兰州理工大学

科研成果汇总

学　　号：	221081104002
研 究 生：	马佳林
导　　师：	李策　教授
研究方向：	强化学习
论文题目：	视觉注意力下的深度强化学习关键技术研究
学　　科：	控制理论与控制工程
学　　院：	电气工程与信息工程学院
入学时间：	2022 年 9 月

目　录

图书馆 文献检索报告

兰州理工大学图书馆 报告编号：R2025-1247

机构：**兰州理工大学 电气工程与信息工程学院**

姓名：**马佳林 [221081104002]**

著者要求对其在国内外学术出版物所发表的科技论著被以下数据库收录情况进行查证。

检索范围：

- 科学引文索引（Science Citation Index Expanded）： 1900年-2025年

检索结果：

检索类型	数据库	年份范围	总篇数
SCI-E 收录	SCI-EXPANDED	2025	3

委托人声明：

本人委托兰州理工大学图书馆查询论著被指定检索工具收录情况，经核对检索结果，附件中所列文献均为本人论著，特此声明。

作 者（签字）：

完 成 人 （签字）：陈庆怡

完 成 日 期 ：2025年11月6日

完成单位（盖章）：兰州理工大学图书馆信息咨询与学科服务部

（本检索报告仅限校内使用）

附件一：SCI-E收录

#	作者	标题	来源出版物	文献类型	入藏号
1	**Ma, JL**; Li, C; Feng, ZQ; Xiao, LM	Dynamic Visual Attention-based Neuron Awakening and Shifting in deep reinforcement learning	***ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE*** 2025, 158: 111486.	J Article	WOS:001523368900004
2	**Ma, JL**; Li, C; Hong, L; Wei, KL; Zhao, ST; Jiang, HF; Qu, YY	Vision-based attention deep q-network with prior-based knowledge	***APPLIED INTELLIGENCE*** 2025, 55 (6): 565.	J Article	WOS:001450809200003
3	**Ma, JL**; Li, C; Feng, ZQ; Xiao, LM; He, CD; Zhang, Y	Don't overlook any detail: Data-efficient reinforcement learning with visual attention	***KNOWLEDGE-BASED SYSTEMS*** 2025, 310: 112869.	J Article	WOS:001397829000001
				合计	3

第 1 条，共 3 条：

标题： Dynamic Visual Attention-based Neuron Awakening and Shifting in deep reinforcement learning

作者： **Ma, JL** (**Ma, Jialin**); Li, C (Li, Ce); Feng, ZQ (Feng, Zhiqiang); Xiao, LM (Xiao, Limei)

来源出版物： ENGINEERING APPLICATIONS OF ARTIFICIAL INTELLIGENCE **卷：** 158 **文献号：** 111486 **提前访问日期：** JUN 2025 **子辑：** B **出版年：** OCT 22 2025

入藏号： WOS:001523368900004

文献类型： Article **出版物类型：** J

作者地址： [**Ma, Jialin**; Li, Ce; Feng, Zhiqiang; Xiao, Limei] Univ Technol, Coll Petrochem Technol, Pengjiaping Campus 36 Pengjiaping Rd, Lanzhou City **730050**, Gansu Province, Peoples R China.

通讯作者地址： Li, C (corresponding author), Univ Technol, Coll Petrochem Technol, Pengjiaping Campus 36 Pengjiaping Rd, Lanzhou City 730050, Gansu Province, Peoples R China.

第 2 条，共 3 条：

标题： Vision-based attention deep q-network with prior-based knowledge

作者： **Ma, JL** (**Ma, Jialin**); Li, C (Li, Ce); Hong, L (Hong, Liang); Wei, KL (Wei, Kailun); Zhao, ST (Zhao, Shutian); Jiang, HF (Jiang, Hangfei); Qu, YY (Qu, Yanyun)

来源出版物： APPLIED INTELLIGENCE **卷：** 55 **期：** 6 **文献号：** 565 **出版年：** APR 2025

入藏号： WOS:001450809200003

文献类型： Article **出版物类型：** J

作者地址： [**Ma, Jialin**; Li, Ce; Hong, Liang; Wei, Kailun; Zhao, Shutian; Jiang, Hangfei] **Lanzhou Univ Technol**, Sch Informat Sci & Technol, 36 Pengjiaping Rd, Lanzhou **730050**, Gansu, Peoples R China.; [Qu, Yanyun] Xiamen Univ, Sch Informat, 422 Siming South Rd, Xiamen 361005, Fujian, Peoples R China.

通讯作者地址： Li, C (corresponding author), Lanzhou Univ Technol, Sch Informat Sci & Technol, 36 Pengjiaping Rd, Lanzhou 730050, Gansu, Peoples R China.

第 3 条，共 3 条：

标题： Don't overlook any detail: Data-efficient reinforcement learning with visual attention

作者： **Ma, JL** (**Ma, Jialin**); Li, C (Li, Ce); Feng, ZQ (Feng, Zhiqiang); Xiao, LM (Xiao, Limei); He, CD (He, Chengdan); Zhang, Y (Zhang, Yan)

来源出版物： KNOWLEDGE-BASED SYSTEMS **卷：** 310 **文献号：** 112869 **提前访问日期：** JAN 2025 **出版年：** FEB 15 2025

入藏号： WOS:001397829000001

文献类型： Article **出版物类型：** J

作者地址： [**Ma, Jialin**; Li, Ce; Feng, Zhiqiang; Xiao, Limei] **Lanzhou Univ Technol**, Sch Elect Engn & Informat Engn, Lanzhou **730050**, Peoples R China.; [He, Chengdan; Zhang, Yan] Lanzhou Inst Phys, Sci & Technol Vacuum Technol & Phys Lab, Lanzhou **730050**, Peoples R China.

通讯作者地址： Li, C (corresponding author), Lanzhou Univ Technol, Sch Elect Engn & Informat Engn, Lanzhou 730050, Peoples R China.; He, CD (corresponding author), Lanzhou Inst Phys, Sci & Technol Vacuum Technol & Phys Lab, Lanzhou 730050, Peoples R China.

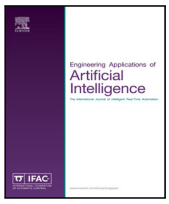
End

Engineering Applications of Artificial Intelligence 158 (2025) 111486

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai



Research paper

Dynamic Visual Attention-based Neuron Awakening and Shifting in deep reinforcement learning

Jialin Ma, Ce Li *, Zhiqiang Feng, Limei Xiao

Lanzhou University of Technology, Pengjiaping Campus: 36 Pengjiaping Road, Qilihe District, Lanzhou City, Gansu Province 730050, China

ARTICLE INFO

Keywords:
Visual attention
Focus on important content
Dynamic visual attention map
Deep reinforcement learning
Industrial applications in dynamic environments

ABSTRACT

Deep reinforcement learning (DRL) has made significant strides in efficient decision-making for complex tasks, yet exploration efficiency remains a primary limitation that affects overall performance. Effective exploration is essential in dynamic environments common to both simulated and real-world scenarios. Inspired by human use of visual information, visual attention mechanisms can help filter irrelevant details in image-based tasks. However, their application in DRL has not achieved expected results. This paper analyzes the limitations of visual attention in DRL and identifies key factors that contribute to performance decline. We propose a new method, "Focus on Important Content" (FIC), to overcome these challenges. First, we address early zero activation of the dynamic visual attention map by resetting the weights of layers reaching a dormancy threshold. Second, we introduce context-aware information to reduce dormant neurons, improving learning efficiency and performance. Our method is tested in the Atari 100K environment, an abstraction of various dynamic scenarios, achieving an Inter Quartile Mean (IQM) of 1.21. Notably, with context-aware mechanisms, FIC reduces dormant neurons by an average of 6% after 100K iterations, showing a significant improvement in exploration efficiency. These findings demonstrate that FIC not only enhances DRL performance but also offers a scalable solution for integrating visual attention mechanisms into DRL, with potential applications in dynamic real-world tasks such as automated inspection, robotic navigation, and object tracking.

1. Introduction

Deep reinforcement learning (DRL) has achieved remarkable success across various domains, particularly in scenarios requiring efficient decision-making and the handling of complex tasks. By emulating human learning mechanisms, it has not only reached but often surpassed human-level performance in addressing diverse problems (Degrave et al., 2022; Ju et al., 2022; Kaufmann et al., 2023). Exploration in DRL refers to the agent's ability to explore unknown parts of the environment, balancing the trade-off between exploiting known strategies and exploring new possibilities. Effective exploration is crucial in environments where optimal solutions are not immediately apparent and require diverse experiences to discover. For example, in applications such as AlphaGo (Silver et al., 2016), OpenAI Five (Berner et al., 2019), and robotics-based autonomous navigation (Dosovitskiy et al., 2017), exploration is not only important but essential for learning effective policies. The exploration cost in these advanced systems is enormous, sometimes equivalent to the efforts of tens of thousands of years of human endeavor. Therefore, improving exploration efficiency is critical for reducing computational resources and accelerating training processes. This need is especially pressing in environments with high-dimensional action spaces and long-term decision-making, such as in robotics, where an agent must explore vast state spaces to discover the most efficient actions. Thus, addressing exploration efficiency in DRL is a central challenge for real-world applications.

Visual deep reinforcement learning (VDRL) (Shi et al., 2022a; Itaya et al., 2021), as a crucial branch of DRL, is extensively applied in various environments such as the arcade learning environment (ALE) (Bellemare et al., 2012) and autonomous driving simulators (Dosovitskiy et al., 2017). In this domain, deep learning (DL) (LeCun et al., 2015; Goodfellow et al., 2016) is tasked with mapping environmental states to value functions or policies (Mnih et al., 2015). Remarkably, Never Give Up (NGU) (Badia et al., 2020) achieves 1344.0% of human-level performance across 57 ALE tasks, Bigger-Better-Faster (BBF) (Schwarzer et al., 2023) achieves human-average exploration efficiency in 26 ALE tasks, and Dense Reinforcement Learning (D2RL) (Feng et al., 2023) accomplishes autonomous driving tasks in multiple scenarios. However, image data contains a significant amount of redundant information regarding the policy. In theory, visual attention (VA) mechanisms (Mnih et al., 2014) can enhance DRL

* Corresponding author.
E-mail addresses: jialinm@lut.edu.cn (J. Ma), lice@lut.edu.cn (C. Li), feng_zq@lut.edu.cn (Z. Feng), xlm@lut.edu.cn (L. Xiao).

https://doi.org/10.1016/j.engappai.2025.111486
Received 28 October 2024; Received in revised form 12 April 2025; Accepted 10 June 2025

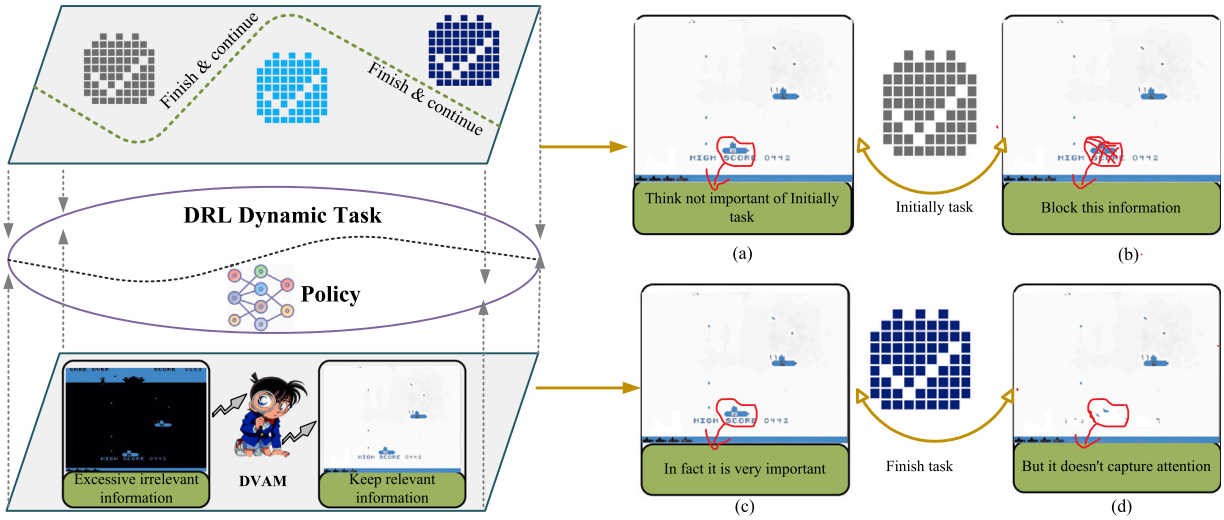


Fig. 1. The impact of DRL on DVAM. When making decisions, humans focus on relevant visual elements. In DRL's dynamic environments, initial tasks must be finished before moving to the next ones, as shown on the left of the figure. However, early Dynamic Visual Attention Maps (DVAM) (in (a)) can hide important information, especially task-related details (in (b)). If DVAM masks key information, it may cause gradient disappearance during updates (as seen in (c) and (d), discussed in Chapter Four), which can slow down training.

efficiency by focusing on task-relevant information. Nevertheless, our research indicates that when applying VA mechanisms to DRL, their performance often falls below expectations and may even lead to policy failures (Sorokin et al., 2015).

The motivation for our study stems from that human vision is inclined to focus solely on essential content, and VA (Leng et al., 2023; Shi et al., 2023; Yuan et al., 2023) can mitigate redundant information in images. Broadbent's Filter Model (Broadbent, 2013) suggests that, during the early stages of information processing, a filter based on physical characteristics (such as color and orientation) selectively screens information, permitting only a fraction to reach the conscious level. This process aims to reduce the demand on cognitive resources. While this model has demonstrated performance enhancement in other domains, its application in the field of DRL has not met theoretical expectations. We explore this discrepancy and discover that, in contrast to supervised learning (Cunningham et al., 2008), the data processed in DRL undergoes changes with variations in the environment (Kaelbling et al., 1996). In supervised learning, data typically originates from the same task or distribution, allowing VA mechanisms to highlight essential information and suppress irrelevant details based on the task. However, in DRL, as illustrated in Fig. 1, information initially deemed irrelevant and suppressed may become crucial later on. Unfortunately, once this information is masked by the VA mechanism, corresponding neurons may enter a dormant state (Sokar et al., 2023a), making reactivation challenging in later stages. This implies that even though this information becomes important later, the neural networks' inability to "see" it eventually leads to training failure.

Building on the considerations above, this paper introduces a method named "Focus on Important Content" (FIC), with the objective of addressing the limitations of VA mechanisms in DRL. To achieve this, we introduce the concept of awakening dormant neurons and combine it with a strategy involving the shift of information between windows. This approach aims to prevent neurons, once awakened, from starting learning from a completely random state. The method not only has the potential to enhance the performance of DRL in visual tasks but also paves the way for new opportunities in improving overall exploration efficiency. In summary, the primary contributions and innovations of this paper transcend a comprehensive analysis of existing issues and solution design. They introduce novel perspectives and advancements in both theory and methodology, encompassing:

- **A thorough analysis of the limitations associated with applying VA in DRL.** This exploration delves into potential reasons for the diminished performance of VA in existing DRL, specifically attributing it to the early blocking of task-irrelevant information caused by dynamic learning tasks, impeding reactivation.
- **The innovative design of a VA mechanism that incorporates the awakening of dormant neurons.** This addresses a critical challenge in VDRL, where significant weights are zero in the early stages, resulting in the dormancy of certain neurons and impeding training. Effective strategies are proposed to overcome this obstacle.
- **The introduction of the information shift between windows strategy, further optimizing the application of the VA mechanism.** This strategic approach aims to diminish the generation of dormant neurons.

The subsequent section in this paper are organized as follows: Section 2 introduces the foundational theory of DRL. Section 3 covers the application of sample efficiency in reinforcement learning and explores the utilization of attention mechanisms in VDRL. Section 4 introduces the origin of the identified issues. Section 5 presents the proposed methodology. Section 6 details the experiments. The final section provides a comprehensive summary of the paper.

2. Background

Reinforcement learning (RL) tasks (Kaelbling et al., 1996; Wiering and Van Otterlo, 2012) are typically described using a Markov Decision Process (MDP) (Baxter, 1995) model, defined by a tuple $(S, A, \mathcal{R}, \mathcal{P})$ where S represents the set of states, A represents the set of actions, $\mathcal{P} : S \times A \to \mathcal{P}(s, a)$ represents the state transition function, and $\mathcal{R} : S \times A \to \mathcal{R}(s, a)$ represents the reward function. The agent's policy is expressed as a mapping from actions to states through $\pi : S \to A$, and the value function $V_\pi(s) := \mathbb{E}_{\pi,\mathcal{P}(S,A)}\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, a_t\right)\right]$ expresses the value π of s, where $\gamma \in (0, 1)$ represents the discount factor, ensuring the boundedness of the value function. In RL, the agent's objective is to find a policy $\pi^\star$ such $V_\pi^\star \geqslant V_\pi$ that for all π.

There are various methods for solving $\pi^\star$ (Silver et al., 2014; Konda and Tsitsiklis, 1999), and our work is limited to model-free value-based approaches (Watkins and Dayan, 1992). In the domain of value-based methods, our focus lies in the estimation of Q-values, which are expressed through Bellman recurrence $Q(s, a) := \mathcal{R}(s, a) + \gamma \mathbb{E}_{s\prime \sim \mathcal{P}(s,a)}[\max_{a\prime \in A} Q(s\prime, a\prime)]$. Furthermore, the optimal policy $Q^\star(s, a)$ is

derived by maximizing $\pi^{\star}$, where $\pi^{\star}(s) := \max_{a \in A} Q^{\star}(s,a)$. The solution for is attained through Bellman temporal difference, as formulated in Eq. (1). Here, $\mathcal{R}\left(s_t,a_t\right)+\gamma \max_{a_{t+1}} Q\left(s_{t+1},a_{t+1}\right)$ is defined as the Bellman target.

$$\Delta Q\left(s_t,a_t\right)=\left(\mathcal{R}\left(s_t,a_t\right)+\gamma \max_{a_{t+1}} Q\left(s_{t+1},a_{t+1}\right)\right)-Q\left(s_t,a_t\right) \tag{1}$$

In the early practices of RL, was commonly represented in tabular form, and it has been demonstrated that the Bellman Temporal Difference (TD) method can converge to a unique solution. For linear relationships, approximators $Q(s,a)$ for can be provided through methods such as linear regression. However, when dealing with visually intensive tasks characterized by highly nonlinear properties, finding effective and efficient nonlinear approximators becomes a crucial challenge. The development of deep learning has introduced new solutions to address this challenge.

Mnih et al. (2015) introduced the deep Q-learning (DQN) method by combining DL with the TD method, achieving human-level performance on the ALE (Bellemare et al., 2012). In the paper, a model comprising three convolutional layers and two fully connected networks is proposed as an approximator for the Q-function, where represents the optimized weights. The majority of value-based work is grounded in DQN, and Hessel et al. (2017) integrated six methods to propose the Rainbow DQN approach. These six methods include prioritized experience (Schaul et al., 2016), n-step learning (Sutton, 1988), distributional RL (Bellemare et al., 2017), double Q-learning (Van Hasselt et al., 2015), dueling architecture (Wang et al., 2016), and NoisyNets (Fortunato et al., 2017).

3. Related work

3.1. Sample efficiency in reinforcement learning

Due to the high cost of interacting with the environment in RL, Kaiser et al. (2020) addressed the challenge of sample efficiency in RL by introducing the Atari 100K benchmark. Kostrikov et al. (2020) devised a sample-efficient RL method named DrQ, incorporating data augmentation, thereby surpassing previous approaches on the Atari 100K. Both Data-Efficient Rainbow (DER) (Van Hasselt et al., 2019) and DrQ(ϵ) (Agarwal et al., 2021) outperformed existing methods simply by adjusting the hyperparameters of prevailing model-free algorithms without introducing any algorithmic innovations. Schwarzer et al. (2020) introduced self-predictive representations (SPR), building upon Rainbow (Hessel et al., 2017), and employed a self-supervised temporal consistency loss based on BYOL (Grill et al., 2020), combined with data augmentation. The integration of SPR with periodic network resets, known as SR-SPR (Nikishin et al., 2022), achieved state-of-the-art performance, with the IQM increasing from 0.337 to 0.631. The effectiveness of this reset mechanism was also demonstrated in ReDO (Sokar et al., 2023b). However, when VA was integrated with resets, performance dropped to 0.38 due to the suppression of key visual information. DOAD (Ma et al., 2025) addressed this by introducing a constraint to keep VA output above 0.5 and applying various resetting techniques to stabilize learning. These methods effectively demonstrated the feasibility of combining VA with resets, though they also increased model complexity, highlighting the need for careful tuning of context-aware mechanisms and resetting strategies to manage complexity in practical applications. Ye et al. (2021) incorporated a self-supervised consistency loss similar to SPR (Chen and He, 2021). EfficientZero (Ye et al., 2021), an efficient variant of MuZero (Schrittwieser et al., 2020), learned a latent dynamic model of discrete actions from environmental interactions and selected actions through forward-looking MCTS in the model's latent space. Micheli et al. (2023) presented IRIS, a data-efficient agent learning within a world model composed of an autoencoder and an autoregressive transformer. Sokar et al. (2023a) identified the issue of neuron dormancy in DRL and proposed a method to awaken dormant neurons. Schwarzer et al. (2023) achieved model-free human-level performance by optimizing the hyperparameters of SR-SPR.

