i_{q\_ref} = K_{pq}\Delta Q + K_{iq}\int_0^t \Delta Q dt \tag{9}$$

Here, K_{pq} and K_{iq} represent the proportional and integral coefficients of the reactive power control loop, respectively.

Because reactive power compensation, voltage regulation, and neutral point balancing all share the dq current control channel, there is significant dynamic coupling between the control loops. When multiple control objectives change simultaneously, conventional cascaded PI control is prone to response hysteresis and local oscillations.

2.4. Analysis of the Limitations of Traditional Multi-Loop Control

Traditional multi-loop PI control for NPC-APF systems offers advantages such as a clear structure and straightforward engineering implementation; however, under complex operating conditions, the following issues persist:

(1) multiple control loops share the dq current regulation channel, resulting in dynamic coupling between control objectives, which can easily lead to response conflicts;

(2) PI parameters are typically tuned empirically, making it difficult to balance dynamic response and steady-state accuracy when system parameters change;

(3) traditional control methods rely on relatively accurate mathematical models, whereas APFs are highly nonlinear, high-frequency switching systems, and model parameters are easily affected by load and grid disturbances;

and (4) multi-loop cascaded structures reduce the overall dynamic response speed of the system, resulting in limited robustness under voltage sags and sudden load changes.

Therefore, it is necessary to introduce intelligent control methods with autonomous learning and global collaborative optimization capabilities to enhance the NPC-APF's comprehensive control performance under complex operating conditions.

3. A COOPERATIVE CONTROL METHOD FOR NPC-APF BASED ON SINGLE-AGENT DDPG

Deep reinforcement learning can autonomously learn control strategies through continuous interaction with the environment. It does not require precise mathematical models. This makes it particularly suitable for high-dimensional, nonlinear, and continuous control problems.

Unlike existing works that directly apply standard DDPG to APFs, our cooperative control framework does not simply replace PI regulators with agent outputs. Instead, it is purposefully designed from two perspectives: multitimescale decoupling and multiobjective conflict resolution.

First, in terms of architecture, we retain the current inner loop — which has the fastest response and the highest safety requirements — under conventional PI control. Only the three outer loops, which are dynamically slower and mutually coupled, are handed over to the agent for unified optimization. This "fixed inner loop + intelligent outer loop" hierarchical scheme avoids the safety risks of direct DRLgenerated PWM pulses. At the same time, it uses the global view of a single agent to eliminate cascade coupling fundamentally among outer loops.

Second, the reward function deliberately adopts the L2 norm rather than simple linear weighting. The three outerloop errors — voltage, neutralpoint, and reactive power — have different physical dimensions and widely varying dynamic ranges. Linear weighting cannot properly balance their priorities under extreme conditions. In contrast, the L2 norm penalizes larger errors quadratically. This forces the agent to adaptively focus its attention on the most severe deviation during training. Consequently, it makes an optimal tradeoff when multiple objectives conflict. This design is substantially different from the generic DDPG frameworks commonly reported in the literature. It rep-

resents a customized innovation that is specifically tailored to the control characteristics of APFs.

Compared to multiagent reinforcement learning, the single-agent structure eliminates the need to design communication and coordination mechanisms between agents. This reduces training complexity and mitigates the issue of state dimension explosion. It also avoids the nonstationary effects associated with multiagent policy updates. Therefore, it is more suitable for highfrequency realtime control scenarios of APFs.

3.1. Markov Decision Modelling for NPC-APF Coordinated Control

The APF control system features highfrequency discrete sampling, continuous action outputs, and strongly nonlinear dynamics. We therefore describe the NPCAPF multiobjective control process as a Markov Decision Process (MDP). This MDP is defined as a tuple (S, A, P, r, γ), where S is the state space; A is the action space; P is the state transition probability; r is the instantaneous reward; and γ is the discount factor.

At each sampling time t, the agent observes state S_t, outputs action a_t, receives reward r_t from the environment, and transitions to a new state S_{t+1}. The control objective is to learn a deterministic policy $\mu(s)$. This policy maximizes the long-term cumulative reward. It must achieve optimal control under multiple constraints, including harmonic suppression, reactive power compensation, and midpoint balancing.

3.1.1. State Space Design

The state space must fully reflect the system dynamics. This ensures that the agent can accurately perceive all control objectives and their deviations. In this study, we select nine strongly correlated state variables to form a highdimensional state space. The observation module outputs the state vector: $S_t = [U_{dc,t}, e_{dc,t}, \int e_{dc,t}dt, U_{dc1,t}, U_{dc2,t}, e_{dc12,t}, \int e_{dc12,t}dt, e_{Q,t}, \int e_{Q,t}dt]^T$, where $U_{dc,t}$ is the DC bus voltage; $e_{dc,t} = |U_{dc\_ref} - U_{dc,t}|$ represents the bus voltage deviation and is used to reflect the energy balance of the system; $\int e_{dc,t}dt$ represents the integral of the bus voltage difference; $U_{dc1,t}$ and $U_{dc2,t}$ represent the upper and lower capacitor voltages, respectively; $e_{dc12,t} = |U_{dc1,t} - U_{dc2,t}|$ is used to characterize the degree of neutral point potential deviation; $\int e_{dc12,t}dt$ represents the integral of the potential difference at the midpoint; $e_{Q,t} = |Q_{ref} - Q_c|$ is the reactive power deviation; $\int e_{Q,t}dt$ represents the integral of the difference in reactive power. Furthermore, the introduction of an error integration term preserves the system's historical dynamic information, thereby enhancing the agent's ability to perceive steady-state deviations and slow-varying dynamic processes, and preventing long-term steady-state errors during the control process.

3.1.2. Action Space

The action space directly determines the continuity and constraints of the agent's output signals. The NPCAPF must simultaneously stabilize the DC bus, balance the neutral point, and compensate reactive power. To meet these requirements, we design a threedimensional continuous action vector: $a_t = [i^*_{du}, i^*_0, i_{q\_ref}]^T$, where i^*_{du} is the fundamental active current on the d-axis, used to maintain the DC bus energy balance; i^*_0 is the compensation current reference value, used to regulate the neutral point potential; i_{q_ref} is the reference value for the qaxis reactive current, used for dynamic reactive power compensation. Because the sinusoidal pulse width modulation (SPWM) process is essentially a continuous control problem, the DDPG continuous action output method avoids the issue of insufficient regulation resolution associated with discrete action control.

To ensure that the control actions comply with the actual hardware operating constraints, boundary limits are set for each control variable. Specifically: $i^*_{du} \in [-40\,\text{A}, 40\,\text{A}]$; $i^*_0 \in [-30\,\text{A}, 30\,\text{A}]$; and $i_{q\_ref} \in [-20\,\text{A}, 20\,\text{A}]$.

