Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.085661

ARTICLE

A Feature-Adaptive Knowledge Distillation Framework for Efficient Offline-to-Online Reinforcement Learning

Baoping Tian, Zhuxiao Wang*, Jiahao Xue, Hong Wang, Ying Zhang and Yun Ju

Department of Computer Science and Technology, School of Control and Computer Engineering, North China Electric Power University, Beijing, China

*Corresponding Author: Zhuxiao Wang. Email: wangzx@ncepu.edu.cn

Received: 15 May 2026; Accepted: 05 August 2026; Published: 15 September 2026

ABSTRACT: Deep reinforcement learning (DRL) has gained significant attention as an essential technology for constructing intelligent agents capable of handling high-dimensional visual observations in complex control environments. With the rapid development of knowledge transfer paradigms, reincarnating reinforcement learning (RRL) has emerged as a promising approach to accelerate policy convergence and alleviate the inefficiency of traditional tabula rasa training by reusing pre-trained teacher policies. However, existing RRL approaches primarily focus on improving knowledge transfer efficiency, while how student networks adaptively regulate and selectively utilize inherited representations during the teacher–student transition remains underexplored. As a result, student agents may indiscriminately inherit environmental background noise and suboptimal representations from the teacher, leading to suboptimal policy convergence and severe performance plateaus. In this study, we introduce a novel feature-adaptive knowledge distillation framework, named FA-QDagger, to address the critical dilemma of selective inheritance. Distinct from existing attention-augmented value networks, our approach embeds the Squeeze-and-Excitation (SE) channel recalibration mechanism directly into the teacher-student distillation loop as a dynamic information bottleneck. Furthermore, through targeted recalibration operations, it adaptively amplifies task-critical decision signals while robustly filtering out the teacher's suboptimal prior knowledge based on varying environmental states. The proposed approach is evaluated on high-dimensional visual control benchmarks across Atari 2600 environments under restricted offline and online interaction budgets. Experimental results on five representative Atari 2600 environments demonstrate that FA-QDagger consistently improves asymptotic performance over representative RRL baselines during online fine-tuning while maintaining the jump-start advantage of teacher-guided distillation. In addition, visual interpretability analysis using saliency maps confirms that the attention-enhanced agent develops more concentrated visual focus on decision-critical regions, indicating improved feature decoupling and robustness against environmental interference.

KEYWORDS: Deep reinforcement learning; reincarnating reinforcement learning; knowledge distillation; offline-to-online learning; channel attention; intelligent agent

1 Introduction

Autonomous agents have become an important component of modern intelligent computing systems, where they are expected to perceive complex environments, process high-dimensional observations, and make sequential decisions with limited human intervention. Deep reinforcement learning (DRL) provides a powerful framework for constructing such agents because it can learn control policies directly from raw sensory inputs and interactively optimize decision-making behaviors through environmental feedback [1–3]. In recent years, reinforcement-learning-based agents have also been explored in context-aware intelligent

environments and multi-agent decision-support systems, further demonstrating the potential of DRL for autonomous reasoning, adaptive behavior modeling, and intelligent system deployment [4]. However, despite these advances, conventional DRL usually follows a tabula rasa training paradigm, in which an agent learns from scratch through massive trial-and-error interactions. This process requires a large number of interaction frames and extensive computational resources, which significantly limits the rapid development and practical deployment of intelligent agents in resource-constrained scenarios.

To reduce the computational burden of training from scratch, reincarnating reinforcement learning (RRL) has emerged as a promising knowledge reuse paradigm [5]. Instead of discarding previously trained agents, RRL aims to reuse existing computational assets, such as teacher policies, saved model parameters, or offline interaction logs, to accelerate the training of a new student agent. Under this framework, representative methods such as QDagger perform offline knowledge distillation followed by online fine-tuning to accelerate policy improvement. Notably, QDagger is a reincarnation learning framework rather than a fixed network architecture and can be instantiated with different value-based backbones. The original implementation adopts a Rainbow agent with an Impala-CNN backbone, whereas the present work employs the classical Nature DQN backbone to investigate the effect of the proposed SE-based feature recalibration module. In this way, the student agent can obtain an initial performance jump-start from the teacher while retaining the opportunity to further improve its policy through subsequent interaction with the environment.

Although Q-Dagger and related RRL methods can effectively alleviate the inefficiency of early-stage exploration, most existing studies mainly focus on distillation objectives, replay strategies, or teacher-student policy matching. A critical but insufficiently explored issue lies in the architecture of the student network itself. Current RRL implementations commonly adopt standard convolutional backbones, such as the Nature Deep Q-Network (DQN), as the feature extractor of the student agent [6]. During visual feature extraction, these networks process all feature channels in a relatively uniform manner, without explicitly distinguishing task-critical visual cues from irrelevant background information. In complex visual control environments, the knowledge provided by the teacher policy is not always optimal. It may contain overfitted background patterns, redundant visual distractors, and rigid policy preferences inherited from the teacher's own training process.

Fig. 1 presents the five Atari 2600 environments used in this study, including Breakout, Bowling, Seaquest, Asterix, and Space Invaders. These environments contain high-frequency visual changes, moving targets, environmental interference, and delayed decision consequences, making them suitable benchmarks for evaluating whether a student agent can selectively inherit useful policy knowledge while avoiding redundant or misleading visual representations.

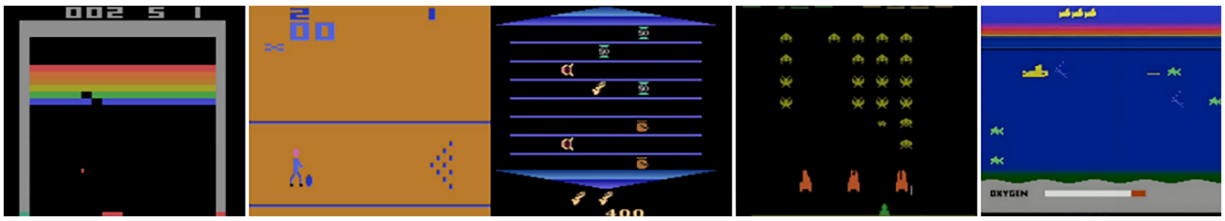

Figure 1: Screenshots of the five Atari 2600 games evaluated in this study: Breakout, Bowling, Seaquest, Asterix, and Space Invaders.

Lacking an intrinsic feature-filtering mechanism, the student agent may passively inherit suboptimal feature representations during offline distillation. This indiscriminate knowledge absorption can cause the student network to overfit redundant environmental features rather than focusing on decision-critical

visual patterns. In value-based reinforcement learning, such redundant feature fitting may further aggravate Q-value overestimation, resulting in unstable online fine-tuning, reduced policy plasticity, and performance plateaus. Consequently, although the student agent may achieve a strong initial jump-start, it may fail to maintain robust long-term improvement or surpass the performance ceiling of the teacher policy during the online adaptation stage.