Despite the significant achievements of the aforementioned methods in improving the sample efficiency of RL, image data contains a considerable amount of redundant information concerning the policy. In theory, VA mechanisms can enhance the efficiency of DRL by focusing on task-relevant information. Therefore, we believe there is another crucial dimension worth exploring: leveraging VA mechanisms to further improve learning efficiency. While there is currently no universally recognized, definitive method for addressing sample efficiency issues in RL through VA, recent years have witnessed a series of studies attempting to apply VA mechanisms to RL. The goal is to enhance learning efficiency and performance in complex visual task environments. In the following, we will explore these methods, with a particular focus on the application of visual saliency in RL.

3.2. Visual attention for reinforcement learning

In Atari Games, an agent represents a vision-based RL ask. Introducing human attentional behaviors allows the agent to focus more on task-relevant information, providing a certain level of interpretability. One approach involves incorporating attention mechanisms at the feature layer, aiming to enhance the agent's ability to focus on specific aspects in visual task. Sorokin et al. (2015) introduced both soft and hard attention on features. Mott et al. (2019) designed an attentional pattern, querying spatial and content information separately. Itaya et al. (2021) implemented attention mechanisms in both the actor and critic branches, expressing the network's focus areas for each branch output. Guo et al. (2021) discovered that various RL algorithms consistently focus on crucial visual features during the learning process, resembling human behavior. Wang et al. (2021) employed an unsupervised method to extract the visual foreground as the input for RL.

Another approach involves directly introducing visual saliency weights into the input, focusing on specific content in the visual input while minimizing the information received by the agent to ensure the retention of all relevant details. Nikulin et al. (2019), utilizing eye-tracking data collected from human gamers, employed an attention module based on Free Lunch Saliency (FLS) to generate DVAMs. These DVAMs visualize the importance of different parts of the input in the agent's current decision-making process. Wu et al. (2021) designed a self-supervised attention method that guides the agent to select interesting regions during the learning process. Shi et al. (2022a) proposed an approach where the agent utilizes DVAM to assign weights to different parts of an image. These weights are determined by the relevance of each part to the task, effectively reducing task-irrelevant redundant information. Shi et al. (2022b) designed a dedicated causality discovery network to generate high-resolution and clear attention masks, highlighting task-relevant spatiotemporal information. These masks constitute the most critical evidence for a visual-based RL agent to make continuous decisions.

In addition to these foundational studies, recent work has applied attention mechanisms across various domains, highlighting their relevance in improving RL performance. For example, the Multi-Dimensional Attention Fusion Network has been used in tasks like terahertz image super-resolution (Wu et al.). Attention mechanisms have also been explored in transportation mode detection (Merikhipour et al., 2025) and domain adaptation for Atari environments (Carr et al., 2018), demonstrating their broader utility. Other applications, such as resilient RL for voltage control in microgrids (Laskar et al., 2023) and AI-driven customer service (Chaturvedi and Verma, 2023), further emphasize the versatility of attention-based strategies.

Given the surplus of redundant information in image data concerning policies, VA mechanisms theoretically hold the potential to

Table 1
Comparison of key contributions.

Method	World model-based	SPR	Visual attention	Reset mechanism
BBF (Schwarzer et al., 2023)	✗	✓	✗	Complete Reset
Dreamer V3 (Ye et al., 2021)	✓	✗	✗	None
EfficientZero (Ye et al., 2021)	✓	✗	✗	None
DOAD (Ma et al., 2025)	✗	✓	✓	Complete Reset
ReDO (Sokar et al., 2023b)	✗	✗	✗	Dormancy Reset
FIC (our method)	✗	✗	✓	Dormancy Reset + Shift

improve the efficiency of DRL by concentrating on task-relevant information. However, our observation reveals that the introduction of visual saliency leads to the phenomenon of dormant neurons, resulting in the suppression of early task-irrelevant patterns in DRL. Consequently, this impedes the subsequent updates of DVAMs in later stages. Addressing this issue has become the primary focus of our research. Table 1 summarizes the main contributions of FIC compared to other reported works.

4. Problem of visual attention in deep reinforcement learning

Enhancing task-relevant content and reducing or blocking task-irrelevant information from visual input through saliency weights can enhance the performance of RL training. However, in our early investigations, we observe that the direct introduction of VA mechanisms can yield adverse effects. Therefore, we conduct an initial analysis and formulated a hypothesis regarding this phenomenon, attributing it to the limited reward signals received during early tasks, leading to the incorrect blocking of task-related information. Once neurons enter dormancy early on and cannot be reawakened, there is a lack of pertinent information in later stages, resulting in training failure.

4.1. Unable to awake dormant neurons

Utilizing DRL to guide VA, reducing task-irrelevant redundant information, and consequently enhancing the performance of DRL is a promising approach. Nevertheless, the application of this study in DRL faces limitations, primarily in the agent's challenge of effectively learning strategies from this pattern. To analyze this phenomenon, we define the DVAMs obtained through convolutional operations, as shown in Eq. (2).

$$Z(i,j) = \sum_m \sum_n X(i+m, j+n) \cdot W(m,n) \tag{2}$$

where, X represents the input image, W denotes the convolutional kernel, and Z is the resulting feature map. We apply the ReLU activation function to further process this feature map, as illustrated in Eq. (3).

$$Y(i,j) = \max(0, Z(i,j)) \tag{3}$$

During $Y(i,j) = 0$, Neumann etal [21] referred to this state as neuron dormancy. In this state, the observed image parts $X(i,j) \cdot Y(i,j)$ are set to zero. In the process of weight updates, the update of the weight W is represented by Eq. (4).

$$\frac{\partial Y}{\partial W} = \frac{\partial(\mathrm{ReLU}(Z))}{\partial Z} \times \frac{\partial Z}{\partial W} \tag{4}$$

For the case of $\mathrm{ReLU}(Z) = 0$, the derivative $\frac{\partial(\mathrm{ReLU}(Z))}{\partial Z} = 0$ implies that the weight W corresponding to the feature map parts of $Y(i,j)$ will not be updated.

In a supervised learning context, if the saliency score for a certain part $X(i:i{+}m, j:j{+}n)$ learns to be 0, it is generally considered correct behavior, as it indicates that this part's pattern is irrelevant to the task. However, in RL, the state-optimal policy distribution $D : S \to \pi^\star$ changes over the learning process. The agent provides a policy through $\pi(s)$. Consider a scenario: when $\mathrm{DVAM}(X_{sub}) = 0, \pi(\cdot) \sim D$, i.e., part X_{sub} in the image is irrelevant to the current policy distribution D. In this situation, the agent can still offer an optimal policy $\pi^\star$ even without information about X_{sub}, effectively reducing redundancy in visual input. However, if the distribution $\tilde{D} : S \to \pi^\star$ changes, and the agent continues to block X_{sub}, it may lead to training failure. If X_{sub} becomes important in the new task $\pi(\cdot) \sim \tilde{D}$, the agent cannot observe information about X_{sub}, resulting in an inability to adapt and potentially causing training failure.

To illustrate this point concretely, we design a continuous two-scenario task. In the first scenario, the agent needs to avoid the red elements while treating the blue elements as passable, as depicted in Fig. 2(a). In the second scenario, the situation is reversed, with the blue elements becoming obstacles to avoid, and the red elements becoming passable, as shown in Fig. 2(b). In the first scenario, the agent should learn to assign a saliency score of 0 to the blue elements, indicating that no salient blue elements should be present in the image after saliency weighting. However, when the agent transitions to the second phase, due to the dormancy effect of neurons, its parameters cannot be updated for the new scenario. Consequently, it continues to ignore the blue elements, fails to reawaken the neurons, and directly results in training failure in the new scenario.

4.2. Lost information

The Pong game involves a player controlling the right paddle, competing against a computer-controlled left paddle. The objective is to continually deflect the ball away from one's goal and score points by propelling it into the opponent's goal. Over 21 rounds of the game, the difference between the player's victories and defeats serves as the evaluation metric for player performance. We attempt to apply saliency weights to the original images. The agent utilized nature Convolutional Neural Networks (CNN) (Mnih et al., 2015) to extract visual features and underwent training using the DQN-Adam method (Raffin et al., 2021a). The agent's exploration rate gradually decreased from 1 to 0.05 over the initial 0.1 million steps, maintaining a constant exploration rate of 0.05 thereafter.

In two training sessions, the agent achieved a score of −21 at 10 million steps. We monitor its DVAMs, as shown in Fig. 3.

We are surprised to discover that saliency weights obtained during the training process are consistently zero. After 10 million steps, the agent with attached saliency weights exhibits a performance drop from 17.89 to −20.97 compared to its performance without attached saliency weights. At 0.1 million steps, the agent's score drastically plummets to a remarkably low -21, and prior to the 0.1 million steps, the agent's highest score is −20.50 under a completely random policy. In the Pong game, the agent receives a reward of 1 only upon winning, with all other sparse rewards of 0. This implies that during the early training process, any part of the image has no influence on the outcome, i.e., $\frac{\partial Q}{\partial \mathrm{DVAM}(X(i,j))} = 0, \forall i,j$. Throughout training, under task-driven conditions of $\mathrm{DVAM}(X(i,j))$, is independent of the policy, as $\mathrm{DVAM}(X(i,j)) \to 0$. Once $\mathrm{DVAM}(X(i,j)) = 0$, in the case of neuron dormancy, the agent can no longer access information about $X(i,j)$, resulting in the failure of the VDRL task with introduced saliency weights.

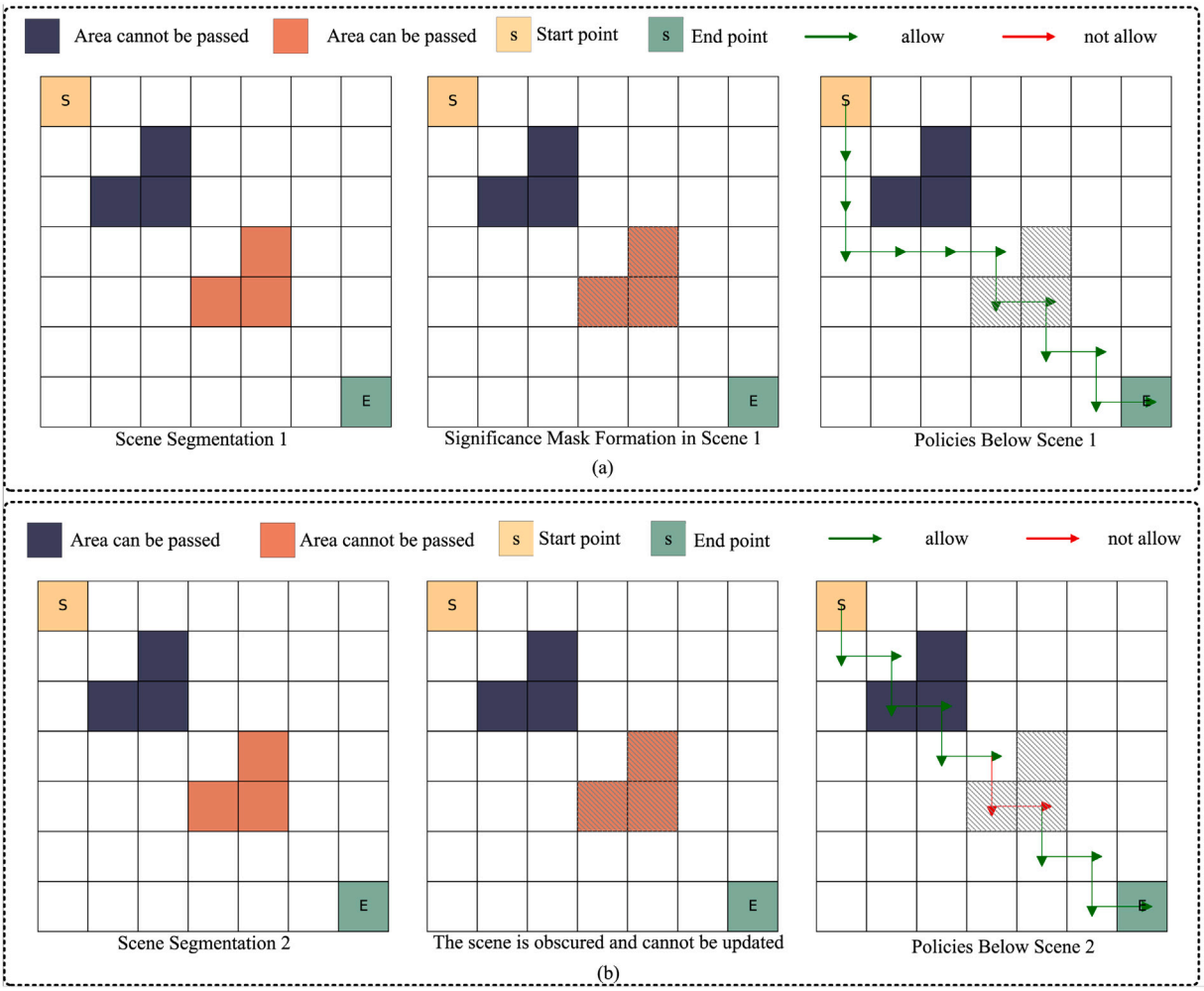


Fig. 2. Path planning for different task scenarios. In Scenario one, the orange block is the starting point and the green block is the endpoint. The agent navigates through obstacles, with red indicating passable areas and blue indicating impassable areas. In Scenario two, the same color scheme applies, but the areas are reversed: blue represents passable regions and red indicates impassable regions.

5. Focus on important content method

In the early stages of training, this study observed the phenomenon of neuron dormancy during the extraction of DVAMs, resulting in the concealment of crucial visual information by the DVAMs and subsequently affecting learning efficiency and performance. To address this challenge, drawing inspiration from the research conducted by Kaelbling et al. (1996), our study introduces two novel strategies aimed at optimizing the VA mechanism in DRL:

- **Awaken Dormant Neurons:** Within the generation process of DVAMs, a strategy is implemented to awaken dormant neurons. This approach facilitates the reactivation of information that is prematurely suppressed, enabling the agent to reassess the relevance of this information to the ongoing task.
- **Information Shift Between Windows Strategy:** To tackle the challenge of relearning weights for awakened dormant neurons, we incorporate the information shift between windows strategy in the encoder. This operation reduces the proportion of dormant neurons during the training process by leveraging contextual aware information from surrounding regions.

5.1. Visual attention Q-network

To tackle the issue of the inefficacy of directly introducing VA mechanisms, we devise a Q-network grounded in VA. This network initially extracts DVAMs from images using an encoder–decoder structure. The encoder employs a multi-layered convolutional network to grasp deep features of the image, while the decoder reconstructs the DVAMs using these features. Each convolutional layer in the encoder is tailored to capture various aspects of the image, spanning from basic edges and textures to more intricate shapes and patterns. A ReLU activation function is incorporated after each convolutional layer in the encoder, and the first deconvolutional layer in the decoder is likewise followed by a ReLU function. Subsequently, saliency weights are amalgamated with the original image through element-wise multiplication, i.e., $\tilde{x} = \mathrm{DVAM}(x) \odot x$, where the intensity of each pixel adjusts based on its saliency weight to underscore key features. Following this, Q-values are computed through a feature extractor and the Rainbow Q-network, i.e., $\mathcal{Q}(x) := \mathcal{Q}_{\theta_Q}(\tilde{x})$. The structure is illustrated in Fig. 4.

Due to the ReLU function inducing neuron dormancy when its output is zero, i.e., negative inputs are mapped to zero, there is a risk in the early stages. If the feature map generated by the encoder assigns a saliency score of 0 to crucial patterns in the image, the neurons responsible for this part remain dormant and cannot be reactivated, as

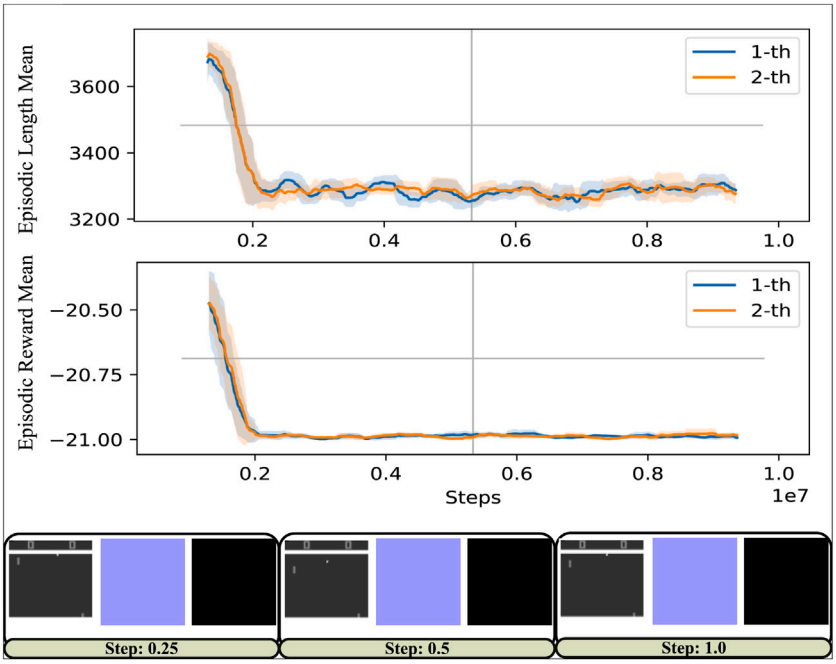


Fig. 3. Pong game's average reward and steps. Illustrates the average reward and game steps per round in the Pong game at 10 million steps. Below, state images at 25%, 50%, and 100% are displayed, accompanied by the obtained DVAMs and the weighted DVAMs.

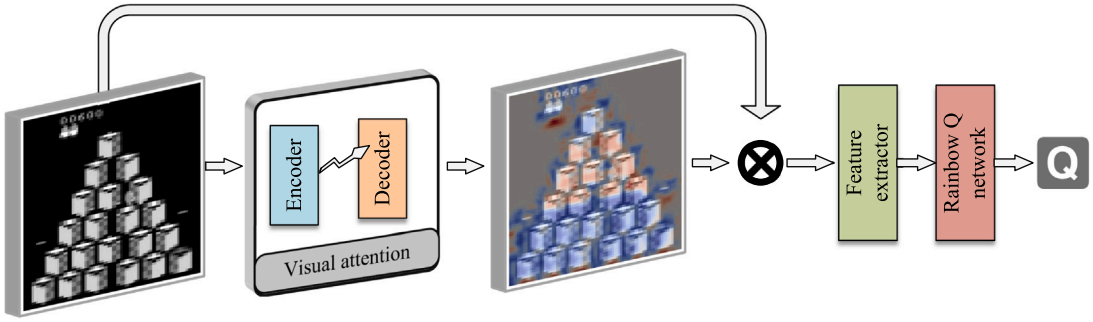


Fig. 4. Visual attention Q-network. The network first extracts DVAMs from images using an encoder–decoder. The saliency weights are then combined with the original image through element-wise multiplication. The color of the saliency weights ranges from blue to red, with values between 0 and 1, reflecting the strength of the agent's attention.

shown in Fig. 5. Consequently, the DVAMs output directly obstructs information from this portion of the image in our structure. For instance, in the Pong game, if this information corresponds to the ping-pong ball, the information processed by the agent will lack this crucial content, rendering it incapable of making effective downstream task decisions and causing catastrophic impacts on training.

To address this challenge and consider the potential dormancy of neurons when the ReLU output is zero, we employ a strategy to awaken dormant neurons during the training process. Additionally, we reinitialize dormant neurons using the Xavier uniform distribution (Glorot and Bengio, 2010). Specifically, we identify neurons in the lth convolutional kernel that satisfy the following Eq. (5) and then reinitialize each found dormant neuron through the Xavier uniform distribution for each layer.

$$\sum_{l}\sum_{m}\sum_{n} X(i+m, j+n)\cdot W(m,n) \le 0 \tag{5}$$

here, l represents the convolutional kernel, while m and n denote the number of convolutions. This formula signifies that for a given window $X(i : i+m, j : j+n)$, when the convolution operation of the lth convolutional kernel, denoted as$W(0 : m, 0 : n)$, results in a value less than 0, it is termed as dormancy, indicating that neuron $W(0 : m, 0 : n)$ produces dormancy for neuron $X(i : i+m, j : j+n)$. Although this method effectively awake dormant neurons, ensuring the stability of the DVAMs and the integrity of information, thereby optimizing the entire reinforcement learning process, the reinitialized weights of dormant neurons may influence overall learning efficiency and performance.

5.2. Reducing dormant neurons in dynamic attention tasks

To address the aforementioned issue, we are considering the design of a method to reduce the occurrence of neuron dormancy. Initially, we review the reasons for the generation of dormant neurons. For a specific window $X(i : i+m, j : j+n)$, the convolution operation of the $W(0 : m, 0 : n)$ convolutional kernel for each layer is required to be less than 0, as shown in Eq. (6), where l represents the lth convolutional kernel, denoted as dormancy of neuron $W(0 : m, 0 : n)$ for neuron $X(i : i+m, j : j+n)$.

$$Z(i,j) = \sum_{l}\sum_{m}\sum_{n} X(i+m, j+n)\cdot W_l(m,n) < 0 \tag{6}$$

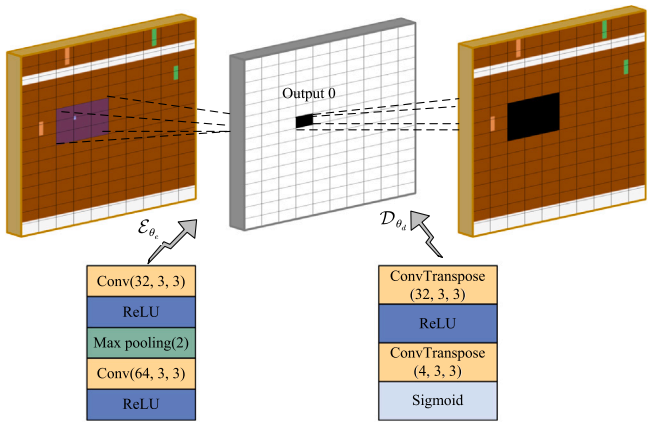


Fig. 5. Chat of dormant neuron in VA. If the feature map generated by the encoder assigns a saliency score of 0 to crucial patterns in the image, the neurons responsible for this part remain dormant and cannot be reactivated.