These ranges account for the APF-rated capacity, switching device current stresses, and system dynamic response requirements.

3.1.3. Multi-Objective Cooperative Reward Function

The reward function is central to an agent's learning of cooperative control strategies. This study employs the L2 norm to quantify errors; compared to the L1 norm, which has a squaring effect on large deviations [20], the L2 norm is better suited to prioritising the suppression of significant voltage fluctuations.

$$\mathbf{e}(t) = [e_{dc}(t) \;\; e_{dc12}(t) \;\; e_q(t)]^T \tag{10}$$

where $e_{dc}(t) = |U_{dc\_ref} - U_{dc,t}|$ is the scalar DC voltage tracking error; $e_{dc12}(t) = |U_{dc1,t} - U_{dc2,t}|$ is the scalar capacitor voltage balancing error; and $e_q(t) = |Q_{ref,t} - Q_{c,t}|$ is the reactive power tracking error. Quantization was performed using the square term of the L2 norm, ensuring that large deviations incur a penalty far greater than small deviations, thereby guiding the agent to prioritize the elimination of significant voltage fluctuations. The quantization expression is shown in (11):

$$H_t = \|\mathbf{e}(t)\|_2 = \sqrt{e^2_{dc}(t) + e^2_{dc12}(t) + e^2_q(t)} \tag{11}$$

where H_t represents the comprehensive error penalty term. The L2 norms of the DC voltage tracking error, capacitive voltage balancing deviation, and reactive power tracking error, together with the system stability constraint term, form the core of the reward function. The L2 norm is used to precisely quantify voltage errors and implement gradient-based error penalization, and the system stability constraint term suppresses overshoot in the agent output and prevents system oscillations. The quantification expression is as follows:

$$r_t = 100 - \mu H_t - \beta \cdot |a_t| \tag{12}$$

where r_t represents the reward function; μ ($\mu > 0$) is the comprehensive error weighting coefficient, which governs the penalty intensity for regulating both DC voltage stability and midpoint voltage balance; βa_t is the system stability constraint term, where a_t is the control action vector output by the agent, and β is the action penalty coefficient, which suppresses excessive control actions. The values of these two coefficients are strictly tied to the APF hardware constraints and dynamic characteristics of the system. Through simulation and experimental

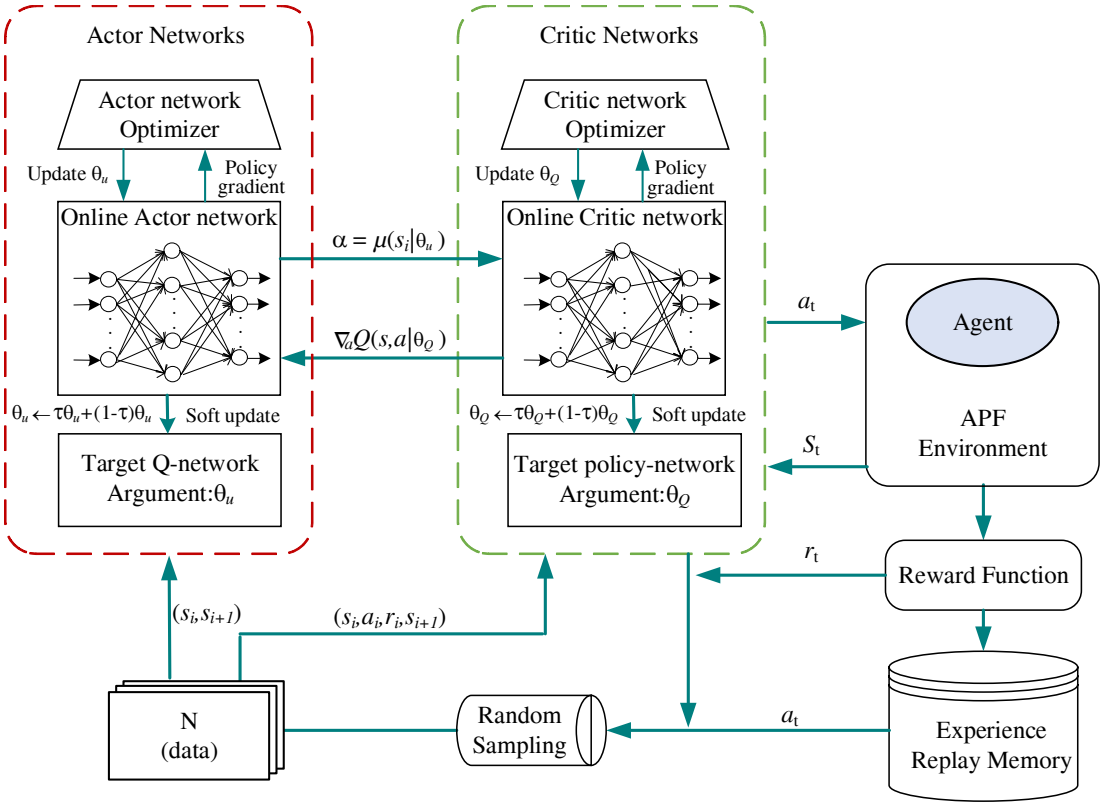

FIGURE 3. NPC-APF cooperative control framework based on single-agent DDPG.

calibration, this study sets $\mu = 0.4$ and $\beta = 1$, which satisfies the requirements for steady-state accuracy and dynamic response simultaneously.

3.2. Network Architecture of the DDPG Algorithm

The NPCAPF control problem involves continuous actions. Thus, we adopt the actorcritic-based DDPG algorithm. As shown in Fig. 3, it directly outputs continuous control signals. This avoids quantization errors common in discrete action reinforcement learning (RL) for high frequency PWM control.

The algorithm consists of four deep neural networks: the online policy network, the target policy network, the online Q-network, and the target Qnetwork.

(1) Online Policy Network (Actor): Based on the policy gradient method, it takes the system state s as input and outputs a deterministic control action $\alpha = \mu(s_i|\theta_u)$, where θ_u is the parameter of the actor network; the network optimizes the control action through policy gradient updates.

(2) Online Critic Network: Based on the temporal difference method, it takes as input the system state s and the action a output by the actor network, and outputs the q-value of the action $\nabla_\alpha Q(s, a|\theta_Q)$, where θ_Q are parameters of the critic network. The network is updated by minimizing the temporal-difference error, providing the gradient update basis for the actor network.

(3) Target Network: Used to stabilize the training process, corresponding to both the actor and critic networks, it reduces training oscillations through soft updating methods.