To address this problem, this paper proposes a feature-adaptive knowledge distillation framework, named FA-QDagger, for efficient offline-to-online reinforcement learning. The core idea is to enhance the student network with an adaptive feature recalibration mechanism, enabling it to selectively absorb useful teacher knowledge while suppressing redundant background noise. Specifically, we integrate the lightweight Squeeze-and-Excitation (SE) channel attention mechanism into the standard DQN backbone and construct a feature-aware network named SE-NatureDQN [7]. By embedding SE modules into multiple convolutional stages, the student agent can dynamically model channel-wise feature dependencies and recalibrate visual representations according to varying environmental states.

In the proposed framework, the SE module functions as both a dynamic information bottleneck and a feature-level regularizer. During offline distillation, it helps the student agent emphasize task-relevant channels and avoid indiscriminate imitation of noisy teacher representations. During online fine-tuning, it further preserves policy plasticity by suppressing redundant features and promoting more conservative value estimation. As a result, the proposed method aims not only to maintain the initial jump-start advantage of teacher-guided distillation but also to improve asymptotic performance and training robustness during the subsequent online interaction phase.

The main contributions of this paper are summarized as follows:

1. Feature-Adaptive Distillation Paradigm: We propose FA-QDagger, a novel reincarnating reinforcement learning (RRL) framework. Instead of treating attention merely as a feature extractor for standard RL, we uniquely introduce channel recalibration into the teacher-student distillation loop. This paradigm addresses the critical problem of "selective inheritance," empowering the student agent to adaptively filter suboptimal prior knowledge and background noise provided by the teacher during the offline-to-online transition.
2. Dynamic Information Bottleneck via SE-NatureDQN: We design a feature-aware value network, SE-NatureDQN, explicitly tailored for knowledge distillation. By embedding the Squeeze-and-Excitation (SE) mechanism as a dynamic information bottleneck within the distillation pipeline, the architecture robustly suppresses Q-value overestimation bias and maintains representational plasticity, breaking the performance ceiling of the suboptimal teacher.
3. We qualitatively discuss the role of the channel attention mechanism from the perspective of feature-level regularization. Rather than a formal theoretical proof, we conceptually explore how embedding this mechanism within the distillation loop has the potential to mitigate Q-value overestimation by acting as a dynamic information bottleneck.
4. Extensive experiments are conducted on five representative Atari 2600 visual control benchmarks under restricted offline and online interaction budgets. The results demonstrate that FA-QDagger consistently improves asymptotic performance while preserving the jump-start advantage of teacher-guided distillation under the evaluated environments. Visual interpretability analysis further shows that the proposed agent develops more concentrated attention on decision-critical regions, indicating improved feature decoupling and resistance to environmental interference.

The remainder of this paper is organized as follows. Section 2 reviews related work on DRL, knowledge transfer, RRL, and attention mechanisms. Section 3 describes the proposed FA-QDagger framework and

SE-NatureDQN architecture. Section 4 reports experimental settings, quantitative results, stability analysis, visual interpretation, and ablation studies. Section 5 concludes the paper and discusses future work.

2 Related Work

This section reviews the research foundations most relevant to the proposed method, including deep reinforcement learning for visual control, knowledge reuse for offline-to-online learning, and attention mechanisms for feature-adaptive intelligent agents. The review highlights that existing RRL approaches mainly emphasize how teacher knowledge is transferred, whereas the representational capability of the student network has received less attention.

2.1 Deep Reinforcement Learning for Visual Control

DRL has significantly advanced autonomous decision-making by combining reinforcement learning with deep neural networks. The Arcade Learning Environment (ALE) provides a standard benchmark for evaluating general agents in high-dimensional visual control tasks [8]. The Deep Q-Network (DQN) demonstrated that convolutional neural networks can learn effective action-value functions directly from raw Atari pixels and achieve strong control performance [6]. Subsequent variants further improved stability and efficiency. Double DQN was proposed to reduce overestimation bias in target value evaluation [9], prioritized experience replay increased sample efficiency by emphasizing transitions with large temporal-difference errors [10], and Rainbow integrated several complementary improvements into a strong value-based agent [11].

Despite these developments, standard DRL agents still require extensive environment interaction and long training cycles. This limitation is particularly problematic when training costs are high, interaction data are limited, or the agent must be deployed in practical intelligent systems. Therefore, reusing prior computational results has become an important research direction for accelerating policy development and reducing training burdens.

2.2 Knowledge Transfer and Offline-to-Online Reinforcement Learning

Knowledge transfer has been widely studied as a way to improve the efficiency of learning agents. Policy distillation transfers action preferences from a teacher policy to a student network, enabling policy compression and reuse [12]. Dataset Aggregation (DAgger) addresses covariate shift in imitation learning by iteratively collecting data from the student distribution while querying expert labels [13]. Offline reinforcement learning further demonstrates the potential of training agents from previously collected datasets without direct interaction with the environment [14,15].

RRL extends these ideas by reusing pre-trained agents, teacher policies, or offline buffers to initialize a new agent and accelerate subsequent learning [5]. In value-based settings, Q-Dagger uses soft Q-value guidance from a teacher during offline distillation and then allows the student to improve through online interaction. Recent studies have also examined offline-to-online knowledge transfer and balanced replay strategies to mitigate the distributional mismatch between static datasets and online exploration [16–18].

However, most existing methods primarily focus on loss functions, replay schemes, or policy-level imitation. They generally assume that a standard student network can absorb teacher knowledge effectively. In complex visual environments, this assumption may be insufficient because teacher policies can encode redundant visual correlations and suboptimal feature patterns. Therefore, improving the feature extraction and filtering ability of the student agent is necessary for robust offline-to-online transfer.

2.3 Attention Mechanisms for Feature-Adaptive Agents

Attention mechanisms allow neural networks to adaptively emphasize informative components of input features [19]. The SE module is a lightweight channel attention mechanism that explicitly models channel-wise dependencies and recalibrates feature responses with limited computational overhead. By strengthening useful channels and suppressing less informative ones, SE blocks have been successfully used to improve representation learning in computer vision tasks.

While combining attention mechanisms with neural networks and DQN-style value networks has been widely explored to improve resource allocation and interpretability [20–22], these applications fundamentally differ from our focus. For instance, early pioneering work such as the Deep Attention Recurrent Q-Network (DARQN) successfully incorporated soft and hard attention into DQN to highlight task-relevant regions in Atari games [23]. Following this trajectory, recent studies have developed various attention-augmented value networks to accelerate feature convergence in traditional tabula rasa (scratch-trained) reinforcement learning. For example, SENet-style attention has been directly embedded into Double DQN (AMUW-DDQN) for traffic signal control [24], and soft, top-down spatial attention mechanisms have been utilized to create information bottlenecks for interpretable RL agents in visual environments [25].