In response to this phenomenon, we enhance the structure of the encoder and decoder by introducing contextual aware information from feature maps, as illustrated in Fig. 6.

In the encoder section of the VA network, after two convolution operations, we introduce a shift module. The shift module is implemented by four operations, as shown in Eq. (7).

$$\begin{aligned} X_{\text{shift right}}(i:i+m, j:j+n) &= X(i:i+m, j-1:j+n-1) \quad (a)\\ X_{\text{shift left}}(i:i+m, j:j+n) &= X(i:i+m, j+1:j+n+1) \quad (b)\\ X_{\text{shift down}}(i:i+m, j:j+n) &= X(i-1:i+m-1, j:j+n) \quad (c)\\ X_{\text{shift up}}(i:i+m, j:j+n) &= X(i+1:i+m+1, j:j+n) \quad (d) \end{aligned} \tag{7}$$

By performing a pixel-wise shift operation (moving one pixel up, down, left, or right), the model can capture more robust feature representations. Even if the value of a particular pixel might lead to neuron dormancy (output less than 0), the values of surrounding pixels may provide complementary information, facilitating the awakening of that neuron. This reduction in dormancy, attributed to individual pixel values, is achieved by leveraging the shift operation. It enables the use of contextual aware information to optimize the weight learning process when reactivating neurons, mitigating the efficiency and performance degradation discussed in Section 5.1. Thus, it effectively addresses the issue of dormancy resulting from the deactivation of individual pixel references, as illustrated in Fig. 7.

Using the 3×3 convolutional kernel, we extract the saliency map within the $X(i+1:i+5, j+1:j+5)$region and define nine adjacent feature maps as DVAM$(i+2:i+4, j+2:j+4)$. Here, the information in the middle is assigned a weight of 0 saliency, i.e., DVAM$(i+2, j+2) = 0$. When the saliency map is attached to the region, the information of $X(i+2, j+2)$ is lost, and due to neuron dormancy, it remains as DVAM$(i+2, j+2) = 0$ throughout subsequent training. Due to the nature of the sliding window in convolution, adjacent windows contain related information, and the saliency score of this information is closely related to the central information, i.e., $|\text{DVAM}(i+2, j+2) - \text{DVAM}(i+m, j+n)| < \varepsilon, m = (1,3), n = (1,3)$. By performing the shift operation on feature maps of different channels, involving pixel-wise shifts in all directions, the context aware information of a feature map with specific content is dynamic due to the dynamic input of RL. Even if the output of the shifted feature map becomes 0, it will still yield a non-zero output after the dynamic input changes, as illustrated in Fig. 8.

This approach not only awake dormant neurons but also enhances the smoothness of the DVAMs. Simultaneously, it ensures that, while awakening dormant neurons, there is no significant alteration in the input information acquired by the agent. Consequently, it effectively enhances the overall information retention and processing capabilities throughout the DRL process. In summary, this paper introduces a method termed Focus on Important Content: Visual Attention-based Neuron Awakening and Shifting, abbreviated as FIC. The specific implementation steps of the FIC algorithm are outlined in Algorithm 1.

Algorithm 1 Focus on Important Content (FIC)

Initialization: Replay buffer B, online network parameters θ, target network parameters $\hat{\theta}$.
Define a network structure that allows for cyclic utilization of neurons and layers.
Training process:
for each episode **do**
  for each time step **do**
    Generate the DVAM for state s, then apply the DVAM: $s \leftarrow \text{DVAM} \times s$.
    Select action a using ϵ-greedy strategy based on the enhanced state DVAM(s) and online network parameters θ.
    Execute action a, observe reward r, next state s', and termination signal d.
    Store transition (s, a, r, s', d) in replay buffer B.
    Sample a batch of transitions (s, a, r, s', d) from B.
    Generate DVAMs for sampled states s and next states s', and apply: $s \leftarrow \text{DVAM} \times s$, $s' \leftarrow \text{DVAM} \times s'$.
    Preprocess the enhanced states DVAM(s) and DVAM(s').
    Compute the target values for the batch using target network parameters $\hat{\theta}$.
    Update online network parameters θ using the loss function with $Q(\text{DVAM}(s), a; \theta)$ and the target values.
    Update $\hat{\theta}$ with θ at a set frequency.
  end for
  Check for dormant neurons in the feature extractor and Rainbow Q network, and reset them.
  Check positions masked by DVAM(s), reset neurons causing dormancy to Xavier uniform distribution.
  Adjust ϵ-greedy strategy as needed to optimize exploration process.
end for

6. Experiment

In order to verify whether our method can achieve visual saliency weighting, we will conduct experiments in this section to validate our idea.

6.1. Experimental design

In our experiments across 26 ALE games, the agent observes continuous sequences of 4 frames, which are transformed to a scale of 84×84 and subjected to grayscale processing. The experiments are conducted under the Atari 100K mode, indicating learning over 100K iterations spanning 400K frames. The action repetition for all games is set to 4, and the frame stack is established at 4. We compare our FIC method against several baselines, including model-free methods such as DER (Van Hasselt et al., 2019), DrQ(ϵ) (Agarwal et al., 2021), IRIS (Micheli et al., 2023), SPR (Schwarzer et al., 2020), SR-SPR (Nikishin et al., 2022), BBF (Schwarzer et al., 2023), and the lookahead search method EfficientZero (Ye et al., 2021). The Rainbow Network serves as the agent's value network, employing the Nature CNN as its core architecture. The detailed experimental setup, including all relevant parameters, is provided in Table 2(Nikishin et al., 2022). These parameters, including the discount factor γ, number of atoms, learning rate, batch size, and other critical settings, were chosen based on prior research and tuned to ensure consistency across experiments.

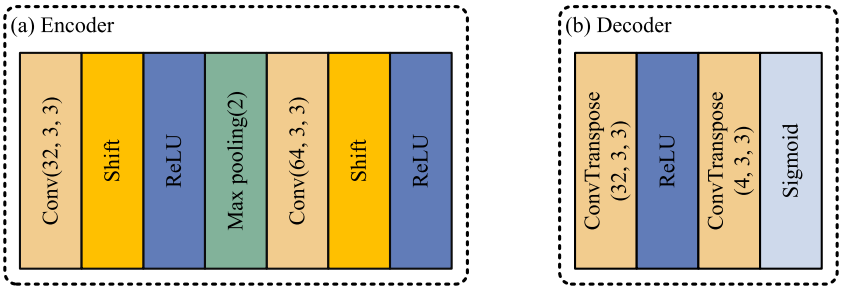


Fig. 6. VA structure. The structure of the encoder and decoder are enhanced by introducing contextual aware information from feature maps.

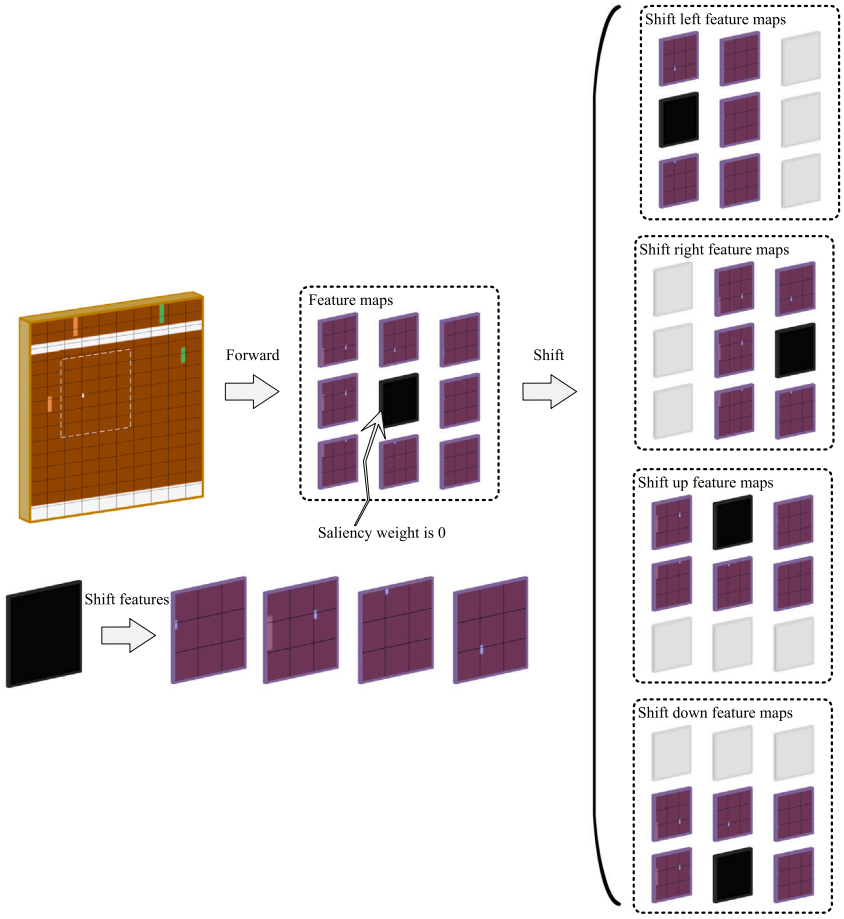


Fig. 7. Pixel-wise shift operation for feature representation. This operation allows the model to capture more robust feature representations. Even if a specific pixel's value causes a neuron to be inactive, the values of surrounding pixels can offer complementary information, helping to reactivate that neuron.

Table 2
Experiment settings.

Parameter	γ	Atoms	n-step	Target update	Learning rate	Batch size	Iteration	RR	Dormant Neuron threshold	Epsilon
Value	0.99	51	3	8000	0.0000625	32	15	2	0.1	0.01

6.2. Experimental result

6.2.1. Human-normalized score

We first compare our method with the baseline method using the most intuitive method, and the results are shown in Table 3.

It is evident that our method achieved the best results in 14 out of 26 games. Our approach secured the top rank in the Intrinsic Quality Metric (IQM) across all methods and obtained the highest Median score among model-free methods, just slightly below the model-based EfficientZero. In terms of mean results, our method secured the third position. Overall, our approach outperforms the baselines in the comprehensive score. However, it is important to note that FIC does not always outperform other methods in every game. In some games, particularly in environments that require precise prediction of future

Table 3
Comparison of our method and baselines on 26 Atari 100K games. The best result in each row is marked in red, and the second best result is marked in blue.

Games	Random	Human	DER	DrQ(E)	SPR	IRIS	SR-SPR	EfficientZero	BBF	FIC
Alien	227.8	7127.7	802.3	865.2	841.9	420	1107.8	808.5	1173.2	1295.68
Amidar	5.8	1719.5	125.9	137.8	179.7	143	203.4	148.6	244.6	584.5
Assault	222.4	742	561.5	579.6	565.6	1524.4	1088.9	1263.1	2098.5	1532.07
Asterix	210	8503.3	535.4	763.6	962.5	853.6	903.1	25557.8	3946.1	3873.48
BankHeist	14.2	753.1	185.5	232.9	345.4	53.1	531.7	351	732.9	1086.07
BattleZone	2360	37 187.51	8977	10 165.3	14 834.1	13 074	17 671	13 871.2	24459.8	27085.53
Boxing	0.1	12.1	−0.3	9	35.7	70.1	45.8	52.7	85.8	86.96
Breakout	1.7	30.5	9.2	19.8	19.6	83.7	25.5	414.1	370.6	39.88
ChopperCommand	811	7387.8	925.9	844.6	946.3	1565	2362.1	1117.3	7549.3	3757.08
CrazyClimber	10 780.5	35 829.4	34 508.6	21 539	36 700.5	59 324.2	45 544.1	83940.2	58 431.8	110034.92
DemonAttack	152.1	1971	627.6	1321.5	517.6	2034.4	2814.4	13003.9	13341.4	2563.98
Freeway	0	29.6	20.9	20.3	19.3	31.1	25.4	21.8	25.5	33.32
Frostbite	65.2	4334.7	871	1014.2	1170.7	259.1	2584.8	296.3	2384.8	3905.09
Gopher	257.6	2412.5	467	621.6	660.6	2236.1	712.4	3260.3	1331.2	2603.18
Hero	1027	30 826.4	6226	4167.9	5858.6	7037.4	8524	9315.9	7818.6	28087.19
Jamesbond	29	302.8	275.7	349.1	366.5	462.7	389.1	517	1129.6	857.89
Kangaroo	52	3035	581.7	1088.4	3617.4	838.2	3631.7	724.1	6614.7	8309.91
Krull	1598	2665.5	3256.9	4402.1	3681.6	6616.4	5911.8	5663.3	8223.4	6276.67
KungFuMaster	258.5	22 736.3	6580.1	11 467.41	14 783.2	21759.8	18 649.4	30944.8	18 991.7	21 007.22
MsPacman	307.3	6951.6	1187.4	1218.1	1318.4	999.1	1574.1	1281.2	2008.3	2337.16
Pong	−20.7	14.6	−9.7	−9.1	−5.4	14.6	2.9	20.1	16.7	19.72
PrivateEye	24.9	69 571.3	72.8	3.5	86	100	97.9	96.7	40.5	321.23
Qbert	163.9	13 455	1773.5	1810.7	866.3	745.7	4044.1	14448.5	4447.1	13673.74
Roadrunner	11.5	7845	11 843.4	11 211.4	12 213.1	9614.6	13 463.4	17 751.3	33426.8	40221.33
Seaquest	68.4	42 054.7	304.6	352.3	558.1	661.3	819	1100.2	1232.5	7667.16
UpNDown	533.4	11 693.2	3075	4324.5	10 859.2	3546.2	112450.3	17264.2	12 101.7	6222.48
Games > Human	0	0	2	3	6	9	9	14	12	14
IQM (t)	0	1	0.183	0.28	0.337	0.501	0.631	1.02	1.045	1.21
Median	0	1	0.189	0.313	0.396	0.289	0.685	1.116	0.917	1.05
Mean	0	1	0.35	0.465	0.616	1.046	1.272	1.945	2.247	1.67

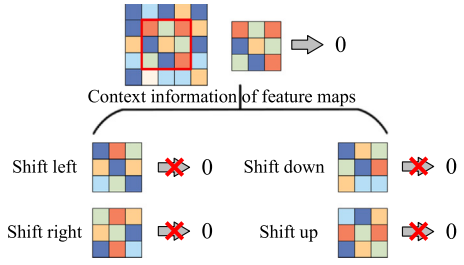


Fig. 8. Dynamic shift operation. By performing the shift operation on feature maps of different channels, even if the output of the shifted feature map becomes 0, it will still yield a non-zero output after the dynamic input changes.

states, such as Breakout and DemonAttack, FIC's ranking was lower than expected. This can be attributed to FIC's lack of state prediction capabilities, which are present in methods like world models and SPR. In games like Breakout, where the agent needs to predict the trajectory of moving objects (such as the ball), the world model-based methods can better predict future states and adjust actions accordingly. As a result, these methods achieve more accurate decisions, leading to better performance in such scenarios. FIC, lacking this predictive mechanism, is less effective in environments requiring precise long-term state forecasting. This limitation may explain its performance in certain games where future state prediction is crucial for optimal decision-making.

This validates the effectiveness of our method in extracting visually salient weights relevant to the task, thereby reducing redundant information in visual input, and enhancing the sample efficiency of VDRL. To further investigate the efficiency performance of our method, we present the scores during the training process for the 26 games, as shown in Fig. 9.

From Fig. 9, it is evident that all games achieve higher scores with increasing training steps. Unlike the scenario depicted in Fig. 3, where the introduction of visual saliency in the early stages led to the blocking of critical information and resulted in training failure, no such phenomenon is observed here.

6.2.2. Algorithm efficiency

Additionally, we aim to understand how many samples our method requires to achieve human-level's IQM performance. With exploration over 100K iterations, we conduct a total of 16 iterations. Therefore, we present the results at 16 checkpoints, as shown in Fig. 10.

From the graph, it is evident that the IQM metric first reaches human-level performance at Step=9 and maintains at human-level performance from Step ≥ 11 onwards. This validates our initial hypothesis that the agent's exploration efficiency improves when it can effectively reduce irrelevant information to the task.

6.2.3. DVAM results

In this study, we explore the VA mechanisms employed by the agent while processing various Atari games. To achieve this, we adopt a comprehensive visualization strategy, providing detailed representations of DVAMs for all games. Given that the agent receives input comprising consecutive four-frame images, Figs. 11 and 12 showcase the DVAMs observed by the agent across consecutive four frames, offering an intuitive depiction of how the agent processes visual information over time.

Through the analysis of these DVAMs, we observe that the agent dynamically adjusts its attention to different elements in various game scenarios based on the current task requirements. For instance, in Demon Attack, as shown in Fig. 11 (specifically in the row corresponding to DemonAttack), the agent not only distinguishes between the game background and the approaching enemy spaceships but also enhances saliency scores for the spaceships that are directly relevant to the "attack" task. This dynamic allocation of attention reflects the agent's ability to prioritize critical objects during task execution.

Similarly, in Pong, as depicted in Fig. 12, the agent demonstrates heightened awareness of the position of the ping-pong ball and paddle. Notably, in the frames where the ball is in close proximity to the paddle, the agent shows a marked increase in attention to these key elements,

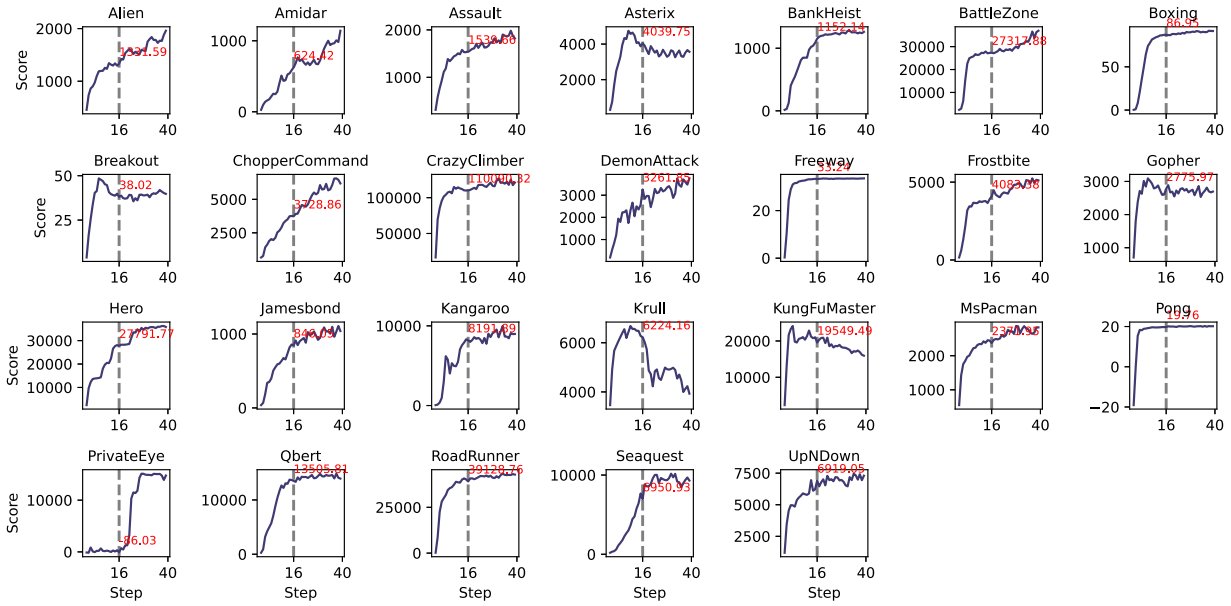


Fig. 9. Relationship between 26 game iterations and rewards. Step=16 represents performance at 100K, and Step=40 represents performance at 250K.

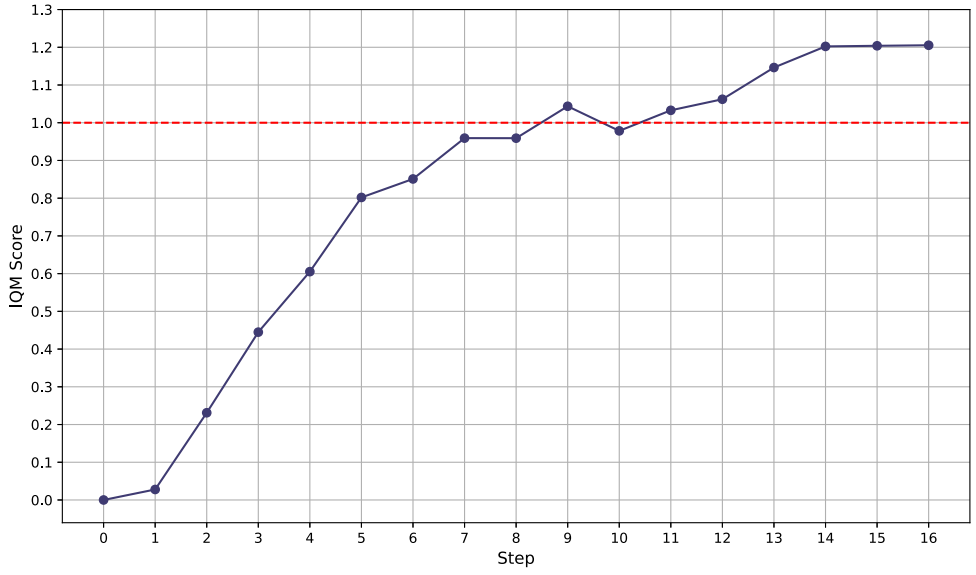


Fig. 10. Human-normalized score. Line graph of Human-normalized score regarding IQM during training process.

underscoring the agent's ability to focus on essential game features during decision-making moments.