4. SIMULATION VALIDATION AND CONTROL PERFORMANCE ANALYSIS

4.1. Simulation Model and Parameter Settings

To validate the effectiveness of the proposed single-agent DDPG cooperative control strategy, a discretized simulation platform for a three-phase, three-level NPC-APF was established by MATLAB/Simulink, as shown in Fig. 5. The simulation model comprises a nonlinear load, grid-side model, PWM module, and APF control system, enabling the simulation of actual harmonic compensation operations in distribution networks.

Considering the typical operating conditions in industrial low-voltage distribution systems, the system's rated phase voltage was set to 220 V and the grid frequency to 50 Hz. The output filter inductance of the APF was set to 2 mH to balance the high-frequency current ripple suppression capability with the system dynamic response speed. The PWM switching frequency was set to 8 kHz, which strikes a balance between reducing switching losses and ensuring current tracking performance. The DC-side split capacitors were all set to 3000 μF to meet the requirements for busbar energy storage and neutral potential stabilization.

The DDPG network training was run on a platform comprising an Intel Core i5-14460H CPU, a 32 GB of RAM, and an NVIDIA RTX 3060 GPU. The algorithm hyperparameter settings are shown in Table 1, where the discount factor is set to 0.94 to enhance the weighting of long-term rewards, and the

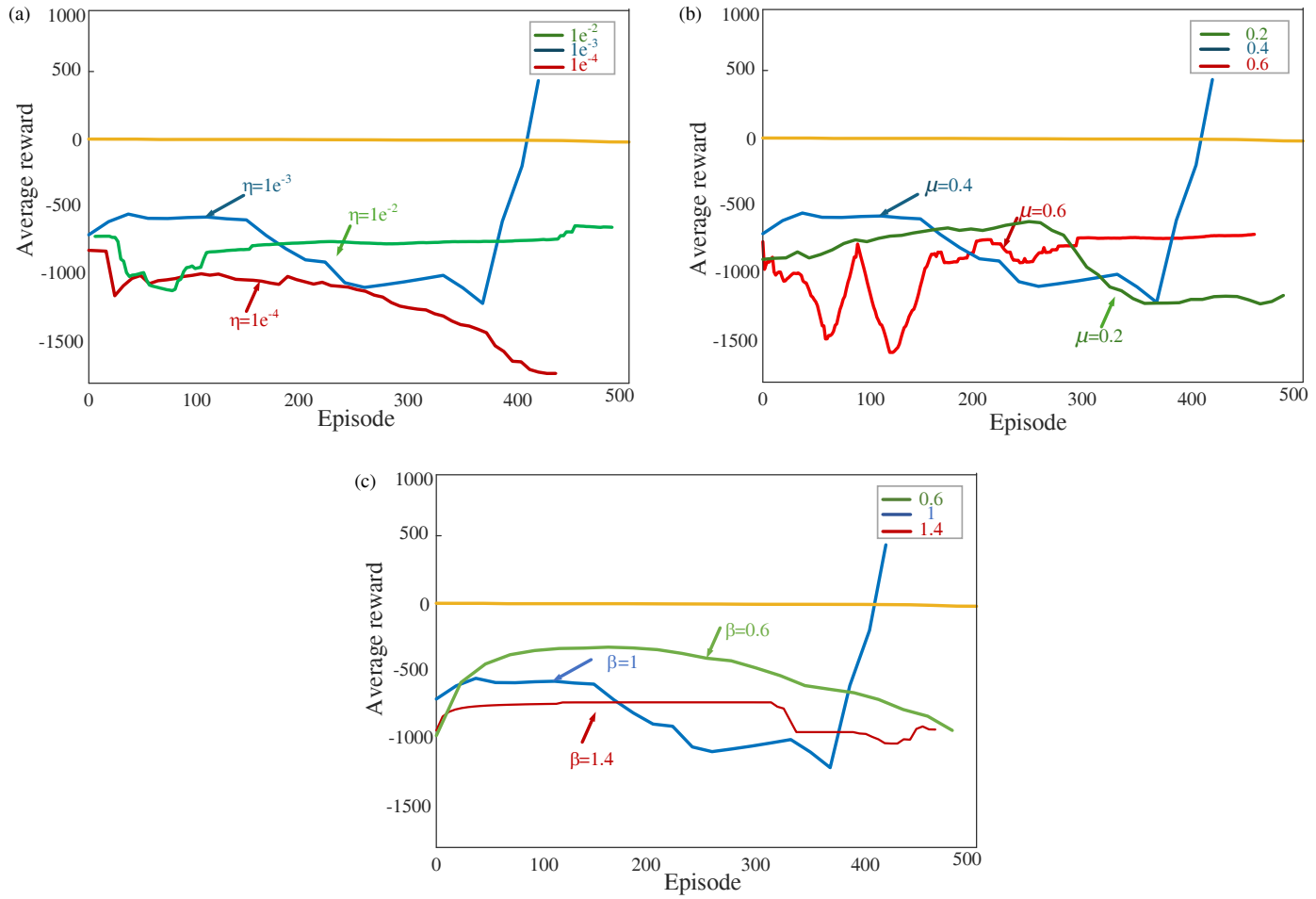

FIGURE 4. Sensitivity analysis of DDPG hyperparameters. (a) Effect of the learning rate on the average reward, (b) effect of the reward weighting coefficient on the average reward, and (c) effect of the action penalty coefficient on the average reward.

TABLE 1. DDPG algorithm hyperparameter settings.

Parameters	value
Experience pool capacity	100000
Discount factor	0.94
Action network learning rate	0.0001
Evaluating the learning rate of the network	0.001
Sampling time	0.001
Soft update coefficient	0.001
Batch size	128

experience replay pool capacity is set to 100,000 to improve sample diversity and training stability.

4.2. Sensitivity Analysis

To determine the optimal training configuration of the DDPG algorithm for the coordinated DC-link control of the neutral-point-clamped (NPC) three-level APF, a sensitivity analysis was conducted on three key hyperparameters, namely the learning rate (η), reward weighting coefficient (μ), and action penalty coefficient (β). The influence of different parameter settings on the algorithm's convergence characteristics and control performance was systematically investigated to identify the optimal parameter combination that achieves a desirable balance among training stability, convergence speed, and control performance. The corresponding results are presented in Fig. 4.