However, these existing studies primarily utilize attention mechanisms to process environmental observations during training from scratch. Our contribution lies elsewhere: we investigate the role of attention within the cross-model knowledge distillation loop of Reincarnating RL. Specifically, we utilize SE-based channel recalibration to solve the "selective inheritance" dilemma. Rather than simply extracting spatial features from the environment, our FA-QDagger employs the SE block as a dynamic information bottleneck to actively filter out suboptimal representations and overestimation biases inherited from the pre-trained teacher. Therefore, our work delineates a novel perspective: leveraging feature-level recalibration to safeguard the offline-to-online knowledge transfer process, rather than merely introducing attention into DQN itself.

Beyond reinforcement learning algorithms, the concept of knowledge-guided optimization has recently been extended to software engineering. For example, recent studies have explored reinforcement learning for knowledge-enhanced software refinement and performance-driven software development, demonstrating that prior knowledge can effectively guide sequential optimization processes in software engineering tasks [26,27]. Although these studies target software engineering rather than reinforcement learning policy learning, they share the common objective of leveraging prior knowledge to improve decision quality and optimization efficiency. Compared with these application-oriented approaches, FA-QDagger focuses on feature-adaptive knowledge transfer within the offline-to-online reincarnating reinforcement learning framework by selectively filtering inherited teacher representations during policy distillation.

3 Materials and Methods

This section presents the proposed FA-QDagger framework. The method is designed to solve the feature inheritance problem in RRL by combining teacher-guided knowledge distillation with a channel-attention-enhanced student backbone. The overall process includes problem formulation, two-stage learning, SE-NatureDQN modeling, knowledge distillation loss design, and theoretical interpretation of the SE-based feature filtering mechanism.

3.1 Problem Formulation and Two-Stage Learning Pipeline

Following the general reinforcement-learning-based autonomous agent modeling paradigm in intelligent environments, the interaction between the student agent and the visual control environment is

formulated as a Markov decision process (MDP), denoted by M = (S, A, P, R, γ). Here, S is the state space, A is the action space, P is the transition probability, R is the reward function, and γ is the discount factor. At each control step t, the agent observes a state s_t constructed from stacked visual frames, selects an action at according to its policy, receives a reward r_t, and transitions to the next state s_{t+1}.

The proposed framework follows a two-stage offline-to-online learning pipeline, as shown in Fig. 2. In the offline stage, the student agent learns from static interaction data generated by a pre-trained teacher policy. This stage aims to provide an initial policy jump-start without requiring additional environmental interaction. In the online stage, the student agent interacts with the environment and refines its policy using hybrid learning signals. The main challenge is that the student must retain useful teacher knowledge while correcting inherited biases and adapting to new state distributions encountered during online exploration.

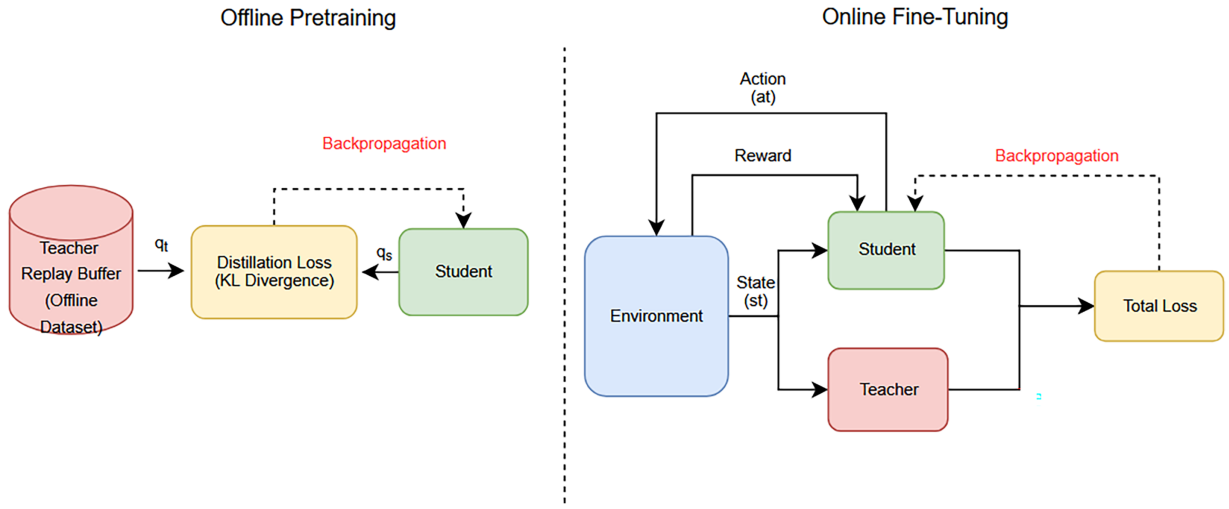

Figure 2: Two-stage knowledge distillation framework of FA-QDagger, including offline teacher-guided policy jump-start and online feature-adaptive policy refinement.

Unlike the original implementation of QDagger, which adopts a Rainbow agent with an Impala-CNN backbone for large-scale evaluation, this work intentionally employs the classical Nature DQN backbone as the student network. This simplified implementation provides a common backbone for both the baseline and the proposed method, allowing the contribution of the SE-based feature recalibration module to be evaluated independently from improvements introduced by more advanced value-learning algorithms.

3.2 SE-NatureDQN Backbone

A key limitation of standard RRL implementations is that the student network usually adopts a conventional convolutional architecture without explicit feature filtering. To address this limitation, this study constructs SE-NatureDQN by embedding SE blocks into the Nature DQN backbone. The network receives an $84 \times 84 \times 4$ stacked visual input and processes it through three convolutional layers followed by a fully connected layer and an action-value output layer. After each convolutional layer, an SE block is inserted to perform channel-wise recalibration.

The architecture is illustrated in Fig. 3. The convolutional layers extract spatial representations from visual observations, while the SE modules dynamically evaluate the importance of feature channels. This design allows the student agent to amplify channels related to task-critical objects, such as balls, enemies,

targets, and boundaries, while suppressing redundant background channels that may be inherited from teacher demonstrations.