It is worth emphasizing that these DVAMs are generated using an unsupervised learning approach. In a completely unlabeled setting, FIC is trained in a task-driven manner. This demonstrates the agent's capability to effectively filter out irrelevant frame content within consecutive four-frame inputs and concentrate attention on crucial information in frames relevant to the task. This finding not only deepens our understanding of the agent's VA mechanisms but also provides essential experimental evidence for the future development of more efficient and task-adaptive artificial intelligence systems.

6.3. Ablation experiment

In pursuit of understanding the key contributor to the substantial performance improvement, we conduct ablation experiments. Within these experiments, attention is introduced atop the Rainbow baseline, and subsequently, we separately integrate the methods of activating neurons and shifting within this structure. Our objective is to ascertain whether activating neurons effectively improve performance or if enhancements arise from the introduction of context-aware mechanisms.

The results are depicted in Table 4. It is evident that after introducing attention, the IQM of 0.13 is inferior to DER's 0.18. In certain games like Boxing, Breakout, and CrazyClimber, the performance is only marginally better than a completely random control strategy. This implies that the agent has not acquired effective strategies. We visualize the DVAMs for these three games, as shown in Figs. 11 and 12.

We visualize 16 frames under continuous control in a sequence of four frames during the process. It is conceivable that early DVAMs block essential information patterns, causing the trained agent to miss acquiring visual information from these segments.

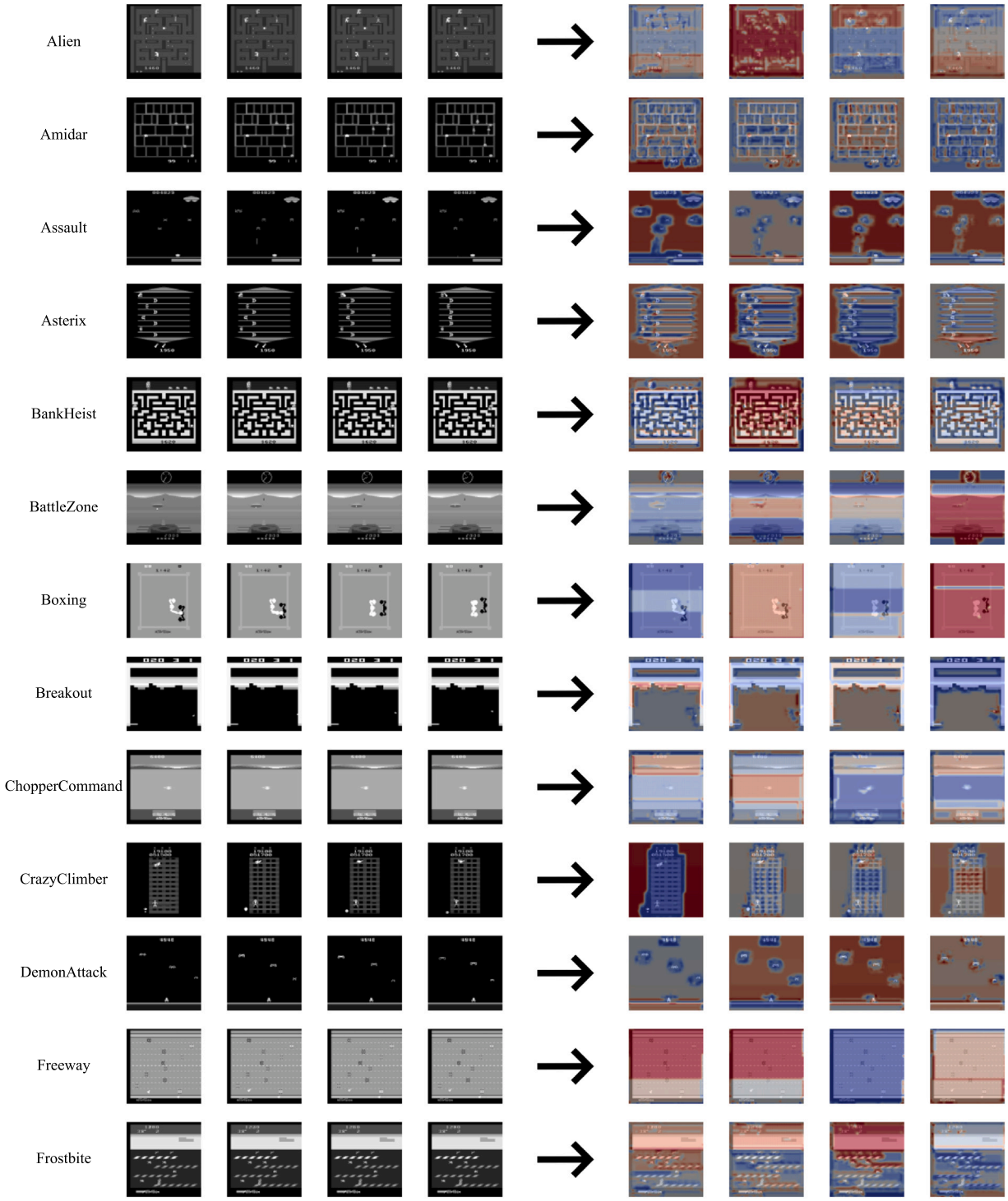


Fig. 11. DVAMs results. The DVAMs of Atari game obtained by agent, with each game displaying four consecutive frames of raw images and DVAMs (the first 13 games).

Furthermore, we introduce context awareness to determine whether this contributes to the substantial performance boost. However, it is evident that this only results in a marginal increase of 0.01 IQM. In the Freeway game, the score even remains at a complete 0, indicating that context awareness does not resolve the issue of the early disappearance of DVAMs.

Upon introducing the mechanism to awake dormant neurons, IQM reaches 0.96, and the agent surpasses human-level performance in 13 out of 26 games. This validates our hypothesis that by awaking dormant neurons in DVAMs, early-blocked patterns can be restored to an unblocked state.

Finally, under the mechanism of awaking dormant neurons with the introduction of the shift, we track 16 checkpoints during the 100K process, as depicted in Fig. 13. It is evident that introducing the shift mechanism effectively provides contextual aware information to the randomly awake dormant neurons, offering certain prior knowledge. This prevents downstream tasks in DRL from encountering a relatively unfamiliar visual input, thus accelerating the training effectiveness.

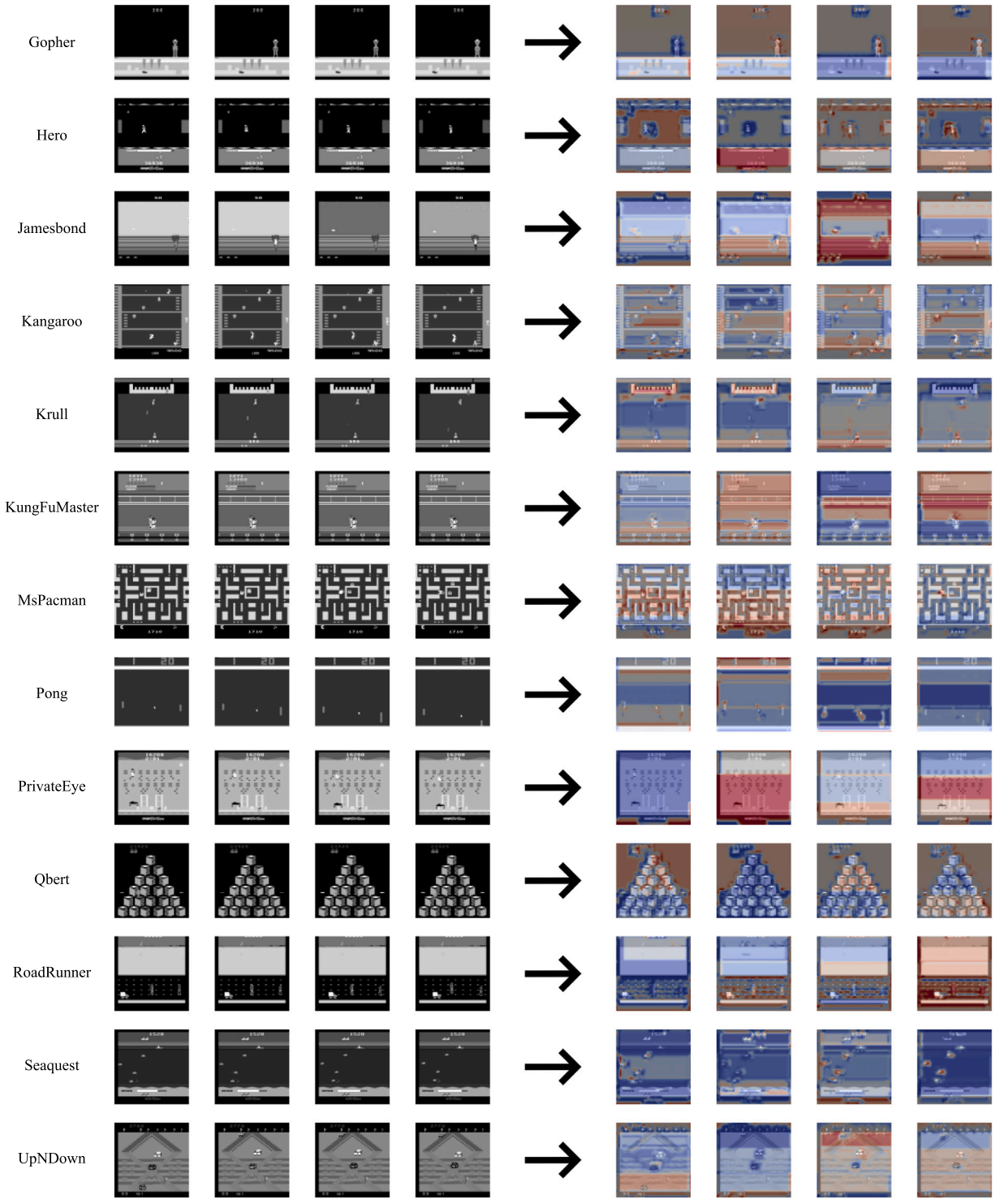


Fig. 12. DVAMs results. The DVAMs of Atari game obtained by agent, with each game displaying four consecutive frames of raw images and DVAMs (the last 13 games).

Throughout the 100K tasks, the model consistently outperforms the one without the shift mechanism.

6.4. The effect of shift on improving dormant neurons in dynamic tasks

To assess the impact of the shift operation on reducing dormant neurons within the attention mechanism in dynamic task environments, we conduct a statistical analysis comparing the proportion of dormant neurons introduced during training with and without the shift operation. Fig. 14 illustrates the overall mean and variance, clearly indicating a significant reduction in dormant neurons with the shift operation. Specifically, at the conclusion of training, the proportion of dormant neurons decreased from 26.60% to 20.38%.

In Fig. 15, we further demonstrate the specific performance of 26 games after introducing shift operations, with a significant decrease in the number of dormant neurons for each game.

This result not only confirms that the shift operation effectively reduces dormant neurons, thereby enhancing the adaptability and processing efficiency of neural networks for dynamic tasks but also underscores its universality and effectiveness across different gaming

Table 4
Results of ablation experiments.

Method	FIC's ablation experiments					Random
	Activate	✓	✓	✗	✗	
	Shift	✓	✗	✓	✗	
Game scores	Alien	1295.68	1205.9849	532.5	556.2	227.8
	Amidar	584.5	262.234	71.96	120.13	5.8
	Assault	1532.07	1109.5532	307.86	307.23	222.4
	Asterix	3873.48	3095.5444	364	398.0	210
	BankHeist	1086.07	1000.1613	102.3	105.4	14.2
	BattleZone	27 085.5	2301.2048	7470	7110.0	2360
	Boxing	86.96	60.9778	3.4	1.71	0.1
	Breakout	39.88	48.809	11.35	6.0	1.7
	ChopperCommand	3757.08	3123.1182	950	1021.0	811
	CrazyClimber	110 035	110 044	15 478	14 042.0	10 780.5
	DemonAttack	2563.98	2167.7659	556.35	1035.6	152.1
	Freeway	33.32	32.2857	0	19.53	0
	Frostbite	3905.09	3391.7021	235.2	211.542	65.2
	Gopher	2603.18	2424.6155	636.8	638.1007	257.6
	Hero	28 087.2	13 952.788	3385.15	7432.2998	1027
	Jamesbond	857.89	476.8116	109	56.0	29
	Kangaroo	8309.91	4667.2412	216	222.0	52
	Krull	6276.67	5962.5928	3378.4	4442.7002	1598
	KungFuMaster	21 007.2	19 868.75	8752	5748.0	258.5
	MsPacman	2337.16	2267.9246	749.5	811.1	307.3
	Pong	19.72	18.1702	−5.87	−3.5	−20.7
	PrivateEye	321.23	73.2368	100	44.0	24.9
	Qbert	13 673.7	8647.3682	656.25	383.25	163.9
	Roadrunner	40 221.3	36 770.832	5277	3151.0	11.5
	Seaquest	7667.16	4015	294	221.6	68.4
	UpNDown	6222.48	4854.6377	2341.8	2714.1001	533.4
Human levels	Games > Human	14	13	1	1.0	0
	IQM	1.21	0.960086	0.14167	0.131094424	0
	Median	1.05	0.9390249	0.13298	0.132185993	0
	Mean	1.67	1.3258769	0.21634	0.262358571	0

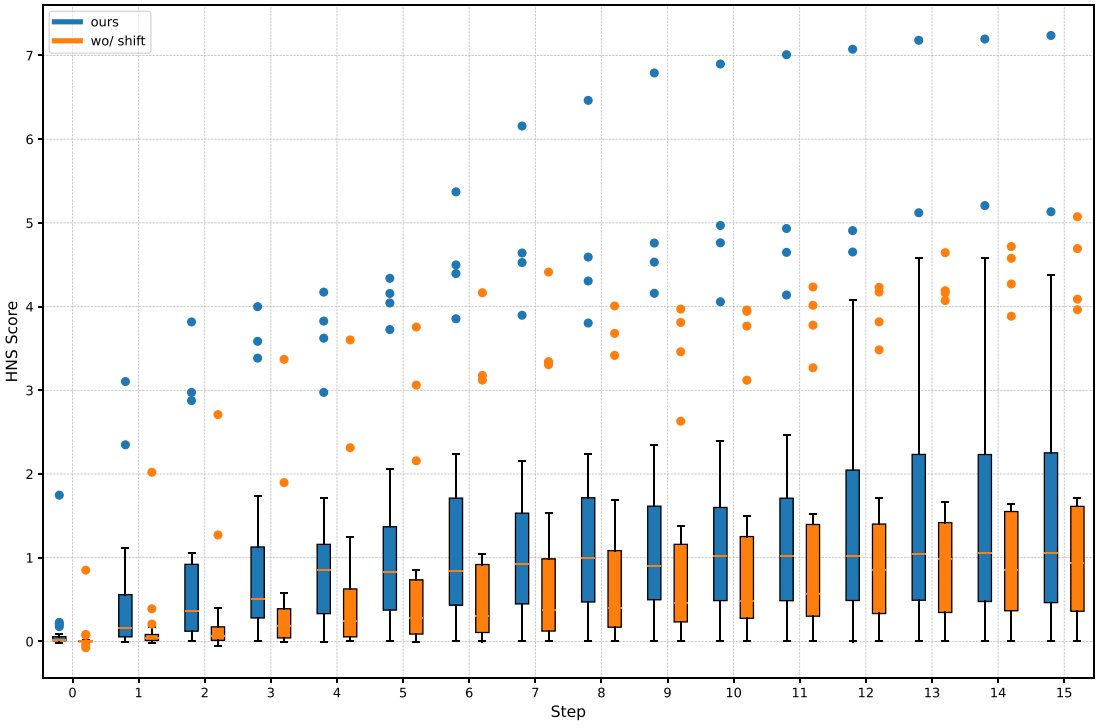


Fig. 13. IQM performance with and without shift under different steps. The x-axis denotes the step, and the y-axis indicates the Human-normalized score (HNS).

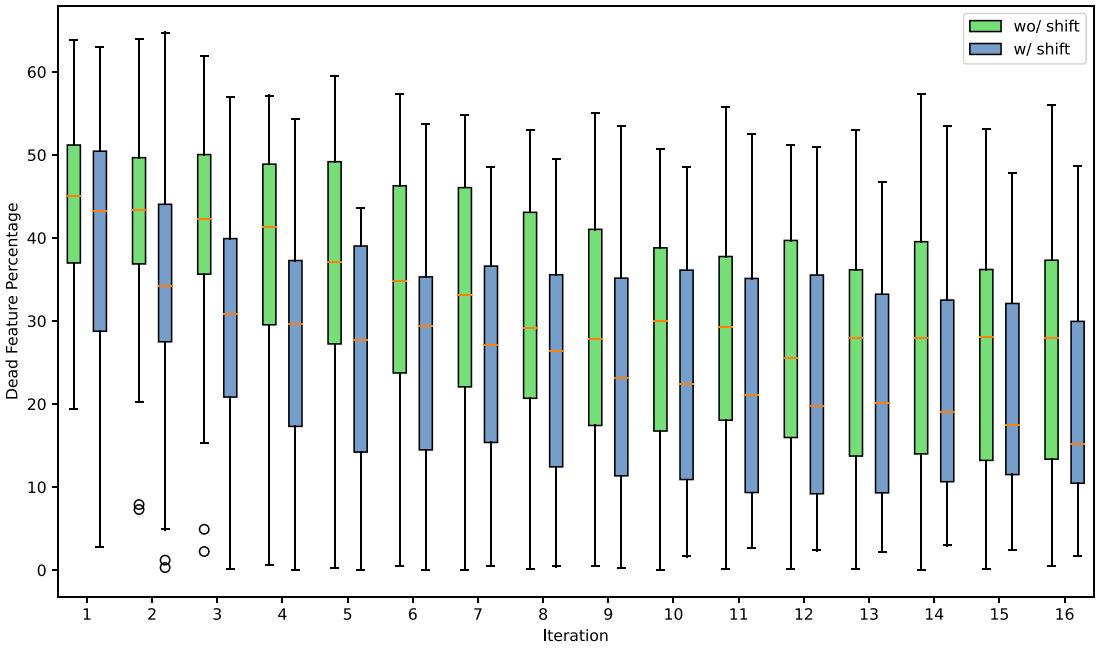


Fig. 14. Percentage of dormant neurons. This plot shows the percentage of dormant neurons (Dead Feature Percentage) across training iterations, comparing scenarios with and without the shift operation. The x-axis indicates the iteration number, and the y-axis shows the percentage of dormant neurons. The box plot summarizes the distribution, highlighting the shift operation's role in reducing dormant neurons.

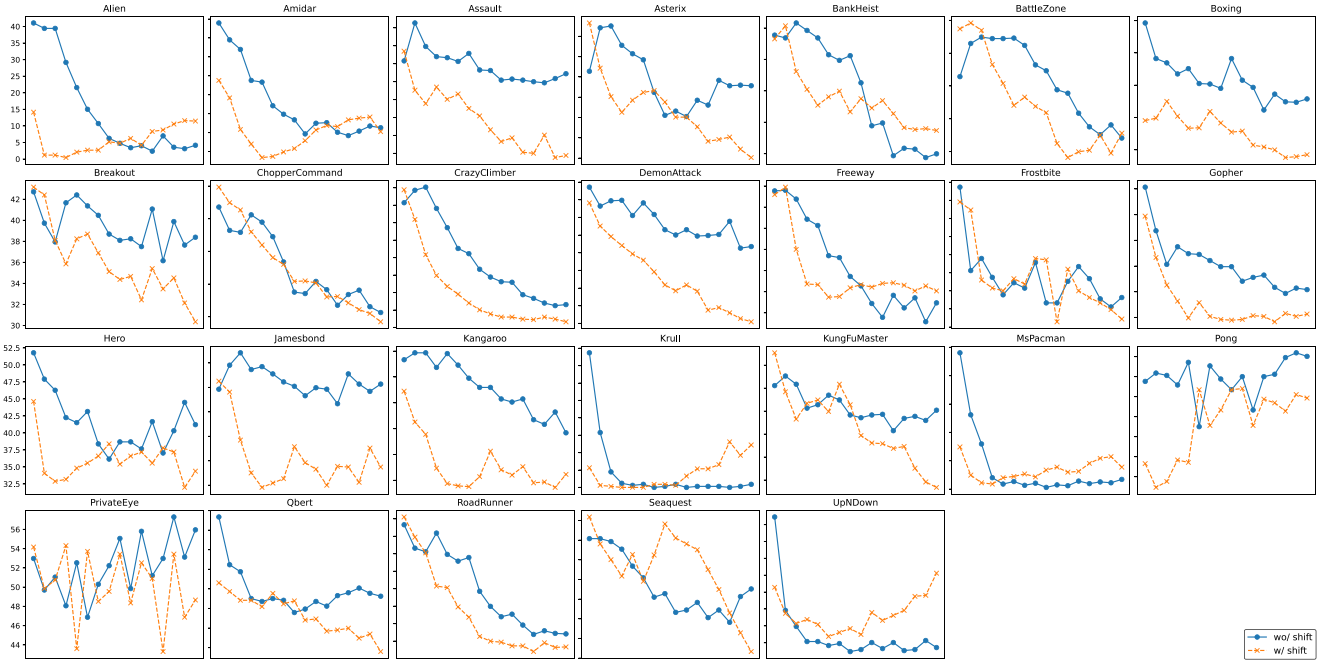


Fig. 15. Dormant neuron proportion changes across 26 games with and without the shift operation. This figure shows how dormant neurons can overlook task-relevant features, emphasizing the shift operation's effectiveness in reducing them. Each subplot represents a game, displaying dormant neuron proportions over training iterations before and after the shift. The x-axis indicates iteration count, and the y-axis shows dormant neuron proportion. The graphs demonstrate that the shift operation significantly lowers dormant neuron proportions in most games, improving neural network adaptability.

tasks. Through the comparative analysis of Figs. 14 and 15, we gain a profound understanding of the crucial role played by the shift operation in strengthening neural network performance.