The learning rate (η) determines the parameter update step size of Actor-Critic networks and plays a crucial role in the convergence speed and training stability of the DDPG algorithm. Its influence on the average reward is illustrated in Fig. 4(a). As can be observed, significantly different convergence behaviors are obtained under different learning rates. When $\eta = 10^{-2}$, the update step is excessively large, causing the average reward to remain at a relatively low level throughout the training process, and the algorithm fails to converge to an optimal policy. In contrast, when $\eta = 10^{-4}$, the average reward improves during the early training stage but gradually deteriorates as training proceeds, eventually exhibiting noticeable performance degradation. This indicates that an excessively small learning rate reduces the efficiency of network parameter updates and increases the likelihood of convergence to a local optimum. By comparison, when the learning rates of the critic and actor networks are set to 10^{-3}, respectively, the training curve converges rapidly after approximately 350 training episodes and

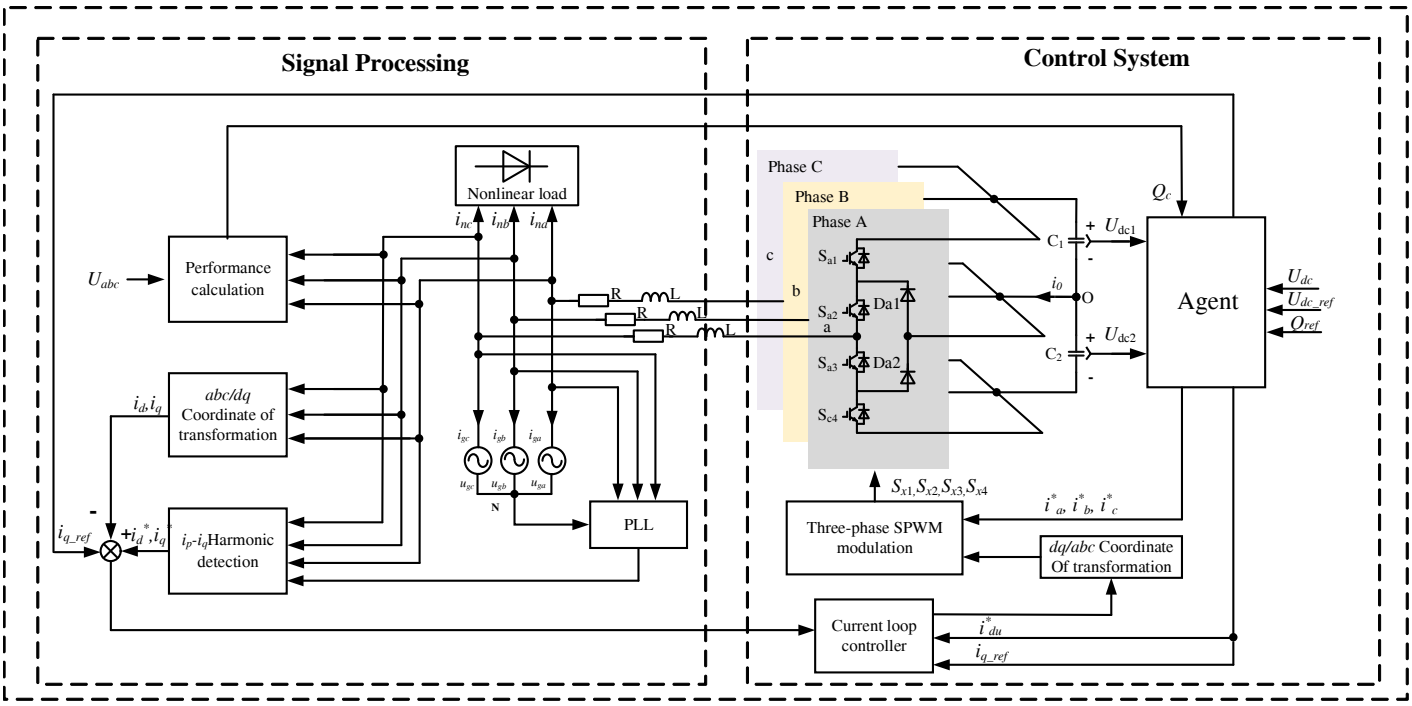

FIGURE 5. Overall control architecture of the NPC-APF system.

achieves the highest average reward among all tested configurations. This demonstrates that the selected learning rates provide an appropriate trade-off between convergence speed and training stability. Therefore, the critic and actor learning rates are finally chosen as 0.001 and 0.0001, respectively.

The reward weighting coefficient (μ) is introduced to balance contributions of the DC-link voltage tracking error and split-capacitor voltage balancing error in the reward function. Its influence on the training performance is presented in Fig. 4(b). Under identical training conditions, comparative experiments were performed with $\mu = 0.2$ and 0.6. As shown in Fig. 4(b), when $\mu = 0.2$, the average reward decreases continuously during the later training stage, indicating poor convergence performance due to insufficient emphasis on coordinated optimization of the two control objectives. When $\mu = 0.6$, relatively large fluctuations occur during the initial training stage, and although the training process gradually stabilizes, the final average reward remains lower than that obtained with the optimal parameter. In contrast, when $\mu = 0.4$, the average reward remains consistently high with relatively small fluctuations, and the algorithm converges rapidly to a stable solution. These results demonstrate that an appropriate reward weighting coefficient effectively balances DC-link voltage regulation and neutral-point voltage balancing, enabling the agent to learn a more robust control policy. Consequently, the reward weighting coefficient is set to 0.4.

The action penalty coefficient (β) is employed to constrain the variation magnitude of control actions, thereby improving the smoothness of control strategy and suppressing excessive control oscillations. The corresponding sensitivity analysis is shown in Fig. 4(c). Comparative experiments were conducted using $\beta = 0.6$, 1.0, and 1.4 while maintaining the remaining parameters unchanged. As illustrated in Fig. 4(c), when $\beta = 0.6$, a relatively high average reward is achieved during the initial training stage; however, the reward gradually declines as training progresses, indicating insufficient action constraints and poor policy stability. Conversely, when $\beta = 1.4$, the average reward remains consistently low, suggesting that an excessively large action penalty restricts the agent's exploration capability and degrades policy optimization. By comparison, when $\beta = 1.0$, the training curve exhibits the best overall stability and achieves the highest average reward in the later training stage, indicating an effective balance between control smoothness and exploration capability. Therefore, the action penalty coefficient is finally selected as 1.0.

4.3. Steady-State Harmonic Compensation and Reactive Power Compensation

To verify the steady-state harmonic compensation capability of the proposed control strategy, a comparative analysis was conducted between the conventional four-loop PI control and the single-agent DDPG cooperative control. Fig. 6 shows the grid-side current waveforms and FFT spectrum results for the two control strategies.