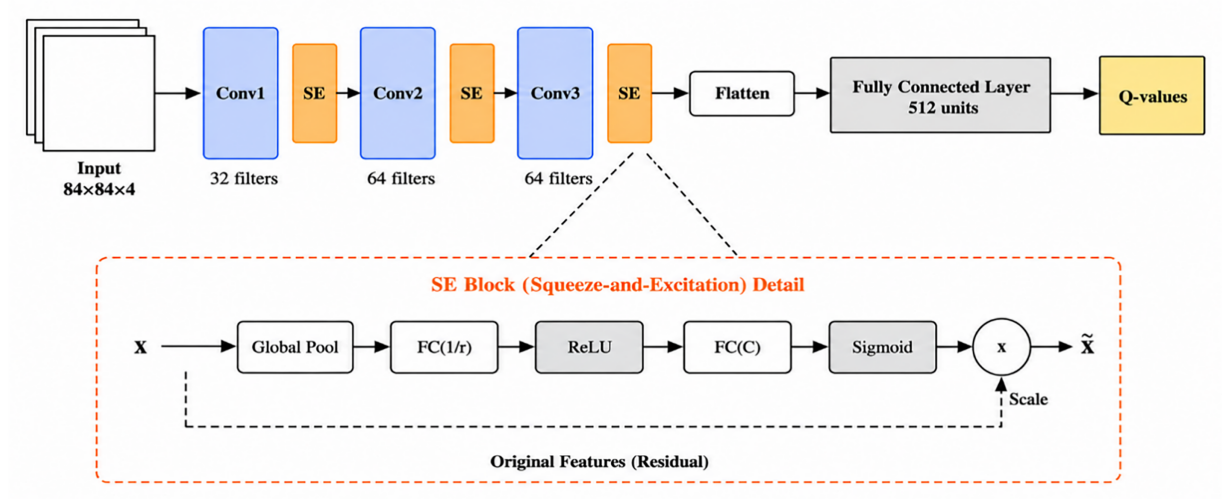

Figure 3: Architecture of SE-NatureDQN and its integration into the two-stage FA-QDagger framework. Cascaded SE blocks recalibrate convolutional features before action-value estimation.

It is worth noting that the proposed feature recalibration mechanism introduces only marginal computational overhead. In our implementation, three SE blocks are inserted into the Nature DQN backbone with a reduction ratio of r = 16. The original Nature DQN contains approximately 1.68 million trainable parameters, while the introduced SE modules add only 1322 additional parameters, resulting in an increase of approximately 0.08% in the total parameter size. This indicates that FA-QDagger achieves adaptive feature selection without substantially increasing the complexity of the student network.

Following the original SENet design, the reduction ratio is fixed to r = 16, which has been shown to provide an effective trade-off between channel interaction capacity and computational efficiency. To ensure consistency with the original architecture and avoid introducing additional hyperparameter tuning, the same configuration is adopted throughout this work.

The detailed layer configuration of the proposed SE-NatureDQN backbone is summarized in Table 1, where SE blocks are embedded after each convolutional layer to perform channel-wise feature recalibration.

Table 1: Parameter settings of the SE-NatureDQN backbone.

Layer	Operation	Kernel/Stride	Output Channels	Activation	Additional Mechanism
Input	State Stacking	–	4 (84 × 84 pixels)	–	–
Conv1	Conv2D	8 × 8/4	32	ReLU	SE-Block (r = 16)
Conv2	Conv2D	4 × 4/2	64	ReLU	SE-Block (r = 16)
Conv3	Conv2D	3 × 3/1	64	ReLU	SE-Block (r = 16)
FC4	Fully Connected	–	512	ReLU	–
Output	Fully Connected	–	$\|A\|$	Linear	–

3.3 Adaptive Feature Recalibration

Let $X \in R^{H\times W\times C}$ denote the feature map generated by a convolutional layer. The SE block recalibrates X through squeeze, excitation, and scale operations. First, global average pooling compresses the spatial dimensions and generates a channel descriptor $Z \in R^{C}$:

$$z_{\mathrm{c}} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathrm{x}_{\mathrm{c}}(i, j) \tag{1}$$

Second, a bottleneck gating structure models non-linear dependencies among channels and produces the channel weight vector s:

$$\mathrm{s} = \sigma(\mathrm{W}_2 \cdot \delta(\mathrm{W}_1 \cdot \mathrm{z})) \tag{2}$$

where δ denotes the Rectified Linear Unit (ReLU), σ denotes the sigmoid function, and W1 and W2 are learnable parameters. Finally, the original feature map is recalibrated by channel-wise multiplication:

$$\tilde{\mathrm{x}}_k = s_k \cdot \mathrm{x}_k \tag{3}$$

Through these operations, SE-NatureDQN obtains a dynamic feature-selection capability. During knowledge transfer, this mechanism can reduce the influence of redundant visual channels and encourage the student to focus on features that are more relevant to policy improvement.

3.4 Training Objective

Guided by the Q-Dagger framework, the hybrid training objective ensures a smooth transition from imitating the teacher to self-evolution for the agent. During the offline distillation phase, the student agent initializes its policy by minimizing the divergence between its policy distribution π_S and the teacher's policy distribution π_{T}. A knowledge distillation loss with a temperature parameter τ is adopted:

$$\mathcal{L}_{distill} = \mathbb{E}\left[\sum softmax(q_T/\tau) \ln\left(\frac{softmax(q_T/\tau)}{softmax(q_S/\tau)}\right)\right] \tag{4}$$

Online Fine-tuning Phase: To maximize environmental rewards while retaining inherited knowledge, we design a hybrid loss function:

$$\mathcal{L}_{total} = \mathcal{L}_{TD} + \lambda \cdot \mathcal{L}_{distill} \tag{5}$$

where $\mathcal{L}_{TD}$ is the standard Temporal Difference error (e.g., Huber Loss) used for precise value estimation. The parameter λ is a dynamically decaying coefficient that balances temporal difference learning and continued imitation of the teacher. To ensure optimal sample efficiency during the online fine-tuning phase, we employ a performance-based adaptive decay schedule rather than a fixed step-based linear decay. Specifically, the coefficient is initialized to $\lambda_0 = 1.0$. During training, the dynamic multiplier λ_t is determined by an affine function of the student's normalized performance score, formulated as $\lambda_t = \lambda_0 \times \max(1.0 - s_t, 0.0)$, where s_t represents the student agent's current score normalized against the teacher's performance. This dynamic mechanism guarantees that the student's reliance on the distillation loss diminishes precisely in proportion to its own competence. Once the student matches or surpasses the teacher's level ($s_t \geq 1.0$), λ_t automatically drops to zero. Additionally, a fallback hard cutoff is implemented to force $\lambda_t = 0$ after a maximum predefined period to guarantee fully independent policy refinement.

3.5 Qualitative Discussion on Feature-Level Regularization and Overestimation Control

While providing a formal mathematical bound for overestimation within deep attention networks remains theoretically challenging, we offer a qualitative analysis of how the SE module influences the student network's behavior. In the context of policy-to-value distillation, we hypothesize that the SE block functions conceptually as a dynamic information bottleneck.

By adaptively recalibrating channel-wise feature responses, this mechanism focuses the network's capacity on task-relevant representations while suppressing less informative or noisy channels. Consequently, we argue that this feature-level regularization has the potential to mitigate the accumulation of Q-value overestimation. When the student network is exposed to sub-optimal policies from the teacher, the attention mechanism may help alleviate the blind inheritance of these overestimation biases, empirically leading to more stable policy refinement during the online learning phase.