6.5. Evaluating FIC in other environments

6.5.1. Environments

To assess the adaptability and performance of the FIC method in more complex environments, we conduct experiments using the Car Racing environment. This environment involves dynamic and continuous action spaces, providing a suitable challenge for evaluating DRL. The observation space is composed of RGB images with a resolution of $96 \times 96 \times 3$, where each pixel has values ranging from 0 to 255. To capture the motion characteristics of the task, we utilize a stack of four consecutive frames as the input to the agent. The action space is discrete, consisting of five possible actions: The action space in the Car Racing environment is discrete and consists of five possible actions available to the agent. The first action (Action 0) involves no operation,

Table 5
Hyperparameters.

No.	Parameter	Value
1	learning rate	1×10^{-4}
2	buffer size	100 000
3	learning starts	50 000
4	batch size	32
5	τ	1.0
6	γ	0.99
7	train freq	4
8	gradient steps	1
9	target update interval	10 000
10	exploration fraction	0.1
11	exploration initial eps	1.0
12	exploration final eps	0.01

Table 6
The game source of FIC.

No.	VA	Activate	Shift	200K	600K	1M
1	✗	✗	✗	31.12	620.6	859.4
2	✓	✗	✗	−31.42	297.5	656.3
3	✓	✓	✗	−26.25	599.2	795.9
4	✓	✓	✓	11.26	707.5	918.4

meaning the agent does nothing. Action 1 corresponds to steering the vehicle to the left, while Action 2 involves steering the vehicle to the right. Action 3 is the acceleration action, causing the agent to increase the vehicle's speed. Finally, Action 4 corresponds to braking, where the agent reduces the vehicle's speed. These discrete actions provide the agent with the ability to control the vehicle's movement in the simulation.

The agent receives a fixed penalty of -0.1 for each frame, encouraging faster completion of the task. Additionally, the agent earns a reward of $+1000/N$ for each new track tile visited, where N is the total number of tiles visited during the episode. For example, if the agent completes the task in 732 frames, the total reward would be calculated as $1000 - 0.1 \times 732 = 926.8$. The episode terminates when all tiles have been visited or when the car drives too far off the track, resulting in a penalty of -100 and the immediate termination of the episode.

6.5.2. Experiment setup

The experiment utilizes the DQN (Mnih et al., 2013) from the Stable-Baselines3 (Raffin et al., 2021b), with the hyperparameters summarized in Table 5. The meaning of the parameters is consistent with the definition of Stable-Baselines3. Training is conducted on an Nvidia RTX 4090 device to accelerate the computation and ensure efficiency.

6.5.3. Performance evaluation

In addition to the parameters listed above, the rest of the experiment setup remains consistent with the FIC method. The performance results of the method, with and without the VA mechanism, are shown in Table 4. The performance results of FIC is shown in Table 6.

As shown in Table 6, the basic introduction of the VA module in DQN yielded the worst performance. However, this did not lead to training failure, as DQN performs a limited number of training steps per frame, preventing premature loss of visual information. With the introduction of the activation mechanism, although training difficulty increased, performance improved compared to the simple VA integration, particularly at 600K iterations where the score increased by over 300 points.

The shift mechanism in FIC addresses the loss of weights in some channels after resets. The highest performance was observed after 200K iterations, suggesting that the active mechanism is beneficial when applied to Visual DQN, particularly when combined with the shift mechanism. This indicates that the method can focus on task-relevant visual information while mitigating issues related to attention loss and weight degradation after resets.

6.6. Discussion and analysis

The experimental results of this study present a comprehensive comparison between our proposed method and state-of-the-art baseline models across multiple Atari 100K games. Through these comparisons, our approach demonstrates significant performance improvements in most games, particularly excelling in those with higher demands for VA. This section explores the underlying reasons behind these results and their implications for future research directions.

Our agent dynamically adjusts its VA based on the current task's requirements, allowing flexibility in handling visual information in various game scenarios, thereby enhancing adaptability and generalization. By reducing attention to task-irrelevant information, our method effectively diminishes redundant information in visual input, improving learning efficiency.

To further analyze the reduction of redundant information in decision-making tasks, we conduct a statistical analysis of the visual information quantity in DVAMs for each game. Specifically, we measure the information content of DVAMs. The results, depicted in Fig. 16, illustrate the information quantity of DVAMs across 26 games.

We observe remarkable effectiveness in DVAMs. In the Seaquest game, with the lowest information content, DVAMs retain only 13% of the original image's information. Even with this decrease, the game attained an impressive score of 7667.16 points, surpassing the second-ranking game by 5.22 times, which scored 1232.5 points. This outcome underscores that minimizing task-irrelevant redundant information in input images significantly boosts the training efficiency of the agent.

In addition to its application in visual-based environments such as Atari 100K, the model can also be applied to non-visual industrial optimization tasks, such as flexible job shop scheduling, vehicle routing problems, and capacitated vehicle routing problems. In these scenarios, although there is no visual input, Rainbow DQN can serve as the actor component of the agent. The dormant neuron reset mechanism, a key feature of our approach, can be applied to these combinatorial optimization problems to reduce the proportion of dormant neurons in the attention model, thereby potentially improving the agent's exploration efficiency. This helps ensure the agent avoids getting stuck in suboptimal solutions, leading to better decision-making and faster convergence in such task-driven settings.

7. Conclusion

This study introduces the FIC method, which significantly improves sample efficiency and performance in VDRL. By integrating the awakening of dormant neurons and a strategy for shifting information between windows, FIC demonstrates substantial performance enhancements in the Atari 100K game test, particularly excelling in games with high VA requirements. It outperforms existing baseline models not only in performance but also in sample efficiency. Theoretically, our work conducts a thorough analysis of the limitations associated with applying VA in DRL, guiding future research directions. On a practical note, our method boosts decision quality and efficiency by minimizing interference from task-irrelevant visual information, offering a practical tool for developing task-oriented visual-based RL systems.

While our approach shows clear superiority, it is important to acknowledge several limitations and challenges. One key limitation is the potential sensitivity of the FIC method to inaccurate attention models. Since VA mechanisms rely on distinguishing relevant visual features, errors in attention prediction may lead to suboptimal decision-making, particularly in environments where visual saliency is hard to define. Additionally, while the information-shifting strategy has shown promising results, its general applicability to a wider variety of games and environments, especially those with sparse or highly dynamic visual information, needs further validation. Furthermore, although awakening dormant neurons helps improve the model's sensitivity to

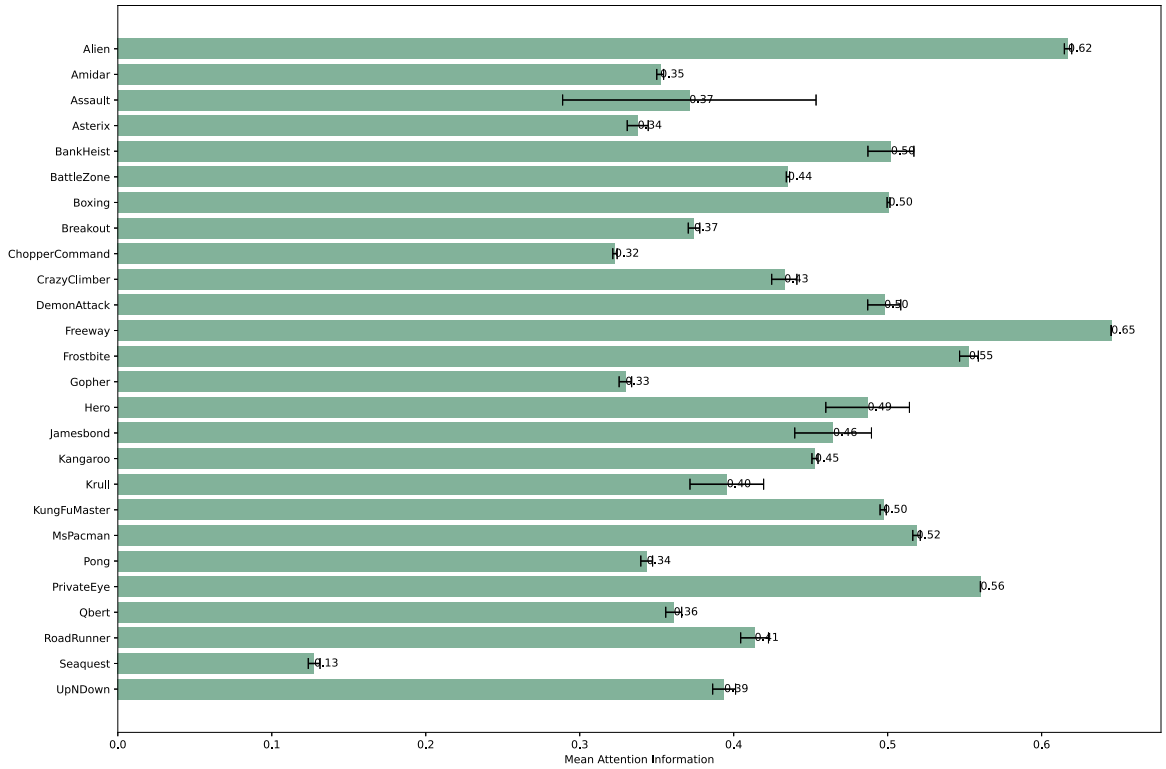


Fig. 16. Information content results of DVAMs for 26 games.

crucial visual information, fine-tuning the awakening process to avoid noise from excessive activation remains a challenge.

To address these limitations, future research will focus on two primary directions. First, we aim to develop more robust and adaptive mechanisms for VA, enhancing their adaptability to a broader range of environments and tasks, and improving the generalization capabilities of the method. Second, we will deepen the theoretical understanding of VA mechanisms in RL, conducting further experimental validation to provide a solid foundation for the design of more efficient learning algorithms.

CRediT authorship contribution statement

Jialin Ma: Writing – original draft, Methodology, Data curation. **Ce Li:** Writing – review & editing, Methodology. **Zhiqiang Feng:** Writing – review & editing. **Limei Xiao:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The work of this thesis is partially funded by the Science and Technology on Vacuum & Cryogenics Technology and Physics Laboratory (No. 61422072305).

Data availability

No data was used for the research described in the article.

References

Agarwal, R., Schwarzer, M., Castro, P.S., Courville, A.C., Bellemare, M., 2021. Deep reinforcement learning at the edge of the statistical precipice. In: Advances in Neural Information Processing Systems, vol. 34, pp. 29304–29320.

Badia, A.P., Sprechmann, P., Vitvitskyi, A., Guo, D., Piot, B., Kapturowski, S., Tieleman, O., Arjovsky, M., Pritzel, A., Bolt, A., 2020. Never give up: Learning directed exploration strategies. In: International Conference on Learning Representations.

Baxter, L., 1995. Markov decision processes: Discrete stochastic dynamic programming. Technometrics 37, 353–353.

Bellemare, M., Dabney, W., Munos, R., 2017. A distributional perspective on reinforcement learning. ArXiv Preprint, arXiv:1707.06887.

Bellemare, M., Naddaf, Y., Veness, J., Bowling, M., 2012. The arcade learning environment: An evaluation platform for general agents. J. Artificial Intelligence Res. 47.

Berner, C., Brockman, G., Chan, B., Cheung, V., Dębiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., 2019. Dota 2 with large scale deep reinforcement learning. ArXiv Preprint, arXiv:1912.06680.

Broadbent, D.E., 2013. Perception and Communication. Elsevier.

Carr, T., Chli, M., Vogiatzis, G., 2018. Domain adaptation for reinforcement learning on the atari. arXiv preprint arXiv:1812.07452.

Chaturvedi, R., Verma, S., 2023. Opportunities and challenges of AI-driven customer service. Artif. Intell. Cust. Serv.: Next Front. Pers. Engag. 33–71.

Chen, X., He, K., 2021. Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758.

Cunningham, P., Cord, M., Delany, S.J., 2008. Supervised learning. In: Machine Learning Techniques for Multimedia: Case Studies on Organization and Retrieval. Springer, pp. 21–49.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., 2022. Magnetic control of tokamak plasmas through deep reinforcement learning. Nature 602 (7897), 414–419.

Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., Koltun, V., 2017. CARLA: An open urban driving simulator. In: Conference on Robot Learning. PMLR, pp. 1–16.

Feng, S., Sun, H., Yan, X., Zhu, H., Zou, Z., Shen, S., Liu, H.X., 2023. Dense reinforcement learning for safety validation of autonomous vehicles. Nature 615 (7953), 620–627.

Fortunato, M., Azar, M.G., Piot, B., Menick, J., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C., Legg, S., 2017. Noisy networks for exploration. p. 1706.10295, CoRR.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings, pp. 249–256.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.
Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., 2020. Bootstrap your own latent-a new approach to self-supervised learning. In: Advances in Neural Information Processing Systems, vol. 33, pp. 21271–21284.
Guo, S., Zhang, R., Liu, B., Zhu, Y., Ballard, D., Hayhoe, M., Stone, P., 2021. Machine versus human attention in deep reinforcement learning tasks. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (Eds.), In: Advances in Neural Information Processing Systems, vol. 34, pp. 25370–25385.
Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., Horgan, D., Piot, B., Azar, M., Silver, D., 2017. Rainbow: Combining improvements in deep reinforcement learning. Proc. AAAI Conf. Artif. Intell. 32.
Itaya, H., Hirakawa, T., Yamashita, T., Fujiyoshi, H., Sugiura, K., 2021. Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning. In: International Joint Conference on Neural Networks. pp. 1–10.
Ju, H., Juan, R., Gomez, R., Nakamura, K., Li, G., 2022. Transferring policy of deep reinforcement learning from simulation to reality for robotics. Nat. Mach. Intell. 4 (12), 1077–1087.
Kaelbling, L.P., Littman, M.L., Moore, A.W., 1996. Reinforcement learning: A survey. J. Artificial Intelligence Res. 4, 237–285.
Kaiser, Ł., Babaeizadeh, M., Miłos, P., Osiński, B., Campbell, R.H., Czechowski, K., Erhan, D., Finn, C., Kozakowski, P., Levine, S., Mohiuddin, A., Sepassi, R., Tucker, G., Michalewski, H., 2020. Model based reinforcement learning for atari. In: International Conference on Learning Representations.
Kaufmann, E., Bauersfeld, L., Loquercio, A., Müller, M., Koltun, V., Scaramuzza, D., 2023. Champion-level drone racing using deep reinforcement learning. Nature 620 (7976), 982–987.
Konda, V., Tsitsiklis, J., 1999. Actor-critic algorithms. In: Advances in Neural Information Processing Systems, vol. 12.
Kostrikov, I., Yarats, D., Fergus, R., 2020. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. ArXiv Preprint, arXiv:2004.13649.
Laskar, M.T.R., Bari, M.S., Rahman, M., Bhuiyan, M.A.H., Joty, S., Huang, J.X., 2023. A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. arXiv preprint arXiv:2305.18486.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Leng, J., Mo, M., Zhou, Y., Ye, Y., Gao, C., Gao, X., 2023. Recent advances in drone-view object detection. J. Image Graph. 28 (09), 2563–2586.
Ma, J., Li, C., Feng, Z., Xiao, L., He, C., Zhang, Y., 2025. Don't overlook any detail: Data-efficient reinforcement learning with visual attention. Knowl.-Based Syst. 310, 112869.
Merikhipour, M., Khanmohammadidoustani, S., Abbasi, M., 2025. Transportation mode detection through spatial attention-based transductive long short-term memory and off-policy feature selection. Expert Syst. Appl. 267, 126196.
Micheli, V., Alonso, E., Fleuret, F., 2023. Transformers are sample-efficient world models. In: International Conference on Learning Representations.
Mnih, V., Heess, N., Graves, A., 2014. Recurrent models of visual attention. In: Advances in Neural Information Processing Systems, vol. 27.
Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M., 2013. Playing atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.
Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A., Veness, J., Bellemare, M., Graves, A., Riedmiller, M., Fidjeland, A., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., Hassabis, D., 2015. Human-level control through deep reinforcement learning. Nature 518, 529–533.
Mott, A., Zoran, D., Chrzanowski, M., Wierstra, D., Jimenez Rezende, D., 2019. Towards interpretable reinforcement learning using attention augmented agents. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'AlchBuc, F., Fox, E., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 32.
Nikishin, E., Schwarzer, M., D'Oro, P., Bacon, P.-L., Courville, A., 2022. The primacy bias in deep reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 16828–16847.
Nikulin, D., Ianina, A., Aliev, V., Nikolenko, S., 2019. Free-lunch saliency via attention in atari agents. In: International Conference on Computer Vision. pp. 4240–4249.
Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N., 2021a. Stable-baselines3: Reliable reinforcement learning implementations. J. Mach. Learn. Res. 22 (1), 12348–12355.
Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N., 2021b. Stable-Baselines3: Reliable reinforcement learning implementations. J. Mach. Learn. Res. 22 (268), 1–8.
Schaul, T., Quan, J., Antonoglou, I., Silver, D., 2016. Prioritized experience replay.
Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., 2020. Mastering atari, go, chess and shogi by planning with a learned model. Nature 588 (7839), 604–609.
Schwarzer, M., Anand, A., Goel, R., Hjelm, R.D., Courville, A., Bachman, P., 2020. Data-efficient reinforcement learning with self-predictive representations. ArXiv Preprint, arXiv:2007.05929.
Schwarzer, M., Ceron, J.S.O., Courville, A., Bellemare, M.G., Agarwal, R., Castro, P.S., 2023. Bigger, better, faster: Human-level atari with human-level efficiency. In: International Conference on Machine Learning. PMLR, pp. 30365–30380.
Shi, W., Huang, G., Song, S., Wang, Z., Lin, T., Wu, C., 2022a. Self-supervised discovering of interpretable features for reinforcement learning. IEEE Trans. Pattern Anal. Mach. Intell. 44 (5), 2712–2724.
Shi, W., Huang, G., Song, S., Wu, C., 2022b. Temporal-spatial causal interpretations for vision-based reinforcement learning. IEEE Trans. Pattern Anal. Mach. Intell. 44 (12), 10222–10235.
Shi, Z., Wu, C., Li, C., You, Z., Wang, Q., Ma, C., 2023. Object detection techniques based on deep learning for aerial remote sensing images: a survey. J. Image Graph. 28 (09), 2616–2643.
Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529 (7587), 484–489.
Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D., Riedmiller, M., 2014. Deterministic policy gradient algorithms. In: International Conference on Machine Learning. PMLR, pp. 387–395.
Sokar, G., Agarwal, R., Castro, P.S., Evci, U., 2023a. The dormant neuron phenomenon in deep reinforcement learning. ArXiv Preprint, arXiv:2302.12902.
Sokar, G., Agarwal, R., Castro, P.S., Evci, U., 2023b. The dormant neuron phenomenon in deep reinforcement learning. In: International Conference on Machine Learning. PMLR, pp. 32145–32168.
Sorokin, I., Seleznev, A., Pavlov, M., Fedorov, A., Ignateva, A., 2015. Deep attention recurrent Q-network. ArXiv Preprint, arXiv:1512.01693.
Sutton, R., 1988. Learning to predict by the method of temporal differences. Mach. Learn. 3, 9–44.
Van Hasselt, H., Guez, A., Silver, D., 2015. Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30.
Van Hasselt, H.P., Hessel, M., Aslanides, J., 2019. When to use parametric models in reinforcement learning? In: Advances in Neural Information Processing Systems, vol. 32.
Wang, Z., Freitas, N., Lanctot, M., 2016. Dueling network architectures for deep reinforcement learning. In: International Conference on Machine Learning, vol.48, pp. 1995–2003.
Wang, X., Lian, L., Yu, S.X., 2021. Unsupervised visual attention and invariance for reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 6677–6687.
Watkins, C.J., Dayan, P., 1992. Q-learning. Mach. Learn. 8, 279–292.
Wiering, M.A., Van Otterlo, M., 2012. Reinforcement learning. In: Adaptation, Learning, and Optimization, vol. 12, p. 729.
Wu, H., Chen, M., He, C., Luo, S., et al., Multi-dimensional attention fusion network for terahertz image super-resolution. In: Meiyun, He, C., Luo, S. (Eds.), Multi-Dimensional Attention Fusion Network for Terahertz Image Super-Resolution.
Wu, H., Khetarpal, K., Precup, D., 2021. Self-supervised attention-aware reinforcement learning. Proc. AAAI Conf. Artif. Intell. 35, 10311–10319.
Ye, W., Liu, S., Kurutach, T., Abbeel, P., Gao, Y., 2021. Mastering atari games with limited data. In: Advances in Neural Information Processing Systems, vol. 34, pp. 25476–25488.
Yuan, X., Cheng, G., Li, G., Dai, W., Yin, W., Feng, Y., Yao, X., Huang, Z., Sun, X., Han, J., 2023. Progress in small object detection for remote sensing images. J. Image Graph. 28 (06), 1662–1684.