Following the connection of a nonlinear load, significant distortion appears in the grid current. Although conventional PI control can achieve basic harmonic compensation, due to dynamic coupling between multiple control loops, the reference current on the dq-axis continues to fluctuate during the steady-state phase, thereby leading to a decline in the tracking accuracy of the compensation current. As shown in Fig. 6(c), the total harmonic distortion (THD) of the grid-side current under

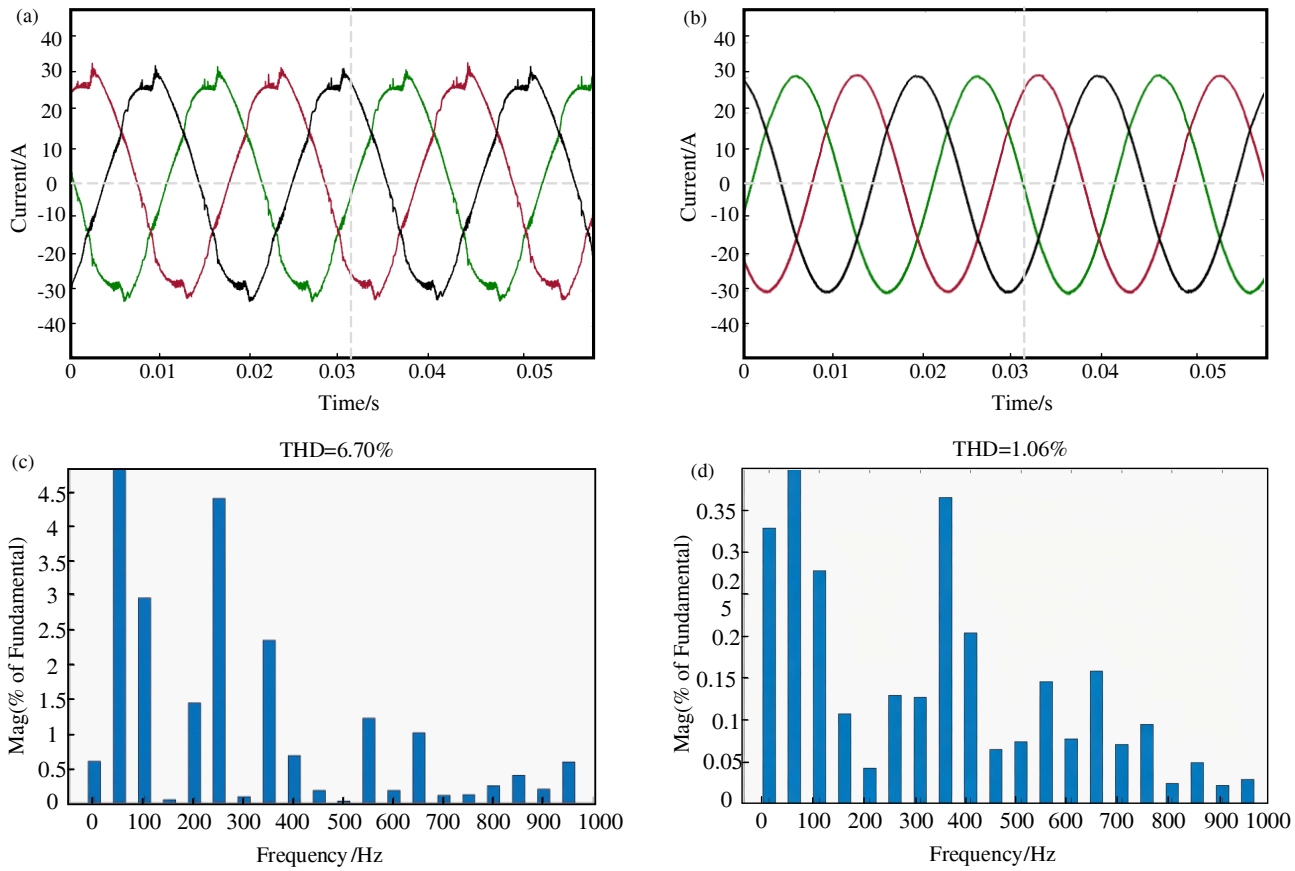

FIGURE 6. Grid-side current waveform and total harmonic distortion. (a) Grid-side current under PI control, (b) grid-side current under DDPG control, (c) current harmonics under PI control, and (d) current harmonics under DDPG control.

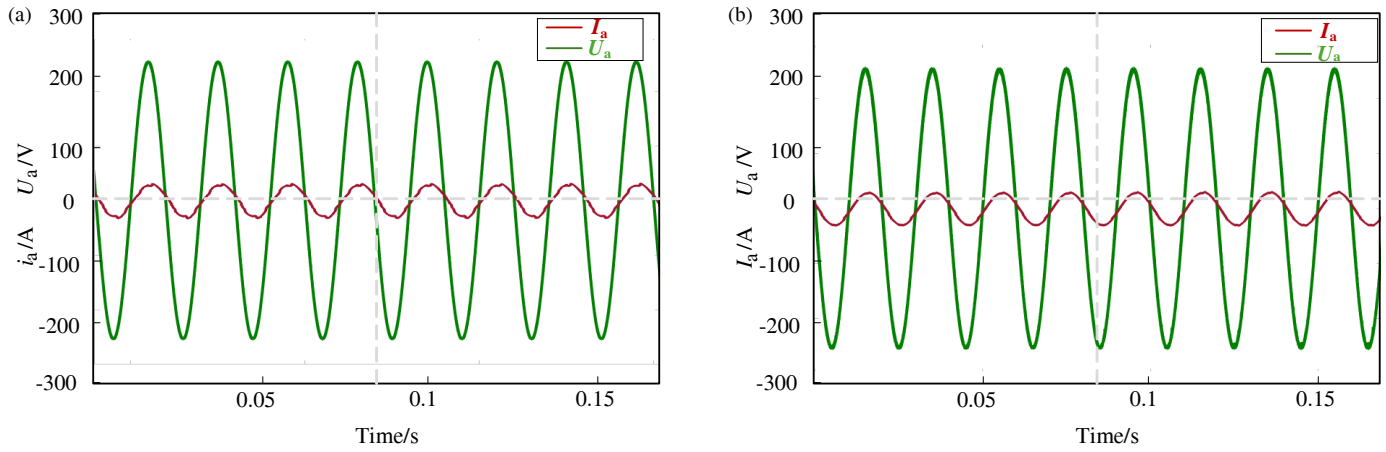

FIGURE 7. Dynamic response of phase A current and phase A voltage on the grid side. (a) Grid-side current and voltage under PI control and (b) grid-side current and voltage under DDPG control.

conventional PI control reached 6.7%, exceeding the harmonic current limits specified in IEEE 519-2014 for public coupling points.

In contrast, the proposed DDPG cooperative control achieves multi-objective joint optimization using a unified reward function, thereby reducing the impact of outer-loop dynamic fluctuations on the current tracking process. Because the agent can autonomously learn the dynamic coupling relationships between different control objectives, fluctuations in the *dq* current components are significantly reduced, resulting in a smoother compensation current waveform. As shown in Fig. 6(d), the THD of the grid-side current under DDPG control was reduced to 1.06%, representing a decrease of approximately 83.7% compared to the conventional PI control.