4 Results and Discussion

4.1 Experimental Setup

The proposed method was evaluated on five Atari 2600 environments: Breakout, Bowling, Seaquest, Asterix, and Space Invaders. These environments were selected because they contain diverse visual dynamics, sparse or delayed rewards, moving targets, and background interference. To increase evaluation difficulty and reduce reliance on deterministic action sequences, sticky actions were enabled with a probability of 0.25 [28].

The offline dataset contained 500K decision steps generated by a suboptimal teacher policy. The online interaction budget was limited to 4M environment frames, corresponding to 1M decision steps under frame skip 4. This restricted budget was used to evaluate whether the proposed method can achieve rapid policy initialization and robust online improvement under constrained computational resources [29,30].

To ensure a fair comparison, all baseline methods were evaluated under the same experimental environment, replay buffer setting, interaction budget, and evaluation protocol. Hyperparameters were configured according to their original papers whenever applicable, while maintaining consistent computational resources and training constraints.

4.1.1 Baselines and Configurations

Four methods were compared. In this study, the Q-Dagger baseline refers to a simplified implementation that follows the original QDagger training protocol while replacing the original Rainbow–Impala backbone with the classical Nature DQN architecture. This design ensures that both the baseline and FA-QDagger share the same backbone, thereby isolating the effect of the proposed SE module. FA-QDagger uses the proposed SE-NatureDQN backbone under the same RRL training protocol. Deep Q-learning from Demonstrations (DQfD) was included as a demonstration-based learning baseline [31], and Jump-Start Reinforcement Learning (JSRL) was included as a representative jump-start learning method [32]. All models were implemented in a JAX-based reinforcement learning framework [33].

To ensure a fair comparison, all baseline methods were trained under the same offline-to-online protocol, and the key hyperparameter settings used in the simulation experiments are listed in Table 2.

4.1.2 Evaluation Metric

Following recent recommendations for reliable reinforcement learning evaluation [34], the interquartile mean (IQM) normalized score was used to reduce the influence of outlier seeds and unstable single-run performance. To ensure a rigorous statistical evaluation, all experiments were conducted using N = 5

independent random seeds across the five evaluated Atari environments. Furthermore, to accurately capture the uncertainty in our aggregate metrics, we computed the 95% stratified bootstrap confidence intervals, which are represented by the shaded regions in all performance curves. Higher IQM normalized scores indicate better overall policy performance across the selected benchmark tasks.

Table 2: Key hyperparameter settings for the simulation experiments.

Category	Parameter	Value	Description
General Settings	Optimizer	Adam	Standard adaptive optimization algorithm
	Learning Rate	1×10^{-4}	Step size for network parameter updates
	Replay Buffer Capacity	1,000,000	Maximum number of transition tuples stored for replay
	Frame Skip & Stacking	4 & 4	Standard spatio-temporal configuration for Atari environments
Phase I: Offline	Offline Training Steps	500,000	Number of offline distillation updates without environment interaction
	Distillation Temperature (τ)	1.0	Hyperparameter controlling the smoothness of soft labels
Phase II: Online	Online Training Steps	1,000,000	Equivalent to 4M environment frames (with frame skip 4)
	Exploration Rate (ϵ) Decay	Linear	Exploration strategy decaying linearly from 1.0 to 0.01

4.2 Quantitative Training Dynamics

Fig. 4 compares the training trajectories of FA-QDagger with the baseline methods. During the offline phase, both FA-QDagger and the standard Q-Dagger baseline rapidly absorb teacher knowledge and achieve strong initial performance. This shows that introducing SE modules does not disrupt teacher-guided policy initialization. The attention-enhanced student can still preserve the jump-start advantage of RRL while adding only a lightweight feature recalibration mechanism.

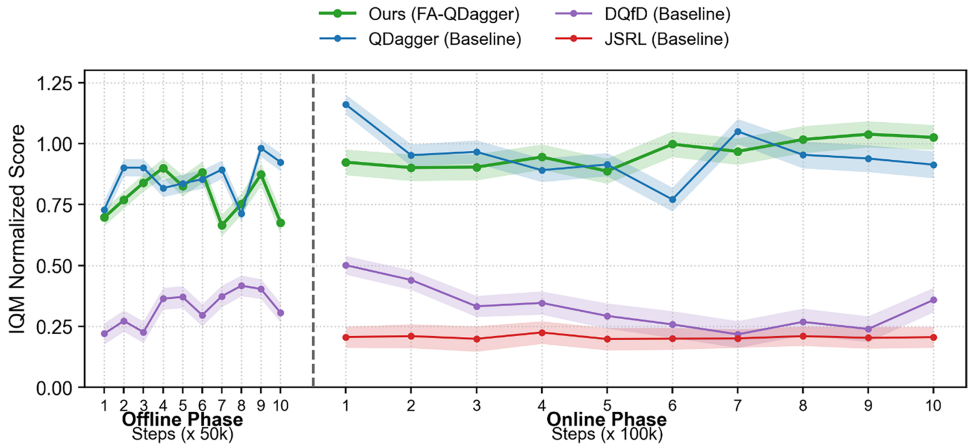

Figure 4: Training dynamics of FA-QDagger and representative baselines during offline distillation and online fine-tuning. The solid lines represent the interquartile mean (IQM) aggregated across 5 Atari environments. The shaded regions denote the 95% stratified bootstrap confidence intervals calculated over N = 5 independent random seeds.

To provide a precise quantitative comparison, Table 3 summarizes the final overall IQM normalized scores of all evaluated methods. As shown, FA-QDagger achieves a final score of 1.04, delivering a 14.3% relative improvement over the standard Q-Dagger baseline (0.91). It is worth noting that the relatively low asymptotic performance observed in certain baseline methods, can be attributed to the heavily constrained experimental settings. Specifically, all methods were evaluated under a strictly limited online interaction budget (4M frames) while initialized with a suboptimal teacher policy. These challenging conditions highlight that traditional methods may struggle to recover from suboptimal priors or early overestimation biases without an extended online exploration phase.

Table 3: Final overall IQM normalized scores and 95% CIs.

Method	IQM Normalized Score	95% CI
DQfD	0.37	0.32–0.39
JSRL	0.22	0.20–0.25
QDagger	0.91	0.87–0.95
FA-QDagger	1.04	0.97–1.09

During the online phase, the difference between the two architectures becomes more apparent. Standard Q-Dagger inherits a strong initial policy but later exhibits unstable fluctuations and a limited asymptotic trend. This suggests that the baseline student may overfit the static teacher distribution and struggle to adapt when teacher guidance is gradually reduced. Across the five evaluated Atari environments, FA-QDagger maintains upward learning momentum and achieves more robust final performance than the compared baselines. These results indicate that channel-wise feature filtering helps the student agent correct inherited feature bias and improve policy adaptation under limited online interaction.