Knowledge-Based Systems 310 (2025) 112869

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Don't overlook any detail: Data-efficient reinforcement learning with visual attention

Jialin Ma [a,1], Ce Li [a,1,*], Zhiqiang Feng [a], Limei Xiao [a], Chengdan He [b,*], Yan Zhang [b]

[a] School of Electrical Engineering and Information Engineering, Lanzhou University of Technology, Lanzhou, 730050, China
[b] Science and Technology on Vacuum Technology and Physical Laboratory, Lanzhou Institute of Physics, Lanzhou, 730050, China

ARTICLE INFO

Keywords:
Visual reinforcement learning
Visual attention
Don't overlook any detail
Reset
Atari 100K

ABSTRACT

With the widespread application of visual reinforcement learning across various domains, the introduction of visual attention mechanisms aims to emulate human visual tasks, enabling deep models to focus on the crucial parts of images and enhancing model performance. However, in situations with limited data samples, solely introducing visual attention mechanisms can exacerbate overfitting in deep reinforcement learning (DRL), deteriorating performance. Herein, we propose a method called 'Don't overlook any detail (DOAD)' to tackle this issue. A two-step training strategy is proposed to increase the training frequency of the visual attention module while avoiding the specification of explicit tasks and fully acknowledging the pivotal role of visual attention in the learning process, rendering the model more adaptable to environmental changes. Furthermore, a conditional network reset method is proposed to simulate the flexibility observed in human learning processes, encouraging the model to adapt more flexibly to new information through regular reset mechanisms without excessively adhering to early knowledge. Finally, extensive experiments were conducted on 26 game environments within the Atari 100K environment. Compared to the baseline with the introduction of visual attention on the interquartile mean (IQM) from 0.44 to 0.37, the introduction of DOAD visual attention methods can improve the IQM to 0.70. DOAD elucidates the internal mechanisms of DRL and offers novel insights for applying visual attention mechanisms in DRL models under limited sample data contexts.

1. Introduction

Dynamic task-based visual attention is a critical approach that may be utilised by humans in visual decision-making tasks, where individuals allocate attention to task-relevant content as the task progresses. Despite visual deep reinforcement learning (VDRL) achieving human-level performance in some tasks [1–3], significant performance degradation occurs when dynamic visual attention mechanisms are introduced.

This phenomenon contrasts with our expectations. Introducing attention mechanisms in deep learning (DL) has shown significant effectiveness in computer vision and natural language processing [4,5]. By assigning a weight to each part of the input feature, representing the level of attention of the model to that part, attention mechanisms partially reduce the redundancy and irrelevant computational load associated with tasks. Dynamically focusing or ignoring information in images through the dynamic task-based saliency map (DTSM) is a direct method applied to images. Early studies Sorokin et al. [6] introduced attention mechanisms in the deep Q-network (DQN) training process. They found them effective only in certain environments, leading to performance degradation in others. Shi et al. [7] approximated the value function of whether DTSM was introduced after deep reinforcement learning (DRL) training, learning DTSM without annotations, which partially improved the performance of VDRL. However, this approach does not use DTSM during the DRL learning process and requires additional training resources. This contradicts the original intention of DRL to achieve human-level performance with limited interactions. DRL excels in tasks requiring large amounts of data collected through almost unlimited interactions with the environment. However, learning from limited interactions remains a significant challenge, as reported in studies on the Atari 100K tasks [8].

In studies on the Atari 100K dataset, the DrQ method was devised by leveraging data augmentation [9], while data-efficient rainbow (DER) [10] and DrQ(ϵ) [11] significantly improved the performance under a 100K setting by adjusting the hyperparameters of existing methods. The self-predictive representation (SPR) method [12], which predicts state self-representations, was introduced. DRL training relies

* Corresponding authors.
E-mail addresses: xjtulice@gmail.com (C. Li), hecd8132@sina.com (C. He).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.knosys.2024.112869
Received 9 May 2024; Received in revised form 2 December 2024; Accepted 10 December 2024
Available online 26 December 2024

heavily on early experiences, making it challenging to utilise later data. Periodic neuron reset operations were proposed to address the risk of overfitting to early data [13]. Schwarzer et al. [14] discusses the hyperparameter adjustments in the SPR method, achieving performance improvements solely through hyperparameter modifications. In these methods, visual images are inputs, and neural network map images are used in value functions in an end-to-end manner. Subsequently, weight updates of the neural network are performed through backpropagation based on the temporal difference error calculated from the value function.

The objective is to effectively diminish task-irrelevant information in images using DTSM, enhancing the performance on the Atari 100K dataset. In particular, DTSM is anticipated to empower agents to process image data flexibly, directing attention towards crucial areas and augmenting sensitivity to task-relevant information. Existing research has used DTSM to enhance interpretability, and some studies involved pre-training networks in a data augmentation format before training or learning DTSM after training. Therefore, the initial focus was to investigate why the performance deteriorates after introducing DTSM during the training phase. Interestingly, a peculiar phenomenon was observed when implementing the DTSM-integrated DQN algorithm in the game of Pong, where DTSM became all-zero after a certain number of training steps, rendering the agent incapable of perceiving any information at that juncture.

Herein, we propose a method called 'Don't overlook any detail (DOAD)' based on the SPR to address the issue of DTSM disappearance during the training process. First, a network structure was designed to extract DTSM and incorporate its weights into the original state images, aiming to emulate the human ability to focus on task-relevant regions in visual decision-making tasks. Second, a training process was devised to integrate DTSM and periodically perform conditional resets during training. This approach ensures the agent does not overly cling to early knowledge during the learning process. Instead, the model is encouraged to adapt more flexibly to new information through regular reset mechanisms. In summary, the aims of this study are as follows:

- Identify and analyse the limitations of integrating DTSM into deep reinforcement learning, highlighting its counterintuitive impact on performance degradation.
- Develop a novel DTSM learning module that enables saliency map extraction during DRL training without requiring additional annotations, enhancing flexibility in visual decision-making tasks.
- Propose an adaptive reset mechanism to mitigate the over-reliance on early knowledge during training, ensuring that the agent remains responsive to new information and improves long-term task performance.

2. Related work

2.1. Atari 100k

Atari games [15] are widely adopted as standard platforms for testing and developing algorithms. This series of classic arcade games, such as Breakout, Pong and Space Invaders, encompasses various levels of difficulty and complexity. Each Atari game possesses a unique state space, which describes the current state of the game, and an action space, which specifies the actions that the agent can execute. In Atari 100K, the environment state consists of four consecutive image frames, each resized to 84 x 84 pixels and converted to grayscale, and the actions are expressed through controller commands. The objective of the agent was to maximise the game score.

Several methods have been developed to train agents on Atari games [16–20], with value-based approaches being one of the most prominent. The DQN Mnih et al. [21], a method grounded in Q-learning [22] that uses DL, was introduced to minimise the loss function described in Eq. (1).

$$\mathcal{L}_\theta = \mathbb{E}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta)\right)^2\right] \tag{1}$$

where θ represents the parameters of the current network; θ^- denotes the parameters of the target network; r signifies the reward; γ is the discount factor; s' represents the next state; and a' is the action at the next step. Several subsequent methods built upon DQN, enhancing its performance in various ways. For example, the prioritised experience replay [23] improves the sample efficiency by ranking important transitions, while n-step learning [24] accelerates learning by bootstrapping multiple steps ahead. The distributional RL [25] captures the distribution of rewards, leading to more stable updates, and double Q-learning [26] addresses the overestimation bias in Q-learning. The duelling network architecture [21,27] decouples state-value and action-advantage functions, improving the accuracy of the value estimation. NoisyNets [28] introduces stochastic noise in the network to promote exploration. These advances were combined into the Rainbow DQN framework [29], showing how their integration can improve performance. However, while these methods achieved substantial improvements in various Atari games, they typically required extensive interactions with the environment to achieve human-level performance, leading to high exploration costs.

The Atari 100K benchmark [8] was proposed to address the challenge of sample efficiency, which limits the agent to 100K environment interactions (400K frames). This benchmark emphasises learning efficiency with limited data. Various methods, such as DER [10,18], DrQ(ϵ) [11], DrQ [9] and SPR [12], were developed as a response. These methods adapt existing DRL techniques to perform well under data constraints and focus on maximising the learning efficiency by augmenting data, optimising network hyperparameters or predicting future states to regularise the learning process.

Although these approaches improved performance with limited data, they often relied on tuning hyperparameters or augmenting training data rather than addressing the core challenge of dynamically managing the visual information most relevant to decision-making. In addition, methods like SPR and its variants [12,30] focused primarily on regularising the representation learning process to reduce overfitting to early-stage data. However, they did not directly tackle the challenge of isolating task-relevant information from noisy visual inputs, which can significantly affect the learning efficiency in visual tasks. Methods such as EfficientZero [31,32] and IRIS [33] have also shown remarkable results in limited data settings, but they still face difficulties in dealing with the large amounts of irrelevant information present in visual observations. Therefore, this study introduced effective visual attention mechanisms in VDRL to enhance the DRL performance with limited data.

2.2. Visual attention

Task-driven visual attention (TVA) is a crucial mechanism for focusing selectively on relevant stimuli in visual systems during specific tasks. By filtering out irrelevant information, TVA helps optimise cognitive resources and improves the response efficiency to critical task-related content. Previous studies examined TVA through various approaches. For example, human eye-tracking data were collected during Atari gameplay to train neural networks for predicting visual attention content [34]. While this approach provided valuable insights into human attention patterns, it depends strongly on the quality and availability of human-generated data. Similarly, pre-collected visual attention data were utilised for training networks, but this reliance on extensive labelled datasets introduced challenges related to cost and scalability [35].

This work was extended by incorporating eye-tracking data into decision-making processes [36], which enhanced agent performance. However, the improvement depended on the granularity and relevance of the eye-tracking data, presenting limitations when applied to more complex environments where collecting such data may be infeasible. A selective eye-gaze augmentation network trained on multiple human attention datasets was developed [37], but this approach also faced difficulties with the consistency of human attention data across different tasks and environments. Moreover, visual attention research spans many phenomena [38,39], with some studies focusing on attention in non-gaming industries [40–44]. However, many of these approaches encounter similar barriers regarding the cost and data dependency inherent to human visual attention studies.

Recent methods have addressed these challenges by exploring the ability of neural networks to learn visual attention autonomously. Liu et al. [45] introduces a model that adapts its focus to smaller, more relevant image regions during extended interactions. This autonomous approach reduces the dependency on human data but requires significant interactions, which may not be feasible in all scenarios. A dual attention mechanism is proposed to improve efficiency [46], but the method still requires substantial training data, limiting its generalisation to other tasks.

The impact of visual attention on the learning process was examined by introducing visual masks [47], offering improved interpretability but falling short in performance when compared with attention-free methods under the same data constraints. Guo et al. [48] examined the role of hyperparameters in influencing the DRL attention mechanisms, while Itaya et al. [49] applied masked visual attention to A3C networks, improving spatial awareness at the cost of increased training complexity. Self-attention in transformers was used [50] to enhance state–action representation, but this method remained computationally expensive. Hence, two approaches were introduced [51,52] for spatial and temporal attention, but both methods required extensive training data and often failed to outperform models without attention under identical training constraints.

3. Minor phenomenon

Broadbent's filter model and Treisman's attenuation theory elucidated how humans process sensory inputs from a cognitive psychology perspective, specifically how they selectively attend to certain information among numerous stimuli. When applied to attentional processing in images, our visual attention is represented by Eq. (2).

$$\mathcal{A}(x,y) = I(x,y) \cdot \mathcal{F}(x,y) \tag{2}$$

where $\mathcal{A}(x,y)$ represents the weighted image; $I(x,y)$ is the information on the original image at position (x,y); and $\mathcal{F}(x,y)$ is the visual attention map, with values ranging from 0 to 1. Initially, visual attention was applied to the state to obtain the weighted image, followed by training using the DQN-Adam method for 10 million steps in Pong. The rewards, original images, visual attention maps and weighted images were tracked during the training process, as shown in Fig. 1.

Training the Pong game using Deep RL can achieve a reward of 21. However, the lack of positive rewards in the early stages causes neuron deactivation when visual attention is introduced, leading to the subsequent occlusion of crucial information, such as the ball, ultimately resulting in training failure. After 25% of the steps, the states in the replay experience pass through the visual attention neural network $\mathcal{F}(x,y;\theta_F) = 0$, resulting in $\mathcal{A}(x,y;\theta_A) = 0$, as shown in Fig. 1. At this juncture, the weighted image contains no information, rendering the agent unable to train further. Furthermore, considering the partial loss of visual attention, only the ball in the Pong game is mapped to a loss through the visual attention neural network, as shown in Fig. 2.

The ball in this frame of the image is not lost. Instead, in the visual attention maps obtained by the visual attention neural network for all frames, the information on the ball is lost. Among the 26 games in Atari 100K, not all games exhibit this issue. However, this is also one of the potential reasons for the performance degradation in Atari 100K because of visual attention. The next section proposes an approach to address this phenomenon.

4. Don't overlook any detail

Fig. 3 shows the overall framework diagram of DOAD.

4.1. Network structure

In the proposed DRL architecture, the DOAD module can attain the capability of obtaining DTSM through a neural network, as formulated in Eq. (3):

$$\text{DOAD}: \tilde{s} = \mathcal{A}(s;\theta_a) \tag{3}$$

The input to this network is a four-channel image. Through a series of convolutional and activation layers, the network ensures that the output visual attention map is spatially aligned with the input image at the pixel level while maintaining channel alignment. By employing consecutive 3×3 convolutional layers and ReLU activation functions, the network progressively increases the channel count from 32 to 64 and then reduces it back to 32, effectively extracting spatial features from the image. Finally, a 3×3 convolutional layer reduces the channel count to 4, combined with a Sigmoid activation function, to generate an image directly reflecting the intensity of visual attention, with the channel count consistent with the input. The Sigmoid function restricts the output to the range of 0 to 1, with values close to 1 considered important and those close to 0 considered unimportant. During training, the visual attention gradually approaches 0 for irrelevant content, effectively blocking it from further computation.

The encoder module $\mathcal{E}_o$ and $\mathcal{E}_m$ serve as the primary feature extractors, gradually extracting features from the input states using a meticulously designed three-layer convolutional neural network, as depicted in Eq. (4):

$$\text{encoder}: z = \mathcal{E}(s;\theta_e) \tag{4}$$

The first convolutional layer uses 8×8 convolutional kernels with a stride of 4×4, effectively reducing the size of feature maps while preserving local feature information, resulting in 32 feature maps as output. The second convolutional layer utilises 4×4 convolutional kernels with a stride of 2×2, further refining the features and yielding 64 feature maps. Finally, the last convolutional layer uses 3×3 convolutional kernels with a stride of 1×1, maintaining the size of the feature maps unchanged while increasing the depth of the network, resulting in 64 feature maps. This hierarchical feature extraction process ensures the network can capture feature representations from low to high levels.

The feature representation z_t is mapped from the final convolutional layer to a 512-dimensional feature space, achieved via a fully connected layer. This high-dimensional feature space furnishes abundant information for subsequent decision-making and state transitions.

The Q-module in Q-learning uses a fully connected layer expressed as Eq. (5):

$$\text{Qhead}: q = \mathcal{Q}(z;\theta_h) \tag{5}$$

The 512-dimensional feature vector is mapped to a space defined by the number of actions and value atoms, representing the value distribution for each action. This mapping follows the configuration proposed by Schwarzer et al. [12] in SPR with visual attention.

The transition module integrates action information and the feature representation of the current state. This is achieved by convolving the feature map output by the encoder with the encoded feature representation of the current state and K action information. Subsequently, it

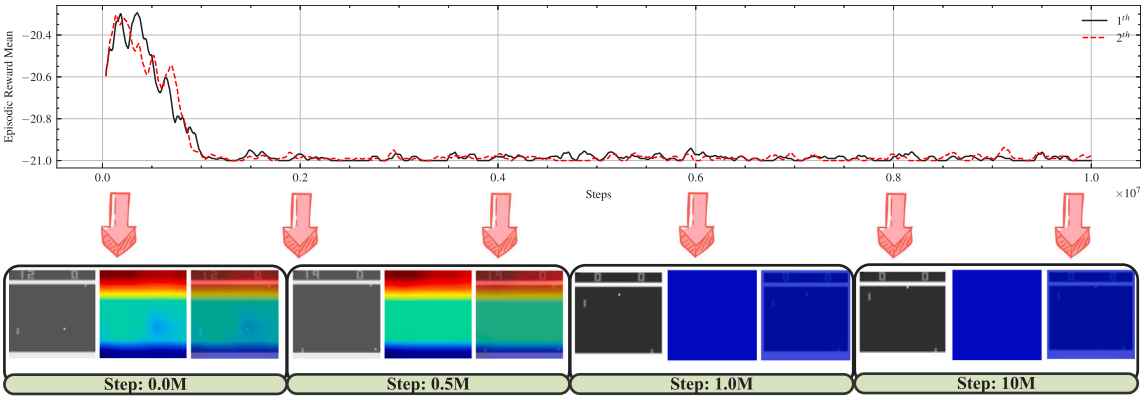


Fig. 1. Training with DQN-Adam in the Pong game for 10 million steps. Pong is a game with 21 rounds between a human and an AI. The top part represents the average reward during the training process, while the bottom part depicts the original image, visual attention map, and weighted image at different training steps. A reward of −21 means the player lost all rounds. During training, the exploration rate dropped to 0.05 after 1 million steps, leading to almost no further exploration. At this stage, the agent gained little reward. Although the initial attention mechanism allowed the agent to create attention maps from pixel values, the lack of reward led it to treat all information as irrelevant, quickly weakening its attention weights. By 0.5 million steps, the agent had stopped focusing on the ball, and by 1 million steps, its attention was nearly gone, causing the training to fail.

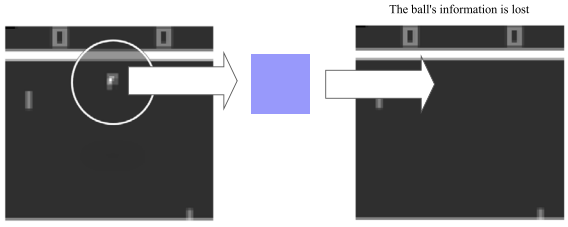


Fig. 2. The visual attention weight for the ball in Pong is 0, indicating a loss of visual attention, which results in the weighted image lacking information about the ball.

predicts multiple successive states and compresses them into a feature representation z_t, which is expressed as Eq. (6):

$$\text{transition}: z_{t+1} = \mathcal{T}\left(z_t, a_t; \theta_t\right) \tag{6}$$

The projection module $\mathcal{P}$ is responsible for further mapping the encoded state representation to a specific feature space using a fully connected layer. An additional mapping $\mathcal{P}_q$ was also established for predictions during the online process, as expressed in Eq. (7).

$$\text{projection}: y = \mathcal{P}\left(z; \theta_p\right) \qquad \text{prediction}: y = \mathcal{P}_q\left(y; \theta_{p_q}\right) \tag{7}$$

4.2. Visual attention, no overlooking

In DRL tasks, attention mechanisms often prematurely shift away from certain features, leading to a loss of important information, as discussed in Section 3. The characteristics of the ReLU activation function exacerbate this challenge, resulting in 'dead ReLU', where neurons become inactive with zero output. In such cases, it is essential to ensure that the visual attention mechanism retains its effectiveness throughout the training process, even when using ReLU.

The behaviour of ReLU and its impact on the distribution of output activations was examined to address this issue. ReLU outputs values in the range $(0, \infty)$, with dead neurons corresponding to zero activations. In contrast, the sigmoid function produces values close to 0.5 for zero input, providing a smoother gradient for learning. The risk of dead neurons was mitigated by applying ReLU after each convolutional layer while ensuring that 'dormant' neurons in the final visual attention map retain minimal weights, close to 0.5, to avoid being prematurely excluded from the attention process.

Therefore, this study defined a function $O(x, y) = \sigma(\mathcal{F}(I(x, y)))$, where σ represents the ReLU activation function. The output is $\sigma(\mathcal{F}(I(x, y))) = 0$, where $I(x, y) = 0$ when the input value is 0. The probability of dead neurons was analysed statistically after the layer output of each layer, denoted as $\mathbb{P}\left(\sigma\left(\mathcal{F}_{\cdot}(I(x, y))\right) = 0\right) = \varphi$, indicating that the proportion of dead ReLU is φ after each layer output, where the positions of dead ReLU correspond to zero outputs in the next layer. Therefore, it can be approximated as a problem, where, in $I(x, y)$, each pixel undergoes n repeated random assignments of zeros with a probability of φ. As n increases, the proportion of zero outputs in $O(x, y)$ also increased. Consequently, dead neurons result in missing or reduced representations in subsequent layers, potentially undermining the ability of the attention mechanism to focus on key information.