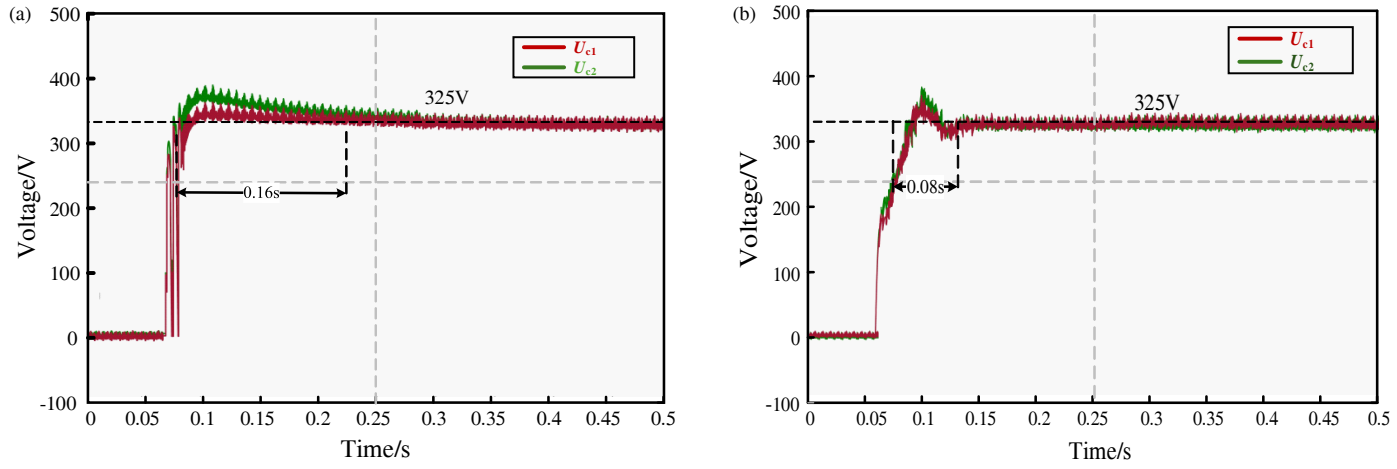

FIGURE 8. Dynamic response of the midpoint capacitance voltage under different control strategies. (a) Midpoint capacitance voltage under PI control and (b) midpoint capacitance voltage under DDPG control.

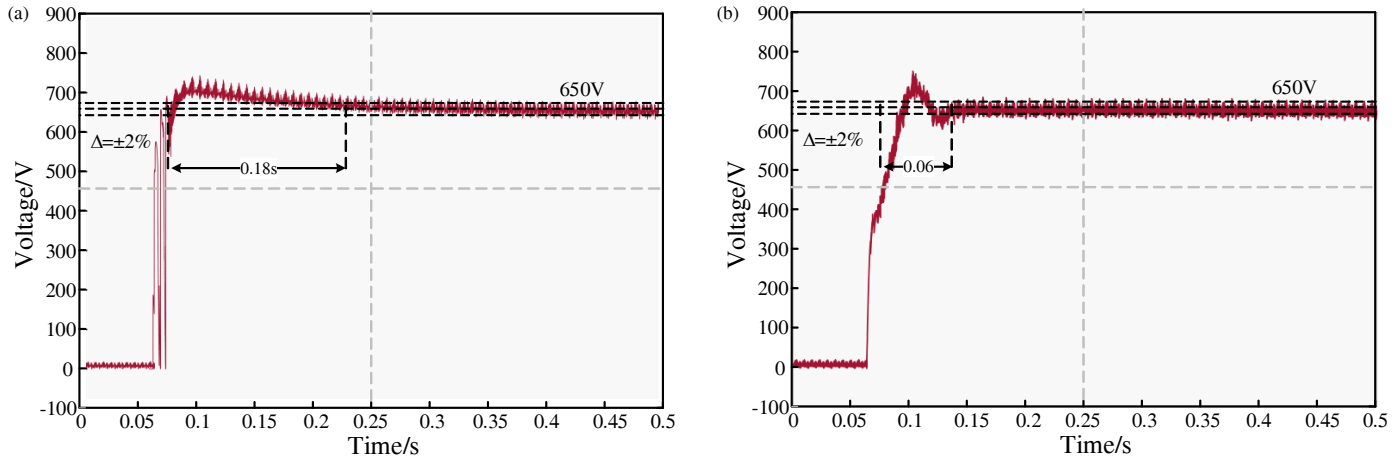

FIGURE 9. Dynamic response of the DC bus voltage under different control strategies. (a) DC bus voltage under PI control and (b) DC bus voltage under DDPG control.

Furthermore, as shown in Fig. 7, the grid current and voltage remained essentially in phase under DDPG control, and the system power factor was significantly improved. This indicates that the proposed method not only effectively suppresses harmonics but also achieves good reactive power compensation performance.

4.4. Comparison of Capacitor Voltage Balancing Performance

Midpoint potential drift is an inherent flaw in the three-phase, three-level, NPC-APF topology. When the voltages across the upper and lower capacitors are unbalanced, this can lead to the injection of even-order harmonics into the output voltage and overvoltage breakdown of the switching devices.

To verify the capacitor voltage balancing capability of the proposed strategy, a series of comparative tests were conducted, and the results are shown in Fig. 8.

Under PI control, the time required to balance the capacitor voltages was 0.16 s. During the regulation process, the waveform exhibited slight oscillations, and a minor steady-state error was present. In contrast, under DDPG cooperative control, the time required to balance the capacitor voltages was only 0.08 s — a 50% reduction compared to PI control. There was no steady-state error or significant waveform jitter, demonstrating that the capacitor voltage balancing performance was significantly superior to that of conventional PI control.

4.5. Comparison of DC Bus Voltage Regulation Performance

The stability of the DC bus voltage determines the APF's energy exchange capability and compensation performance. When load changes or reactive power fluctuations occur in the system, the energy on the DC side changes rapidly; therefore, the controller must possess strong dynamic regulation capability.