4.3 Visual Interpretability Analysis

Saliency-map visualization was used to analyze how the proposed architecture affects visual focus [35,36]. As shown in Fig. 5, the standard Q-Dagger agent produces relatively diffuse activation patterns. Its attention often spreads to static or irrelevant background regions, suggesting that the baseline convolutional backbone may not sufficiently separate decision-critical objects from environmental noise.

In comparison, FA-QDagger generates more concentrated saliency responses on dynamic objects and task-related regions. This result supports the interpretation that SE-based channel recalibration acts as an implicit feature denoising mechanism. By dynamically weighting feature channels, the attention-enhanced student can better decouple useful visual cues from redundant background information, which improves both interpretability and robustness in dynamic visual environments.

4.4 Ablation Study

To evaluate the contribution of SE module placement, three architectural variants were compared: fully cascaded SE insertion, deep-layer-only insertion, and shallow-layer-only insertion. As shown in Fig. 6, the fully cascaded architecture achieves the most robust performance. This result indicates that feature filtering is required at multiple representation levels.

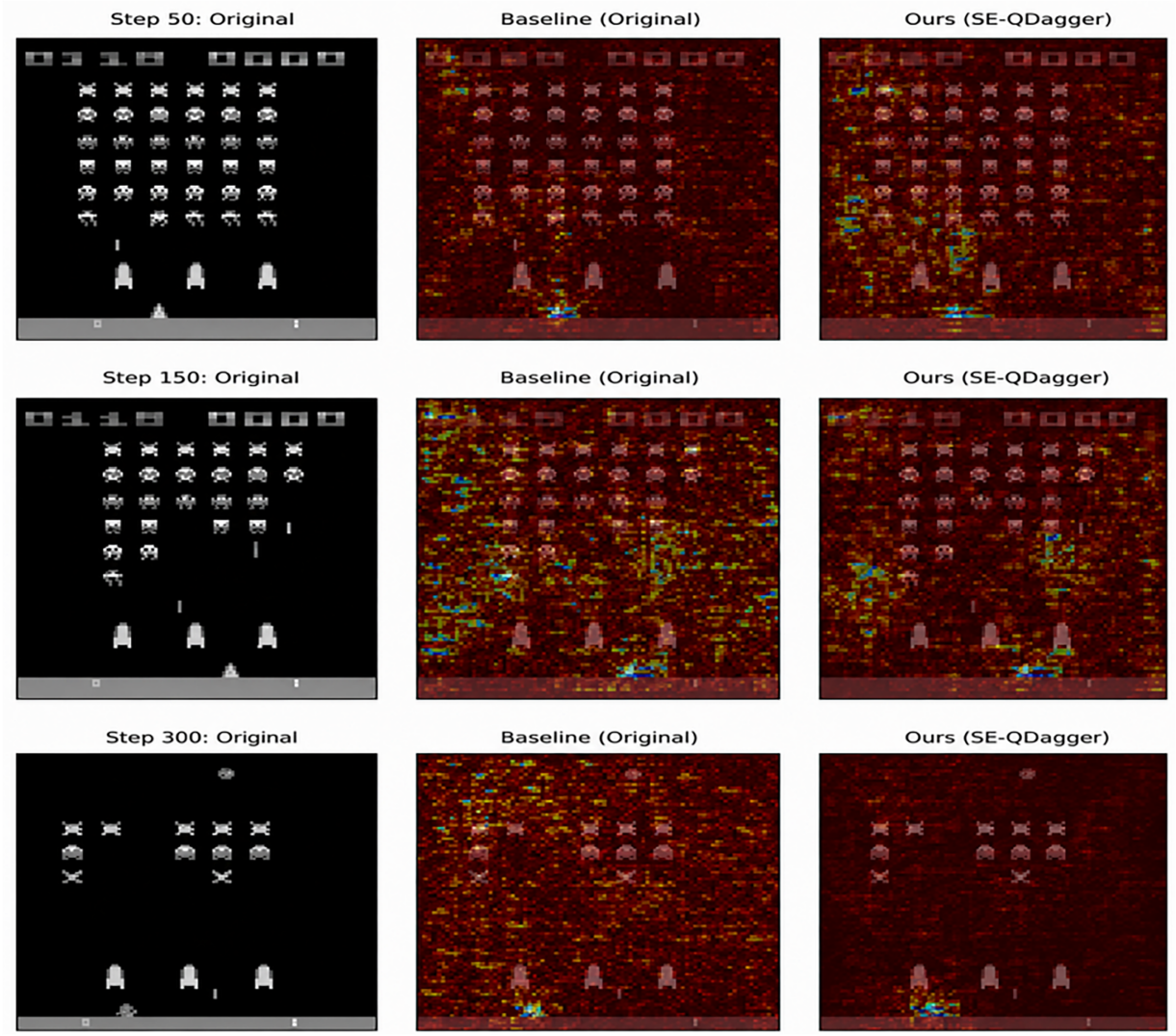

Figure 5: Visual interpretability comparison in Space Invaders. Original frames are shown with saliency maps generated by the baseline Q-Dagger and FA-QDagger agents.

The deep-layer-only variant suffers from instability because pixel-level noise is not sufficiently filtered before entering high-level semantic processing. The shallow-layer-only variant avoids some low-level noise but lacks deep semantic recalibration, leading to a suboptimal asymptotic trend. The fully cascaded design provides a multi-stage information funnel: shallow SE blocks suppress local visual noise, while deeper SE blocks refine abstract decision-related features. This coordinated design explains why FA-QDagger can preserve sample efficiency while improving online robustness.

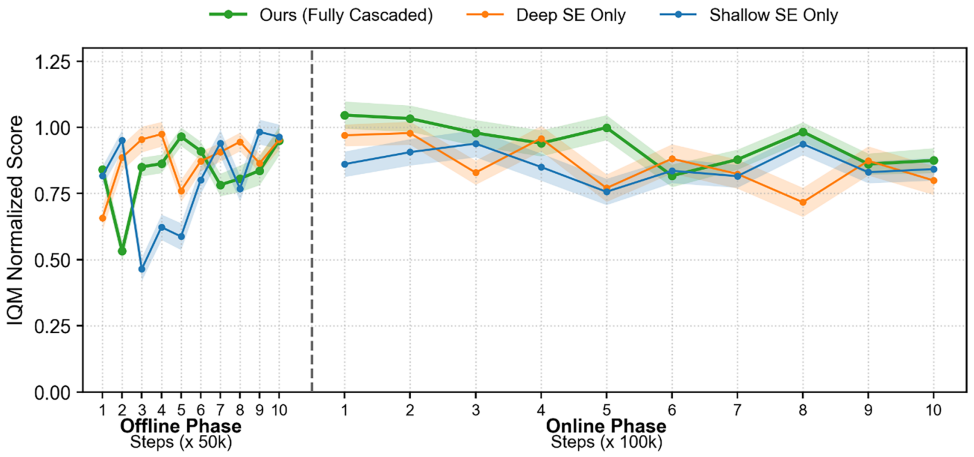

Figure 6: Ablation study on different SE module insertion positions. The solid lines represent the interquartile mean (IQM) aggregated across 5 Atari environments. The shaded regions denote the 95% stratified bootstrap confidence intervals calculated over N = 5 independent random seeds.