The method's name, 'no overlook', was revisited at this juncture. When $O(x, y) = 0$, the output after the sigmoid operation was 0.5. At this stage, the information in this position is overlooked rather than being entirely blocked out. The agent can still acquire this attenuated information during subsequent training.

This study also considered resetting the visual attention entirely, similar to the operation in SPRSR [14], to alleviate overfitting to early data. However, it still results in a decline in performance, possibly because of the requirement for a larger sample size for training visual attention. Moreover, the agent must adapt to a new randomly weighted image when the visual attention is reset. At this point, visual attention does not overlook any content but rather attenuates some information.

4.3. Training method

Building on the SPR framework, the SPRSR [14] method introduces a reset mechanism that significantly improves the model performance, increasing the interquartile mean (IQM) score from 0.44 to 0.61. This shows the efficacy of the mechanism in mitigating the issue of overfitting early data in DRL. A key consideration in this approach is determining which components should be reset to effectively harness the benefits of visual attention without destabilising the training process. Therefore, while a complete reset of the visual attention map might appear advantageous to reintroduce overlooked or previously blocked information, this strategy presents challenges. Resetting visual attention disrupts the weighted image representation, leading the DRL model to misinterpret images that differ from those seen during earlier approximations. This disruption increases the single-step prediction errors, accumulating errors in sequential decision-making processes. Therefore, we reset each of the last layers of three components encoder, transition model and Q-head, which are shown in Fig. 3. Here, 'reset' refers to the periodic re-initialisation of the weights of the network (network parameters), primarily using the Xavier initialisation method. Xavier initialisation is an effective approach for initialising neural networks, designed to ensure that the variance of the outputs

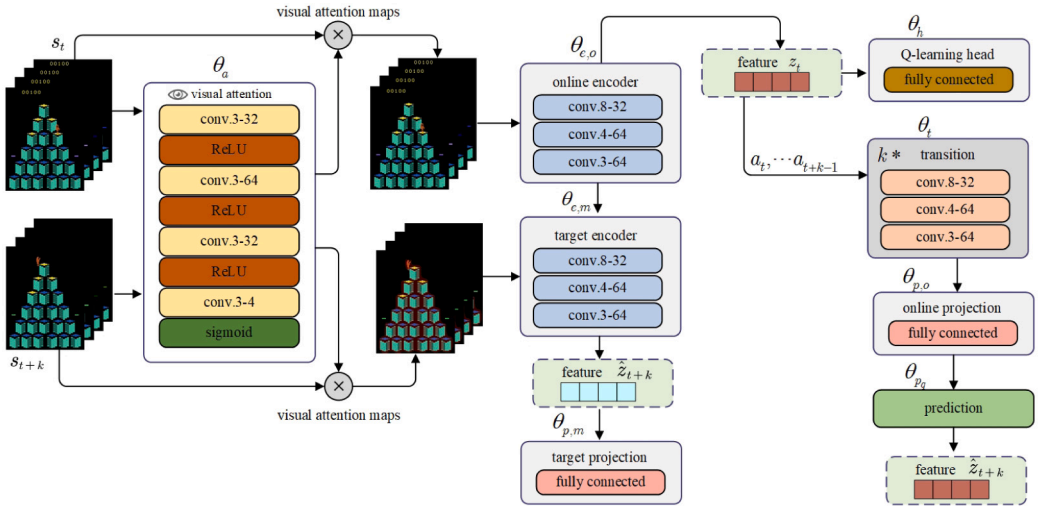


Fig. 3. Overall framework of DOAD. Representations from the online encoder are used in the VDRL task to predict future representations from the target encoder via the transition model. The target encoder and projection head are defined as an exponential moving average of their online counterparts and are not updated via gradient descent. For brevity, this paper illustrates only the kth step of future prediction, but in practice, the loss was calculated over all steps from 1 to K.

remains consistent across layers, facilitating efficient information flow throughout the network.

A two-stage training approach was adopted to maintain training stability. In the first stage, the visual attention mechanism, encoder and transition model were jointly trained to learn effective feature representations. The second stage focused on fine-tuning the visual attention mechanism alongside the Q-head, projection and prediction networks to improve the Q-value estimation and enhance the predictive capabilities of the model. Particularly, visual attention was refined throughout both stages. This continuous training is necessitated by the slow convergence of attention map learning, a challenge exacerbated by the introduction of periodic resets, which may cause instability in the visual attention maps due to intermittent target loss, as described elsewhere [12].

The overall SPR loss into two distinct components was decomposed to address this, as defined in Eqs. (8) and (9) [12], ensuring that the feature representation and value function learning proceed stably and efficiently.

$$\mathcal{L}_{SPR,1}(\theta_a, \theta_{e,o}, \theta_t) = -\sum_{k=1}^{K} \left(\frac{\mathcal{P}(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a)))}{\|\mathcal{P}(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a)))\|_2} \right)^T \cdot \left(\frac{\mathcal{P}_q\left(\mathcal{P}\left(\mathcal{T}\left(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a); \theta_{e,o}), a_{t+k}; \theta_t\right)\right)\right)}{\|\mathcal{P}_q\left(\mathcal{P}\left(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a); \theta_{e,o}), a_{t+k}; \theta_t\right)\right)\|_2} \right) \tag{8}$$

$$\mathcal{L}_{SPR,2}(\theta_a, \theta_{p,o}, \theta_{p_q}) = -\sum_{k=1}^{K} \left(\frac{\mathcal{P}(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a)))}{\|\mathcal{P}(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a)))\|_2} \right)^T \cdot \left(\frac{\mathcal{P}_q\left(\mathcal{P}\left(\mathcal{T}\left(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a)), a_{t+k}; \theta_{p,o}\right)\right)\right)}{\|\mathcal{P}_q\left(\mathcal{P}\left(\mathcal{E}(\mathcal{A}(s_{t+k}; \theta_a)), a_{t+k}; \theta_{p,o}\right); \theta_{p_q}\right)\|_2} \right) \tag{9}$$

The loss function of Q-Rainbow was used to assess the temporal-difference (TD) loss in DRL. The parameters included those of the DOAD module, Encoder and Q-head, as expressed in Eq. (10) [29]:

$$\mathcal{L}_{RL}\left(\theta_a, \theta_{p,o}, \theta_h\right) = \mathcal{L}_{\text{rainbow}}\left(s, a, r; \theta_a, \theta_{p,o}, \theta_h\right) \tag{10}$$

where $\mathcal{L}_{\text{rainbow}}$ represents the loss calculation method in Rainbow DQN. This study is influenced by advanced techniques introduced in DrQ, incorporating random shifting and colour dithering techniques. The aim of applying these techniques is to enhance the robustness of the model to visual inputs, enabling it to adapt better to various gaming environments. Specifically, the same random shifting and colour dithering techniques were used as DrQ to augment the diversity of training samples, enhancing the generalisation performance of the model across different scenarios. In summary, the pseudocode for the DOAD algorithm is shown in the Algorithm 1.

5. Experiments

5.1. Experimental setup

Twenty-six games were selected from the Atari 100K environment as benchmarks for training and testing. Consistent with previous experimental setups, the comparability of the experiments was ensured. The parameter settings during the training phase remained consistent with those used in prior works, such as SPR [12] and SPRSR [14], to ensure the reliability and reproducibility of the experimental results. Throughout the experimental process, metrics such as IQM, median, mean and human performance were used to validate the proposed model. The IQM metric, which calculates the average of the middle portion of the data after removing the extreme values, helps reduce noise and provides a more accurate reflection of the typical model performance. The median, representing the middle value when all samples were sorted in ascending order, better captured the central tendency of the data. The mean, calculated as the sum of all sample values divided by the number of samples, reflects the overall average level of the data. In human-machine comparisons, human performance is often used as a benchmark to evaluate the gap between the model performance and human performance in specific tasks. In many computer vision tasks, human performance is typically considered the upper bound; thus, a model that approaches or exceeds human performance exhibits exceptional ability.

5.2. Experimental result

The IQM was used as the evaluation criterion to compare DOAD with other methods, including [10], DrQ(ε) [11], SPR [12], visual attention, and SPRSR (without visual attention) [30]. IQM can objectively and meticulously analyse the performance of different algorithms in terms of image quality. This approach provided a more comprehensive and in-depth validation of the effectiveness of the proposed DOAD

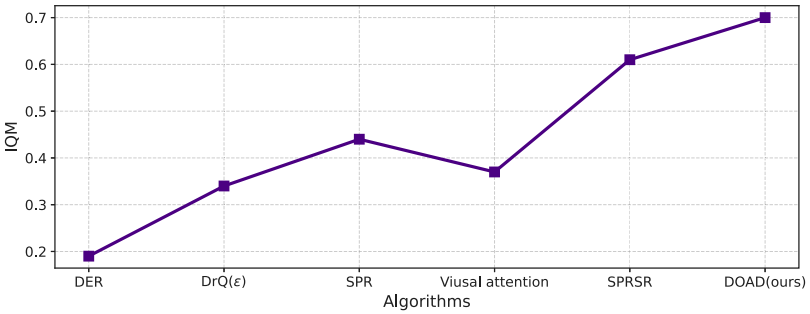


Fig. 4. Comparative IQM results for the 26 Atari 100K games.

Algorithm 1 Don't overlook any detail

Denote parameters of online encoder $\theta_{e,o}$ and projection $\theta_{p,o}$.
Denote parameters of target encoder $\theta_{e,m}$ and projection $\theta_{p,m}$.
Denote parameters of attention model θ_a.
Denote parameters of transition model θ_t, predictor θ_{p_q}, and Q-network θ_h.
Initialize replay buffer B.

while training **do**
  Collect experience (s, a, r, s') with $\theta_a, \theta_{e,o}, \theta_{p,o}, \theta_h$ and add to buffer B.
  Sample a minibatch of sequences $(s, a, r, s') \sim B$.
  $\tilde{s} \leftarrow \mathcal{A}\left(s; \theta_a\right), \tilde{s\prime} \leftarrow \mathcal{A}\left(s\prime; \theta_a\right)$
  for i in range$(0, N)$ **do** // N represents batch size
    if augmentation **then**
      $\tilde{s}^i \leftarrow \text{augment}(\hat{s}^i); \hat{s}^{ri} \leftarrow \text{augment}(\tilde{s}^{ri})$
    end if
    $z_0^i \leftarrow \mathcal{E}\left(\mathcal{A}(\hat{s}_0^i; \theta_a); \theta_{e,o}\right)$ // online representations
    $l^i \leftarrow 0$
    for k in $(1, \ldots, K)$ **do** // K represents prediction depth
      $\hat{z}_k^i \leftarrow \mathcal{T}(\hat{z}_{k-1}^i, a_{k-1}^i; \theta_t)$ // latent states via transition model
      $\tilde{z}_k^i \leftarrow \mathcal{E}(s_k^i; \theta_{e,m})$ // target representations
      $\hat{y}_k^i \leftarrow \mathcal{P}_q(\mathcal{P}(\hat{z}_k^i; \theta_{p,o}); \theta_{p_q}), \tilde{y}_k^i \leftarrow \mathcal{P}(\tilde{z}_k^i; \theta_{p,m})$ // projections
      Computer SPR loss with Eq $\mathcal{L}_{SPR,1}^i$
      Computer SPR loss with Eq $\mathcal{L}_{SPR,2}^i$ // SPR loss
    end for
    Computer RL loss with $\mathcal{L}_{RL}$
    Update parameters with $\mathcal{L}_{RL}$ // RL loss
  end for
  $\mathcal{L}_{SPR,1} \leftarrow \frac{1}{N}\sum_{i=0}^{N}\mathcal{L}_{SPR,1}^i$
  $\mathcal{L}_{SPR,2} \leftarrow \frac{1}{N}\sum_{i=0}^{N}\mathcal{L}_{SPR,2}^i$ // average loss over mini-batch
  Update parameters with $\mathcal{L}_{SPR,1}$
  Update parameters with $\mathcal{L}_{SPR,2}$
  $\theta_{e,m} \leftarrow \tau\theta_{e,o} + (1-\tau)\theta_{e,m}$ // update online parameters
  $\theta_{p,m} \leftarrow \tau\theta_{p,o} + (1-\tau)\theta_{p,m}$ // update target parameters
  if need reset **then**
    Reset $\theta_{e,o}$, θ_t and θ_h // Xavier reset
  end if
end while

algorithm in DRL tasks. Fig. 4 presents the comparative results based on IQM.

In Fig. 4, the performance declined when only the attention module was introduced based on the SPR algorithm. This decline was attributed to the loss of partial information caused by an early visual attention loss. However, overall performance was enhanced by introducing the DOAD mechanism. Table 1 provides a detailed breakdown of the scores for each game, indicating that the proposed method achieves the highest scores in 15 games.

Compared to the algorithms based solely on SPR, this method showed 13%, 6% and 9% improvements in the median metric, mean metric and IQM, respectively. This indicates the effectiveness of the proposed DOAD mechanism in visual attention, as it achieves significant improvements across multiple performance metrics relative to using attention alone.

In the Freeway game, the attention-based method achieved a score of 0.0, approaching a random outcome. Upon inspecting the visual attention maps, without the reset operation, attention tended to focus mainly on the lower part of the screen, as the chicken in this game primarily moves at the bottom during the early stages of the game. In contrast, the proposed method could attend to both the chicken and cars. This phenomenon arises because attention adapts to early overfitting experiences. However, after resetting the end-to-end mapping from states to Q-values, attention refocused on non-early experiences, leading to successful training. This phenomenon was also evident in other games where attention should not focus on early experiences, such as Pong and Bank Heist.

5.3. Training process

The training processes of the 26 games were compared, as shown in Fig. 5. After 100K training steps, SPR, with the introduction of the attention mechanism, learned faster and achieved better results than the reset mechanism in a few games (such as Breakout and MsPacman). However, in more games, the introduction of the attention mechanism leads to a performance decline. The visual attention maps were tracked during the learning process in these games, and no phenomenon of attention map blocking was found. However, visual attention map blocking occurred in games like Freeway and Pong.

The proposed DOAD method showed significant improvement in achieving optimal performance compared to the other methods. Compared to attention without any reset, the proposed method exhibited a periodic decrease in score. However, similar to SPRSR, after each reset, the reward of the next round increased significantly. This phenomenon suggests that the DOAD method can adjust the model parameters more flexibly during the training process, overcoming the potential phenomenon of visual attention blocking introduced by the attention mechanism. Additionally, it enables the agent to focus on important information in visual inputs, achieving more robust performance improvements across multiple game environments.

Furthermore, the loss during the training process of the 26 games was visualised, as shown in Fig. 6.

After each partial network reset, the loss increased and then quickly decreased, as shown in Fig. 6. However, when considering the reward trend in Fig. 6, after each reset and retraining, the agent no longer focused on early knowledge. In addition, as attention was not reset, the

Table 1
The performance for 26 Atari 100K games. Best performance is marked in green, and second best is marked in oragne.

Games	Random	Human	DER [10]	DrQ(ϵ) [11]	SPR [12]	Visual attention	SPRSR(without visual attention) [30]	DOAD (ours)
Alien	227.8	7127.7	802.3	865.2	901.2	780.1	911.2	**1226.4**
Amidar	5.8	1719.5	125.9	137.8	225.4	162.0	201.7	184.6
Assault	222.4	742.0	561.5	579.6	658.6	663.2	953.0	**1081.9**
Asterix	210.0	8503.3	535.4	763.6	**1095.00**	685.5	1005.8	819.5
Bank Heist	14.2	753.1	185.5	232.9	484.7	50.5	547.0	**653.6**
Battle Zone	2360.0	37,187.5	8977.0	10,165.3	10,873.5	9410.0	8821.2	**11,560.0**
Boxing	0.1	12.1	−0.3	9.0	27.6	21.1	32.2	**30.3**
Breakout	1.7	30.5	9.2	19.8	16.9	**25.4**	23.4	22.1
Chopper Command	811.0	7387.8	925.9	844.6	1454.0	1407.0	1680.6	**1780.0**
Crazy Climber	10,780.5	35,829.4	34,508.6	21,539.0	23,596.9	**27,261.0**	**28,936.2**	27,149.0
Demon Attack	152.1	1971.0	627.6	1321.5	1291.7	1520.1	2778.0	**2884.9**
Freeway	0.0	29.6	**20.9**	20.3	9.7	0.0	18.0	**24.9**
Frostbite	65.2	4334.7	871.0	1014.2	1746.2	**2228.7**	1834.3	**2951.1**
Gopher	257.6	2412.5	467.0	621.6	642.4	**757.8**	930.4	**762.2**
Hero	1027.0	30,826.4	6226.0	4167.9	**7554.5**	7208.6	6735.6	**7427.1**
Jamesbond	29.0	302.8	275.7	349.1	383.2	**482.5**	415.7	**442.5**
Kangaroo	52.0	3035.0	581.7	1088.4	**1674.8**	146.0	**2190.6**	732.0
Krull	1598.0	2665.5	3256.9	4402.1	3412.1	4408.7	**4772.4**	**4982.0**
Kung Fu Master	258.5	22,736.3	6580.1	11,467.4	**16,688.6**	**19,917.0**	14,682.1	14,282.0
Ms Pacman	307.3	6951.6	1187.4	1218.1	1334.1	**1920.9**	1324.6	**1493.6**
Pong	−20.7	14.6	−9.7	−9.1	**2.1**	−1.6	−9.0	**7.1**
Private Eye	24.9	69,571.3	72.8	3.5	76.1	**100.0**	82.2	**2936.8**
Qbert	163.9	13,455.0	1773.5	1810.7	3816.2	**4361.8**	3955.3	**4548.3**
Road Runner	11.5	7845.0	11,843.4	11,211.4	**13,588.5**	**15,530.0**	13,088.2	13,135.0
Seaquest	68.4	42,054.7	304.6	352.3	519.7	**769.4**	655.6	**953.0**
Up N Down	533.4	11,693.2	3075.0	4324.5	8873.4	14,336.1	**60,185.0**	**64,478.9**
Median HNS	0.00	1.00	0.19	0.31	0.45	0.28	0.51	**0.64**
Mean HNS	0.00	1.00	0.35	0.47	0.58	0.61	0.90	**0.96**
IQM HNS	0.00	1.00	0.19	0.34	0.44	0.37	0.61	**0.70**
#Games>Human	0	0	2	3	4	5	**7**	**7**
#Games>0	0	26	25	25	26	26	26	26

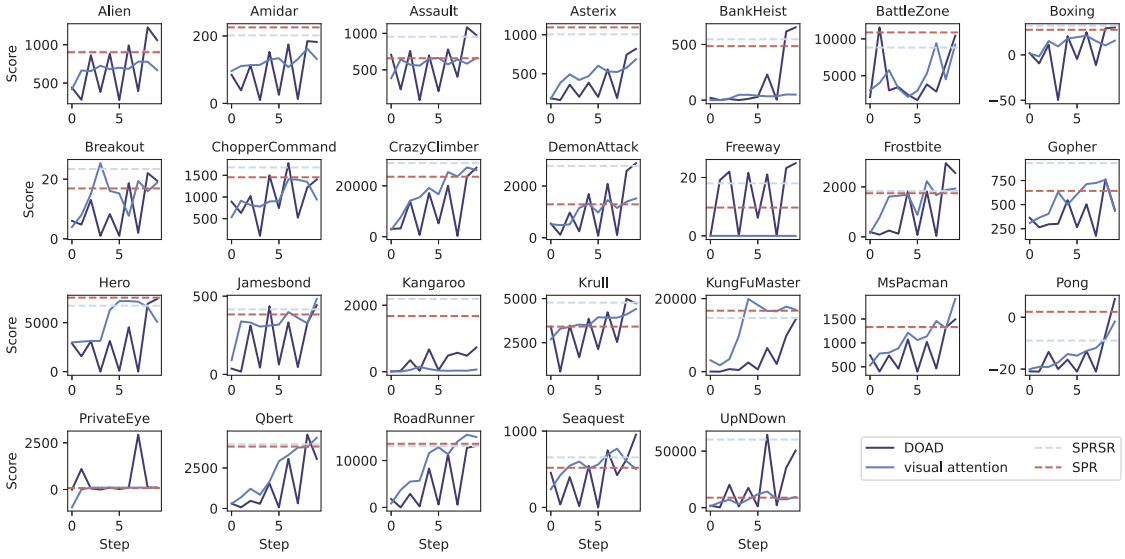


Fig. 5. Training process of different games, where the horizontal axis represents 10 evaluations and the vertical axis represents the score.

agent could concentrate on task-relevant information, thereby accelerating the training efficiency. Similar observations were made regarding the loss of SPR, as shown in Fig. 7.

The SPR loss increased after the encoder reset during the online phase. However, the loss decreased promptly after training, suggesting that even with the incorporation of attention, SPR can effectively learn state representations.

5.4. Ablation experiments

A series of ablation experiments was conducted to delve deeper into the impact of the attention mechanism on SPR performance and the effectiveness of introducing the reset mechanism. These experiments showcased results for multiple human-normalised score (HNS), such as median, IQM, mean and optimality gap, with the figures presenting the specific values of each performance metric in the form of the mean and confidence intervals. In the ablation experiments, the necessity of conducting two-stage training for attention was first assessed, followed by examining whether visual attention should be reset and, finally, whether other components should be reset. Fig. 8 shows the overall results of the ablation experiments, and the final results of each are listed in the tables.