Figure 9 shows the dynamic response of the DC bus voltage for the two control strategies. Conventional PI control, which employs fixed parameters, is susceptible to coupling effects between the midpoint balancing loop and reactive power control loop under system dynamic conditions; consequently, the voltage recovery process exhibits an overshoot. As shown

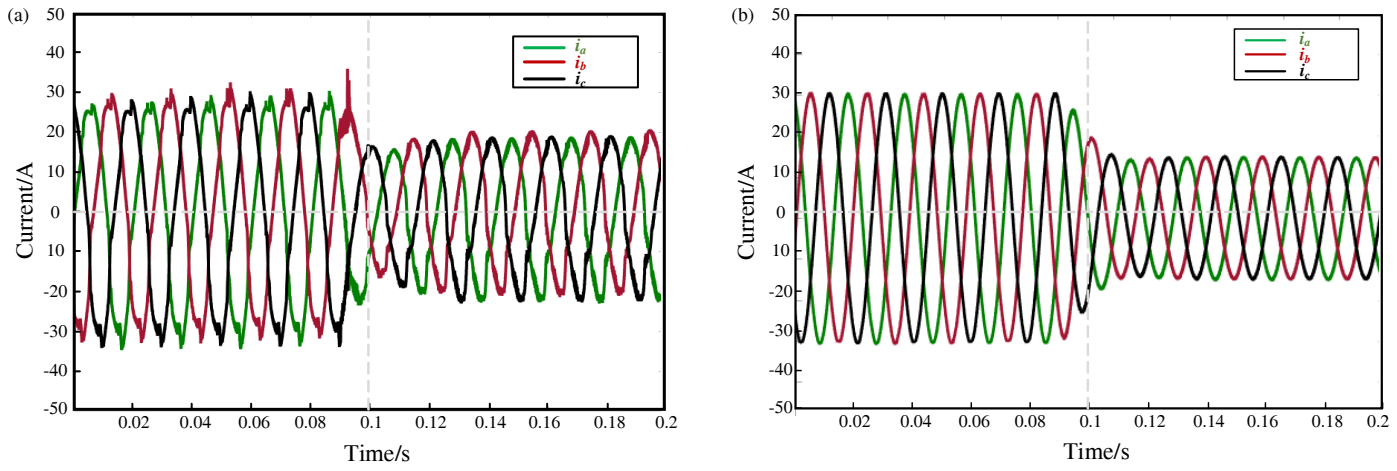

FIGURE 10. Current dynamic response under grid voltage sag conditions, (a) grid-side current under PI control and (b) grid-side current under DDPG control.

in Fig. 9(a), the system under PI control took approximately 0.18 s to reach a steady state.

In contrast, DDPG cooperative control can adjust control actions in real time according to the system state, thereby achieving dynamic optimization of energy distribution on the DC side. Fig. 9(b) demonstrates that the proposed method can complete bus voltage recovery within approximately 0.06 s, reducing the settling time by approximately 66.67% compared with traditional PI control, with no significant overshoot during the dynamic process.

Furthermore, under DDPG control, the bus voltage fluctuation range during the steady-state phase was less than ± 1 V, indicating that the reinforcement learning method can effectively enhance the system's steady-state stability and dynamic response capability.

4.6. Analysis of Grid Voltage Sags

In the actual operation of distribution networks, voltage sags are typical dynamic disturbances. When the grid voltage changes abruptly, the APF must rapidly adjust compensation current to maintain stable system operation; therefore, the controller's dynamic robustness is particularly important.

In this study, the three-phase grid voltage was suddenly reduced to 50% of its rated value, and the dynamic response of the grid-side current under the two control strategies was compared, as shown in Fig. 10.

Owing to its fixed parameters, conventional PI control struggles to adjust the control gain promptly when the system operating point changes abruptly. Consequently, oscillations occurred during the current recovery process, and the system took a relatively long time to stabilize.

In contrast, the DDPG agent can dynamically adjust control actions based on real-time states, thereby maintaining a relatively stable compensation current output even after a voltage dip. As shown in Fig. 10(b), the proposed method restores stable operation within approximately three power-frequency cycles, with no significant overshoot in the current waveform.

These results indicate that the reinforcement learning method has a strong adaptability to changes in system parameters and operational disturbances.

4.7. Analysis of Sudden Load Change Conditions

To further simulate sudden load surge disturbances, the nonlinear load power was increased to 1.5 times its original amplitude during the experiment.

Sudden load changes cause rapid fluctuations in the system's harmonic currents and reactive power demand, thereby placing higher demands on the APF control system in terms of dynamic response. Conventional PI control relies on fixed parameter adjustments, which limits the speed at which the reference current is updated. Consequently, a compensation lag is likely to occur at the instant of a load change.

In contrast, DDPG cooperative control can update control actions in real time according to changes in the system state, thereby enabling faster reconstruction of the compensation current. Fig. 11 demonstrates that, under sudden load-surge conditions, the proposed method can still maintain relatively stable

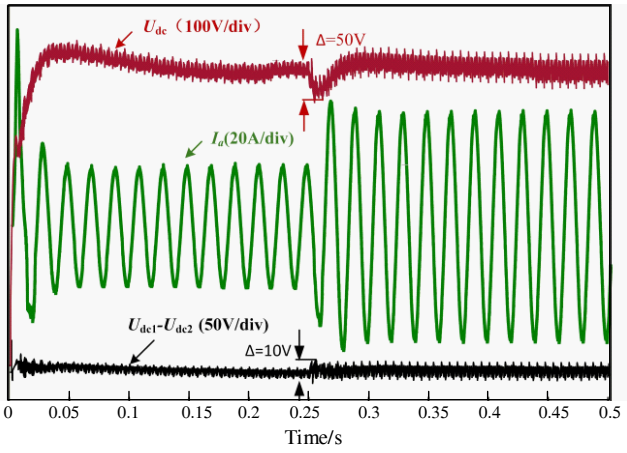

FIGURE 11. Grid-side current response under sudden load changes.

three-phase current waveforms without significant dynamic oscillations.

Taking the results of the disturbance experiments as a whole, it can be seen that reinforcement learning methods can effectively enhance the robustness and adaptability of the NPC-APF under complex dynamic operating conditions.

5. CONCLUSION

To address issues associated with traditional multi-loop PI control of NPC-APF systems, such as complex parameter tuning, severe dynamic coupling among multiple control loops, and insufficient robustness under complex operating conditions, this study proposes a multi-objective cooperative control method based on a single DDPG agent. This method employs a reinforcement learning agent to uniformly coordinate the regulation of the DC bus voltage, midpoint potential balancing, and reactive power compensation. By retaining the inner-loop PI control of the current, the joint optimization of multiple control objectives for the APF is achieved.