4.5 Discussion

The experimental results demonstrate that the main benefit of FA-QDagger lies not only in offline policy initialization but also in the quality of online adaptation. The offline stage ensures that the student obtains an effective initial policy from the teacher. The online stage then reveals whether the student can overcome inherited limitations. Compared with the Q-Dagger baseline implemented with the same Nature DQN backbone, FA-QDagger demonstrates improved performance consistency across the five evaluated Atari environments, suggesting that feature-level regularization may help mitigate unstable knowledge transfer during offline-to-online adaptation.

From the perspective of CMC-oriented intelligent computing, the proposed method can be regarded as a lightweight feature-adaptive agent modeling framework. It improves the efficiency of knowledge reuse without relying on a heavy transformer backbone or complex additional optimization modules. This property is useful for resource-constrained agent training scenarios in which computational cost, sample efficiency, and robustness must be balanced. Because the proposed method is evaluated on a simplified DQN-based implementation of the QDagger framework, the reported performance should primarily be interpreted as demonstrating the effectiveness of the proposed feature-adaptive backbone rather than providing a direct comparison with the official Rainbow–Impala implementation reported in the original RRL literature.

5 Conclusions

This paper proposed FA-QDagger, a feature-adaptive knowledge distillation framework for efficient offline-to-online reinforcement learning. The method addresses the limitation that conventional RRL student networks may indiscriminately inherit redundant background noise and suboptimal teacher representations during offline distillation. By integrating SE channel attention into the standard DQN backbone, the proposed SE-NatureDQN architecture enables dynamic channel-wise feature recalibration and provides the student agent with an intrinsic feature-filtering mechanism.

The proposed framework improves both stages of RRL. During offline distillation, it preserves the initial jump-start advantage by stably absorbing teacher policy knowledge. During online fine-tuning, it suppresses redundant feature inheritance and improves policy plasticity. Experiments conducted on five representative Atari 2600 benchmark environments demonstrate that FA-QDagger consistently improves

asymptotic performance and training robustness under restricted interaction budgets. Although these results validate the effectiveness of the proposed framework on the selected benchmark tasks, evaluating FA-QDagger on a broader range of Atari environments and other reinforcement learning domains remains important future work. Saliency-map analysis and architectural ablation further confirm that the SE module helps the agent develop more concentrated visual attention and more effective feature decoupling.

Future work will extend the experimental evaluation to a broader set of Atari benchmark environments and additional reinforcement learning tasks to further examine the generalization ability of FA-QDagger. First, the method will be generalized to continuous-control algorithms such as Deep Deterministic Policy Gradient (DDPG) and Soft Actor-Critic (SAC). Second, spatiotemporal attention mechanisms may be incorporated to capture long-horizon dependencies in more complex environments. Third, the anti-noise and feature-focusing properties of FA-QDagger will be investigated in sim-to-real transfer tasks, where agents must handle sensor noise, illumination changes, and other real-world uncertainties [37,38].

Although the proposed FA-QDagger demonstrates promising performance under the current experimental setting, several aspects deserve further investigation. In this work, the reduction ratio of the SE module is fixed to r = 16 following the default configuration of the original SENet architecture to ensure consistency and avoid introducing additional hyperparameter tuning. Future work will investigate the influence of different reduction ratios and compare the proposed framework with other lightweight attention mechanisms, such as CBAM and ECA, to further understand the contribution of different feature recalibration strategies. In addition, more comprehensive ablation studies on the individual optimization objectives and evaluations on larger Atari benchmark suites as well as continuous-control tasks will be conducted to further validate the generalization capability of the proposed framework.

Acknowledgement: Not applicable.

Funding Statement: This work is jointly supported by the National Natural Science Foundation of China (No. 52078212, No. 62476086), the Fundamental Research Funds for the Central Universities of Ministry of Education of China of North China Electric Power University (Project No. 2025JG004), State Grid Corporation of China Company Science and Technology Project Research (No. SGSXDKooHLJS2500142), and Beijing Natural Science Foundation (L251012).

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization, Baoping Tian and Zhuxiao Wang; methodology, Baoping Tian and Zhuxiao Wang; software, Baoping Tian and Jiahao Xue; validation, Baoping Tian, Jiahao Xue and Hong Wang; formal analysis, Baoping Tian, Zhuxiao Wang and Hong Wang; investigation, Baoping Tian, Jiahao Xue, Ying Zhang and Yun Ju; resources, Hong Wang, Ying Zhang and Yun Ju; data curation, Baoping Tian and Jiahao Xue; writing—original draft preparation, Baoping Tian; writing—review and editing, Zhuxiao Wang, Hong Wang, Ying Zhang and Yun Ju; visualization, Baoping Tian and Jiahao Xue; supervision, Zhuxiao Wang; project administration, Zhuxiao Wang, Hong Wang, Ying Zhang and Yun Ju; funding acquisition, Hong Wang, Ying Zhang and Yun Ju. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This study used publicly available Atari 2600 benchmark environments from the Arcade Learning Environment. The experimental configurations and implementation details are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

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

References

1. Wang X, Wang S, Liang X, Zhao D, Huang J, Xu X, et al. Deep reinforcement learning: a survey. IEEE Trans Neural Netw Learn Syst. 2024;35(4):5064–78. doi:10.1109/TNNLS.2022.3207346.