The simulation results demonstrate that the proposed method significantly enhances the NPC-APF's overall control performance. Under steady-state conditions, the total harmonic distortion (THD) of the grid-side current was reduced from 6.7% under conventional PI control to 1.06%, thereby meeting the harmonic limitation requirements of IEEE 519-2014 for distribution networks. Concurrent with this, the voltage-balancing time at the neutral-point capacitor and the DC bus voltage regulation time were reduced by 50% and 66.67%, respectively, with no significant overshoot observed during dynamic operation. Further analysis indicates that DDPG cooperative control can autonomously learn the dynamic coordination relationships between multiple control objectives through a unified reward mechanism, thereby mitigating the coupling effects inherent in traditional multi-loop control. Compared with PI control, which relies on fixed parameters, reinforcement learning methods offer stronger online adaptive capabilities; consequently, they can maintain good dynamic stability and robustness even under complex disturbances, such as grid voltage sags and sudden load changes. Furthermore, the proposed method does not require the establishment of high-precision linearized mathematical models, nor necessitates complex parameter tuning for multiple PI controllers, thereby reducing the engineering commissioning difficulty of the APF control system. This method provides a new implementation pathway for applying deep reinforcement learning to power quality management, offering significant engineering reference value for controlling smart distribution systems and high-performance power electronic equipment.

Although this study has validated the effectiveness of the single-agent DDPG in NPC-APF's multi-objective control, certain limitations remain. Owing to the large scale of deep reinforcement learning network parameters, the real-time deployment capability of the proposed method on low-computing-power embedded hardware platforms requires further investigation.

REFERENCES

[1] Tian, S., Z. Ning, P. Li, Y. Guo, Z. Li, and H. Sun, "Siting and capacity allocation for APFs based on optimal reactive power compensation and harmonic mitigation capacities of STATCOMs and MFGCIs," *International Journal of Electrical Power & Energy Systems*, Vol. 173, 111357, Dec. 2025.

[2] Etanya, T. F., P. Tsafack, and D. K. Ngwashi, "Grid-connected distributed renewable energy generation systems: Power quality issues, and mitigation techniques — A review," *Energy Reports*, Vol. 13, 3181–3203, Jun. 2025.

[3] Elkholy, M. M., M. A. El-Hameed, and A. A. El-Fergany, "Harmonic analysis of hybrid renewable microgrids comprising optimal design of passive filters and uncertainties," *Electric Power Systems Research*, Vol. 163, 491–501, Oct. 2018.

[4] Wu, Z., G. Xu, W. Zhu, and G. Sheng, "A dispersing proportional resonance controller for selective harmonic current compensation in a shunt active power filter," *IEEE Transactions on Power Electronics*, Vol. 40, No. 1, 279–289, Jan. 2025.

[5] Smaida, E., J. Saulo, K. Abdellah, M. Ahuna, and A. Teta, "A novel multifunctional shunt active power filter based AI-driven for grid-tied PV system using Z-source inverter," in *2025 4th International Conference on Computing and Information Technology (ICCIT)*, 656–661, Tabuk, Saudi Arabia, Apr. 2025.

[6] Liu, J., Y. Ni, and J. Zhao, "An improved control method of dc voltage for series hybrid active power filter," *Energies*, Vol. 17, No. 14, 3390, Jul. 2024.

[7] Santiprapan, P., K. Areerak, and K. Areerak, "An adaptive gain of proportional-resonant controller for an active power filter," *IEEE Transactions on Power Electronics*, Vol. 39, No. 1, 1433–1446, Jan. 2024.

[8] Gök, R., A. M. Çolak, E. Irmak, and N. Öztürk, "Comparative analysis of fuzzy logic and pi control for harmonic suppression in shunt active power filter," in *2024 13th International Conference on Renewable Energy Research and Applications (ICRERA)*, 1651–1656, Nagasaki, Japan, Nov. 2024.

[9] Zakaria, L. and B. Amar, "Neural network controller for three phase shunt active power filter using self tuning filter," in *2022 19th International Multi-Conference on Systems, Signals & Devices (SSD)*, 333–338, Sétif, Algeria, May 2022.

[10] Daou, N., M. S. Chabani, B. Achour, S. E. Zouzou, V. Puig, M. A. Mossa, and M. M. Almalki, "Robust second-order sliding-mode backstepping control of the parallel active filter for harmonics and reactive power compensation-experimental validation," *Energy Reports*, Vol. 15, 109117, Jun. 2026.

[11] Balouji, E., K. Bäckström, and T. McKelvey, "Deep reinforcement learning based grid-forming inverter," in *2023 IEEE Industry Applications Society Annual Meeting (IAS)*, 1–9, Nashville, TN, USA, Oct. 2023.

[12] Li, B., Q. Wu, Y. Cao, W. Jiao, and C. Li, "Physically informed multi-agent deep reinforcement learning for distributed voltage control in distribution networks," *International Journal of Electrical Power & Energy Systems*, Vol. 174, 111451, Jan. 2026.

[13] Wang, H., J. Li, S. Li, X. Liu, J. Bian, and H. Yu, "Optimal inverter-based droop control in active distribution network with evolutionary strategy-embedded deep reinforcement learning," *Electric Power Systems Research*, Vol. 247, 111789, Oct. 2025.

[14] Dash, D. K., P. K. Sadhu, and A. K. Shrivastav, "Dynamic optimization of active power filters in grid-tied photovoltaic systems using partial reinforcement learning," *Neural Computing and Applications*, Vol. 37, No. 20, 15 605–15 634, May 2025.

[15] Li, P., J. Shen, Z. Wu, M. Yin, Y. Dong, and J. Han, "Optimal real-time Voltage/Var control for distribution network: Droop-

control based multi-agent deep reinforcement learning," *International Journal of Electrical Power & Energy Systems*, Vol. 153, 109370, Nov. 2023.

[16] Wang, J., R. Yang, and Z. Yao, "Efficiency optimization design of three-level active neutral point clamped inverter based on deep reinforcement learning," in *2022 IEEE 6th Conference on Energy Internet and Energy System Integration (EI2)*, 605–610, Chengdu, China, Nov. 2022.

[17] Qashqai, P., M. Babaie, R. Zgheib, and K. Al-Haddad, "A model-free switching and control method for three-level neutral point clamped converter using deep reinforcement learning," *IEEE Access*, Vol. 11, 105 394–105 409, Sep. 2023.

[18] Liang, Z., J. Wang, Y. Wu, and Z. Zhang, "Weight-adaptable disturbance observer for continuous-control-set model predictive control of NPC-3L-Fed PMSMs," *Energies*, Vol. 18, No. 21, 5864, Nov. 2025.

[19] Xu, D. and L. Zhang, "Research on PVPC system based on an improved ip-iq algorithm," in *2021 IEEE Sustainable Power and Energy Conference (iSPEC)*, 1395–1399, Nanjing, China, Dec. 2021.

[20] Hrgović, I. and I. Pavić, "Reward design for intelligent deep reinforcement learning based power flow control using topology optimization," *Sustainable Energy, Grids and Networks*, Vol. 41, 101580, Mar. 2025.