2. Li SE. Deep reinforcement learning. In: Reinforcement learning for sequential decision and optimal control. Singapore: Springer Nature; 2023. p. 365–402. doi:10.1007/978-981-19-7784-8_10.
3. Hoang DT, Huynh NV, Nguyen DN, Hossain E, Niyato D. Markov decision process and reinforcement learning. In: Deep reinforcement learning for wireless communications and networking: theory, applications and implementation. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2023. p. 25–36. doi:10.1002/9781119873747.ch2.
4. Hassan T, Hussain I, Haque HMU, Mirza HT, Ali MN, Kim BS. Semantic knowledge based reinforcement learning formalism for smart learning environments. Comput Mater Contin. 2025;85(1):2071–94. doi:10.32604/cmc.2025.068533.
5. Agarwal R, Schwarzer M, Castro PS, Courville A, Bellemare M. Reincarnating reinforcement learning: reusing prior computation to accelerate progress. In: Advances in Neural Information Processing Systems 35; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 28955–71. doi:10.52202/068431-2099.
6. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33. doi:10.1038/nature14236.
7. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. doi:10.1109/CVPR.2018.00745.
8. Bellemare MG, Naddaf Y, Veness J, Bowling M. The arcade learning environment: an evaluation platform for general agents. J Artif Intell Res. 2013;47(1):253–79. doi:10.5555/2566972.2566979.
9. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. Proc AAAI Conf Artif Intell. 2016;30(1):2094–100. doi:10.1609/aaai.v30i1.10295.
10. Schaul T, Quan J, Antonoglou I, Silver D. Prioritized experience replay. In: Proceedings of the 4th International Conference on Learning Representations (ICLR); 2016 May 2–4; San Juan, Puerto Rico.
11. Hessel M, Modayil J, Van Hasselt H, Schaul T, Ostrovski G, Dabney W, et al. Rainbow: combining improvements in deep reinforcement learning. Proc AAAI Conf Artif Intell. 2018;32(1):3215–22. doi:10.1609/aaai.v32i1.11796.
12. Rusu AA, Colmenarejo SG, Gülçehre Ç, Desjardins G, Kirkpatrick J, Pascanu R, et al. Policy distillation. In: Proceedings of the 4th International Conference on Learning Representations (ICLR); 2016 May 2–4; San Juan, Puerto Rico.
13. Ross S, Gordon GJ, Bagnell JA. A reduction of imitation learning and structured prediction to no-regret online learning. In: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics; 2011 Apr 11–13; Fort Lauderdale, FL, USA.
14. Figueiredo Prudencio R, Maximo MROA, Colombini EL. A survey on offline reinforcement learning: taxonomy, review, and open problems. IEEE Trans Neural Netw Learn Syst. 2024;35(8):10237–57. doi:10.1109/TNNLS.2023.3250269.
15. Agarwal R, Schuurmans D, Norouzi M. An optimistic perspective on offline reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning; 2020 Jul 13–18; Virtual.
16. Li L, Jin Z. Shadow knowledge distillation: bridging offline and online knowledge transfer. In: Advances in Neural Information Processing Systems 35; 2022 Nov 28–Dec 9; New Orleans, LA, USA. doi:10.52202/068431-0046.
17. Lee S, Seo Y, Lee K, Abbeel P, Shin J. Offline-to-online reinforcement learning via balanced replay and pessimistic Q-ensemble. In: Proceedings of the 5th Conference on Robot Learning; 2021 Nov 8–11; London, UK.
18. Zheng H, Luo X, Wei P, Song X, Li D, Jiang J. Adaptive policy learning for offline-to-online reinforcement learning. Proc AAAI Conf Artif Intell. 2023;37(9):11372–80. doi:10.1609/aaai.v37i9.26345.
19. Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning. Neurocomputing. 2021;452:48–62. doi:10.1016/j.neucom.2021.03.091.
20. He N, Yang S, Li F, Trajanovski S, Zhu L, Wang Y, et al. Leveraging deep reinforcement learning with attention mechanism for virtual network function placement and routing. IEEE Trans Parallel Distrib Syst. 2023;34(4):1186–201. doi:10.1109/TPDS.2023.3240404.
21. Itaya H, Hirakawa T, Yamashita T, Fujiyoshi H, Sugiura K. Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning. In: Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN); 2021 Jul 18–22; Shenzhen, China. doi:10.1109/IJCNN52387.2021.9534363.

22. Soydaner D. Attention mechanism in neural networks: where it comes and where it goes. Neural Comput Appl. 2022;34(16):13371–85. doi:10.1007/s00521-022-07366-3.
23. Sorokin I, Seleznev A, Pavlov M, Fedorov A, Ignateva A. Deep attention recurrent Q-network. arXiv:1512.01693. 2015.
24. Zhang H, Fang Z, Chen Y, Dai H, Jiang Q, Zeng X. Traffic signal optimization control method based on attention mechanism updated weights double deep Q network. Complex Intell Syst. 2025;11(5):217. doi:10.1007/s40747-025-01841-9.
25. Mott A, Zoran D, Chrzanowski M, Wierstra D, Rezende DJ. Towards interpretable reinforcement learning using attention augmented agents. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019 Dec 8–14; Vancouver, BC, Canada.
26. Abadeh MN. Knowledge-enhanced software refinement: leveraging reinforcement learning for search-based quality engineering. Autom Softw Eng. 2024;31(2):57. doi:10.1007/s10515-024-00456-7.
27. Abadeh MN. Performance-driven software development: an incremental refinement approach for high-quality requirement engineering. Requir Eng. 2020;25(1):95–113. doi:10.1007/s00766-019-00309-w.
28. Delfosse Q, Blüml J, Gregori B, Kersting K. HackAtari: atari learning environments for robust and continual reinforcement learning. arXiv:2406.03997. 2024.
29. Song Y, Zhou Y, Sekhari A, Bagnell JA, Krishnamurthy A, Sun W. Hybrid RL: using both offline and online data can make RL efficient. In: Proceedings of the 11th International Conference on Learning Representations (ICLR 2023); 2023 May 1–5; Kigali, Rwanda.
30. Guo S, Zou L, Chen H, Qu B, Chi H, Yu PS, et al. Sample efficient offline-to-online reinforcement learning. IEEE Trans Knowl Data Eng. 2024;36(3):1299–310. doi:10.1109/TKDE.2023.3302804.
31. Agarwal R, Schwarzer M, Castro PS, Courville A, Bellemare MG. Deep reinforcement learning at the edge of the statistical precipice. Adv Neural Inf Process Syst. 2021;34:29304–20.
32. Fu Y, Wu D, Boulet B. A closer look at offline RL agents. Adv Neural Inf Process Syst. 2022;35:8591–8604. doi:10.52202/068431-0625.
33. Zhang Y, Liu J, Li C, Niu Y, Yang Y, Liu Y, et al. A perspective of Q-value estimation on offline-to-online reinforcement learning. Proc AAAI Conf Artif Intell. 2024;38(15):16908–16. doi:10.1609/aaai.v38i15.29633.
34. Ghasemipour SKS, Schuurmans D, Gu SS. EMaQ: expected-max Q-learning operator for simple yet effective offline and online RL. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual.
35. Milani S, Topin N, Veloso M, Fang F. Explainable reinforcement learning: a survey and comparative review. ACM Comput Surv. 2024;56(7):1–36. doi:10.1145/3616864.
36. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M. SmoothGrad: removing noise by adding noise. arXiv:1706.03825. 2017.
37. Tang C, Abbatematteo B, Hu J, Chandra R, Martín-Martín R, Stone P. Deep reinforcement learning for robotics: a survey of real-world successes. Annu Rev Control Robot Auton Syst. 2025;8(1):153–88. doi:10.1146/annurev-control-030323-022510.
38. Liang A, Thomason J, Bıyık E. ViSaRL: visual reinforcement learning guided by human saliency. In: Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2024 Oct 14–18; Abu Dhabi, United Arab Emirates. doi:10.1109/IROS58592.2024.10801388.