5.4.1. Ablation experiment of the Parts 1 and 2

Herein, we propose a two-stage training method. In the Part 1, we train the visual attention mechanism, encoder, and transition model. In the Part 2, we train the Q-head, projection, and prediction networks. Ablation experiments were conducted to validate the effectiveness of this approach by replacing the training in Parts 1 and 2 with global training. Fig. 8 presents the overall results, and Table 2 lists the individual game scores.

Fig. 8 and Table 2 show that training attention solely in Part 1 resulted in more effective performance improvements than in Part 2;

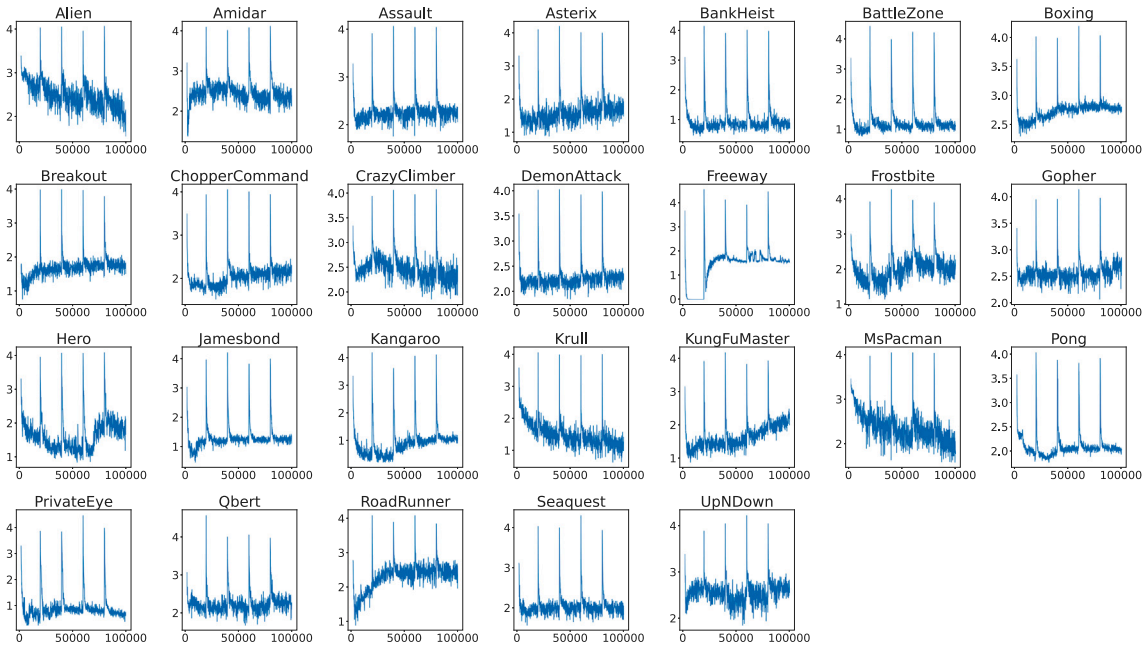


Fig. 6. Trend of DQN loss during the training process of the 26 games.

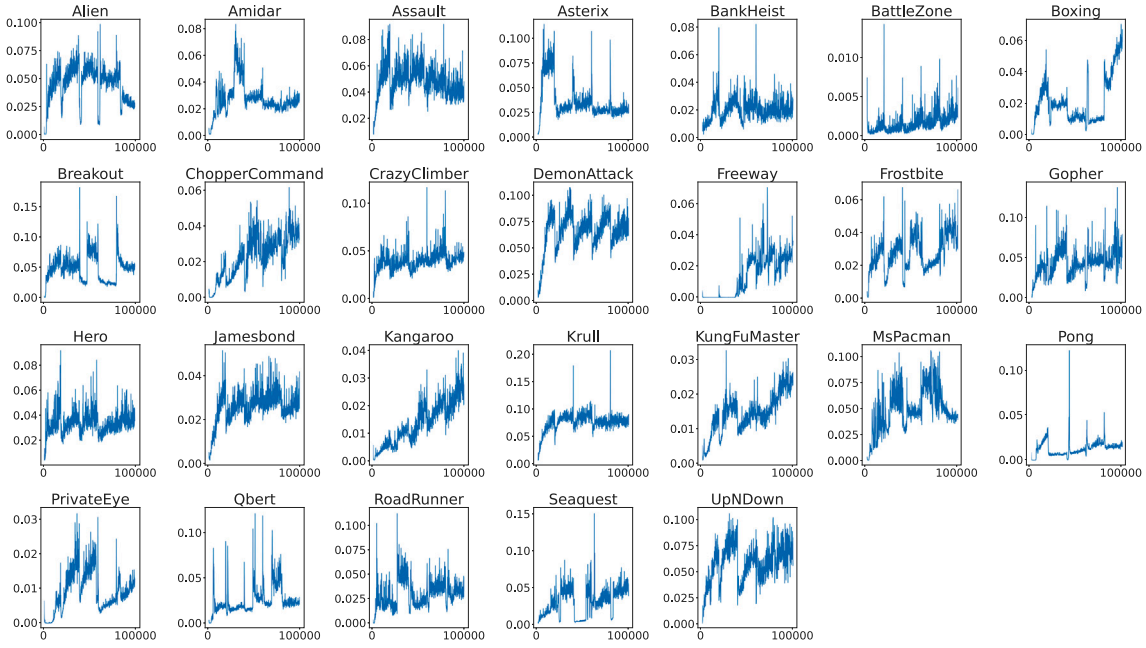


Fig. 7. Trend of the SPR loss during the training process of the 26 games.

it even outperformed training when compared to human-level performance, which includes median HNS, mean HNS, human scores, and IQM HNS. Part 1 focused primarily on the encoding part of the image, where decision-making based on the Rainbow Q-head and projection in Part 2 was relatively weaker. The variations in the attention maps leading to changes in attention images can significantly impact the strategy, explaining the decline in attention performance. The optimal outcomes were attained through training attention in Parts 1 and 2.

5.4.2. Ablation experiment of visual attention resetting

During the training process, this study examined which components should be reset. Ablation experiments were conducted to evaluate the impact of resetting different parts. Fig. 8 shows the overall results, and Table 3 lists the individual game scores.

Fig. 8 and Table 3 show that attention in visual attention DRL necessitates more training. Once trained to generate effective attention images, reducing visual information content enables the agent to process less information, enhancing its performance. Conversely, resetting attention altogether results in a performance decline because the agent cannot train stable attention images because of the insufficient samples. An ablation experiment was also conducted with and without ReLU. The results show that the presence of ReLU ensures that attention is not entirely diminished. Leveraging this mechanism, without altering the convolutional structure, a combination of ReLU and sigmoid was used,

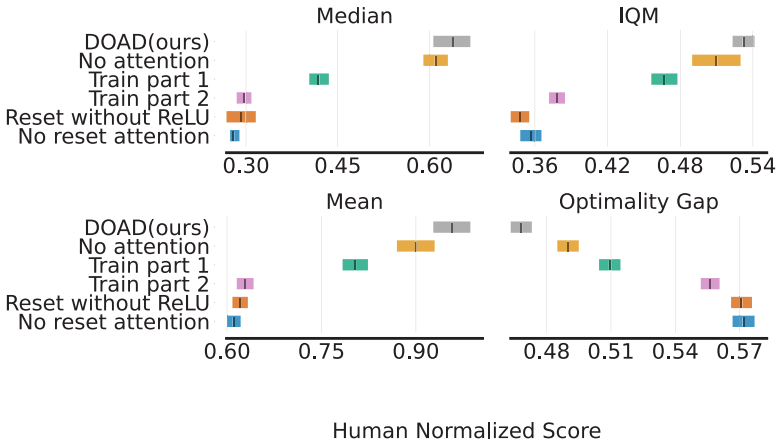


Fig. 8. Ablation experiment results, with each performance metric represented by its mean and confidence intervals. Each colour represents a different method.

Table 2
Training of Parts 1 and 2 of the performance metrics for 26 Atari 100K Games: Including median human-normalised score (HNS), mean HNS, human score, and IQM HNS.

Games	Train Part 1	Train Part 2	DOAD (ours)
Alien	719.7	874.1	1226.4
Amidar	180.78	184.80	184.64
Assault	916.23	749.79	1081.86
Asterix	620.00	659.50	819.50
Bank Heist	59.90	112.00	653.60
Battle Zone	7310.00	5370.00	11,560.00
Boxing	35.31	25.59	30.25
Breakout	18.62	21.81	22.06
Chopper Command	2792.00	1994.00	1780.00
Crazy Climber	23,138.00	22,669.00	27,149.00
Demon Attack	1565.20	2569.30	2884.90
Freeway	26.92	19.99	24.91
Frostbite	924.90	529.40	2951.10
Gopher	638.40	808.00	762.20
Hero	3136.40	9931.35	7427.05
Jamesbond	412.00	400.50	442.50
Kangaroo	4391.00	931.00	732.00
Krull	4304.50	4445.30	4982.00
Kung Fu Master	26,831.00	8299.00	14,282.00
Ms Pacman	1713.80	1793.90	1493.60
Pong	1.84	6.07	7.08
Private Eye	661.48	2674.81	2936.77
Qbert	4731.50	3253.75	4548.25
Road Runner	17,522.00	12,145.00	13,135.00
Seaquest	721.40	919.60	953.00
Up N Down	30,197.40	14,137.80	64,478.90
Median HNS	0.42	0.30	0.64
Mean HNS	0.80	0.63	0.96
#Games > Human	8	7	7
#Games > 0	26	26	26
IQM HNS	0.56	0.41	0.70

Table 3
Visual attention resetting performance for 26 Atari 100K games.

Games	Reset without ReLU	No reset attention	DOAD (ours)
Alien	896.20	780.10	1226.40
Amidar	184.55	161.99	184.64
Assault	757.25	663.16	1081.86
Asterix	723.00	685.500	819.50
Bank Heist	311.20	50.50	653.60
Battle Zone	10,210.00	9410.00	11,560.00
Boxing	31.02	21.05	30.25
Breakout	28.30	25.38	22.06
Chopper Command	1290.00	1407.00	1780.00
Crazy Climber	29,393.00	27,261.00	27,149.00
Demon Attack	1630.15	1520.05	2884.90
Freeway	21.99	0.00	24.91
Frostbite	1309.60	2228.70	2951.10
Gopher	649.80	757.80	762.20
Hero	11,713.80	7208.55	7427.05
Jamesbond	420.00	482.50	442.50
Kangaroo	448.00	146.00	732.00
Krull	4389.60	4408.70	4982.00
Kung Fu Master	4377.00	19,917.00	14,282.00
Ms Pacman	1933.70	1920.90	1493.60
Pong	−10.38	−1.64	7.08
Private Eye	2519.92	100.00	2936.77
Qbert	3161.00	4361.75	4548.25
Road Runner	15,740.00	15,530.00	13,135.00
Seaquest	675.00	769.40	953.00
Up N Down	4321.20	14,336.10	64,478.90
Median HNS	0.29	0.28	0.64
Mean HNS	0.62	0.61	0.96
#Games > Human	5	5	7
#Games> 0	26	26	26
IQM HNS	0.38	0.37	0.70

which ensured that the output was always greater than 0.5, preventing the suppression of visual information.

5.4.3. Ablation experiment of visual attention

Finally, ablation experiments were performed to assess the impact of introducing the visual attention mechanism. Fig. 8 shows the overall results, and Table 4 lists the individual game scores.

Fig. 8 and Table 4 show the effectiveness of the proposed method. From the overarching view of the ablation experiments, similar to learning in human visual tasks, individuals tend to form robust memories for visual attention areas, suggesting that attention is trained multiple times without reset. Regarding decision-making, one might become trapped in local optima if memories of early ideas persist excessively. The optimal approach aims to repeat the process of forgetting and training early knowledge hierarchically.

5.5. Visual attention maps' visualisation

The aim was to understand which parts of the images the agent focused on during the Atari 100K tasks by observing the visual attention maps. The visual attention maps were visualised for selected game frames, as shown in Fig. 9. This visualisation analysis aimed to show the attention patterns of the agent towards image information in different game environments, offering a more intuitive perspective for comprehending the algorithm behaviour.

The agent tends to assign higher weights to the main parts of the game frames and lower weights to non-essential parts (Fig. 9). This saliency distribution strategy helps reduce the amount of information processed in downstream visual tasks, enhancing the efficiency of the agent in handling crucial image information. This observation underscores the effectiveness of the proposed visual attention mechanism,

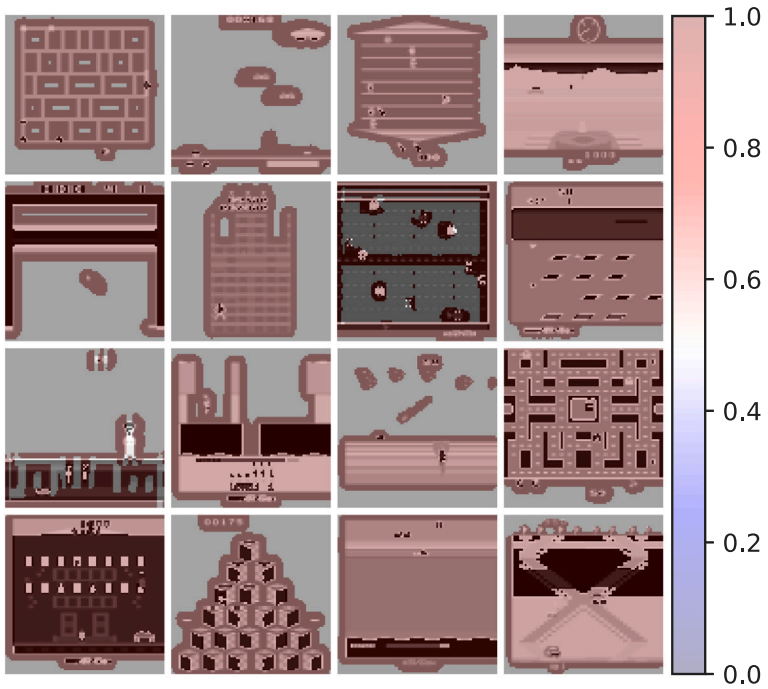


Fig. 9. Visual attention maps for selected games, where shades of blue indicate visual attention close to 0 and shades of red indicate visual attention close to 1.

Table 4
Visual attention performance for the 26 Atari 100K games.

Games	SPRSR (without visual attention)	DOAD (ours)
Alien	911.20	1226.40
Amidar	201.70	184.64
Assault	953.00	1081.86
Asterix	1005.80	819.50
Bank Heist	547.00	653.60
Battle Zone	8821.20	11,560.00
Boxing	32.20	30.25
Breakout	23.40	22.06
Chopper Command	1680.60	1780.00
Crazy Climber	28,936.20	27,149.00
Demon Attack	2778.00	2884.90
Freeway	18.00	24.91
Frostbite	1834.30	2951.10
Gopher	930.40	762.20
Hero	6735.60	7427.05
Jamesbond	415.70	442.50
Kangaroo	2190.60	732.00
Krull	4772.40	4982.00
Kung Fu Master	14,682.10	14,282.00
Ms Pacman	1324.60	1493.60
Pong	−9.00	7.08
Private Eye	82.20	2936.77
Qbert	3955.30	4548.25
Road Runner	13,088.20	13,135.00
Seaquest	655.60	953.00
Up N Down	60,185.00	64,478.90
Median HNS	0.51	0.64
Mean HNS	0.90	0.96
#Games > Human	7	7
#Games > 0	26	26
IQM HNS	0.610	0.70

which can focus on critical information in complex game environments, providing robust support for the agent to accomplish tasks more effectively.

The in-depth investigation aimed to understand the agent's attention on each image frame. This was achieved by conducting a detailed statistical analysis of the scores for each frame in the game; Fig. 10 presents the mean scores for all games. The horizontal axis represents the frame, while the vertical axis represents the saliency score. This analysis provided a more comprehensive understanding of the saliency distribution of the agent throughout the entire game process, providing more concrete data support for further understanding the behaviour of the algorithm.

Surprisingly, the agent focused on the last frame of the image and assigned higher saliency scores to the first two frames, as observed in the Breakout game. This phenomenon may be attributed to the movement of the bricks, enabling the agent to understand the trajectory of the ball by perceiving earlier frames. In contrast, for the Gopher game, the agent assigns higher saliency scores to the last frame as it contains crucial information about the timing of attacking the gopher.

Through an analysis of saliency maps and saliency score distributions, the areas attended to by this method align with human cognition in the game. This confirms the effectiveness of the proposed approach because the agent can simulate the attention patterns similar to humans when processing image information, providing valuable evidence for the robustness of the algorithm in complex tasks.

6. Conclusion

The DOAD method proposed in this paper addresses the issue of early visual attention loss that may arise from utilising the visual attention mechanism in environments with limited data samples. By leveraging the fundamental capabilities of visual attention, this approach ensures that the model prioritises task-relevant information in VDRL tasks. The incorporation of a conditional network reset strategy further enhances the model's adaptability and simulates human-like learning flexibility. This periodic reset mitigates the model's tendency to over-rely on early-stage data and fosters ongoing adaptation to new information. Moreover, the two-stage training process promotes more rapid learning of attention maps and contributes to enhanced stability during training following resets, resulting in significant improvements in model performance. In the Atari 100K benchmark, the DOAD method achieved an impressive IQM score of 0.64, highlighting its effectiveness in constrained data environments. This study provides deeper insights into integrating visual attention mechanisms in DRL and opens new research directions for improving the performance in data-limited settings.

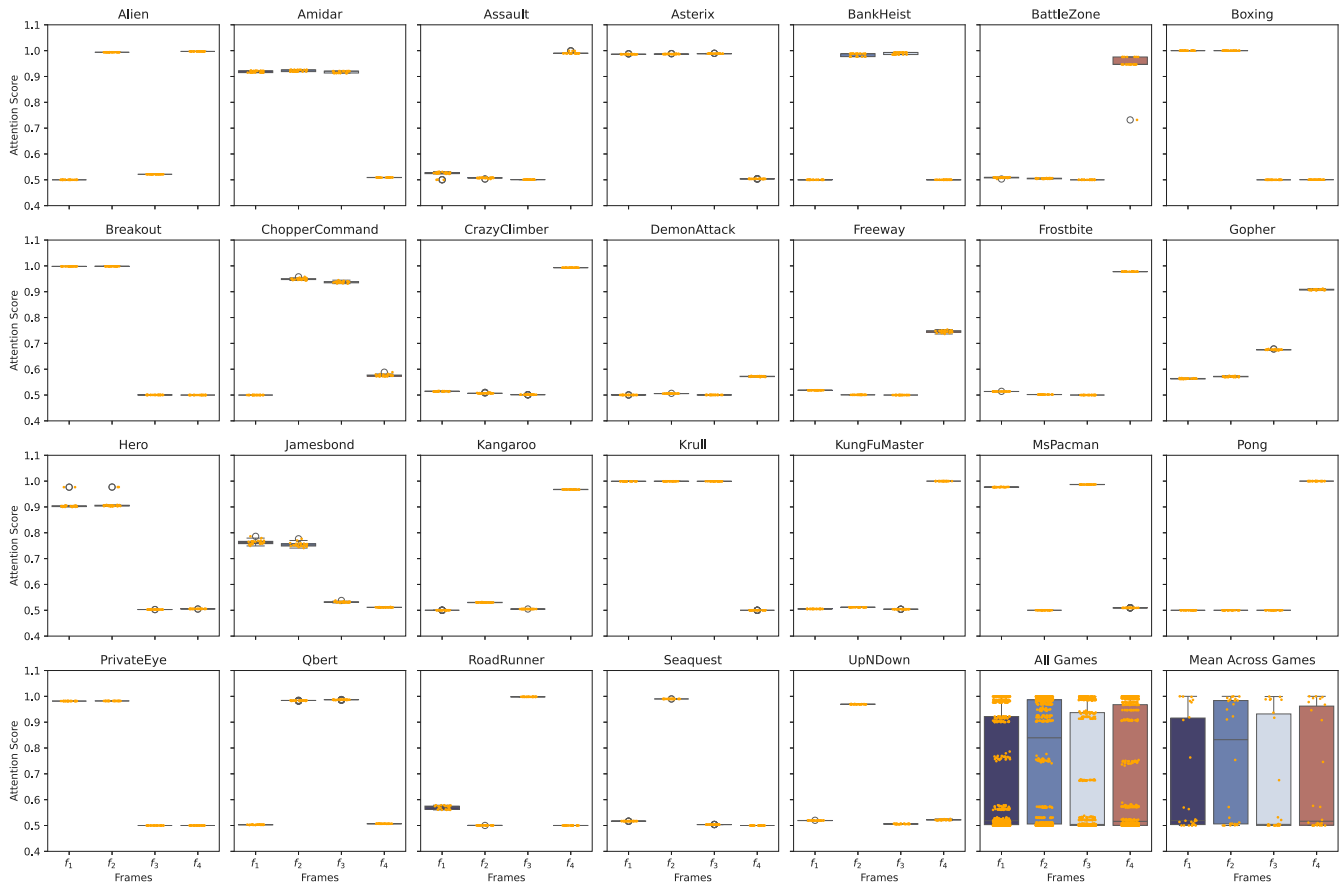


Fig. 10. Distribution of saliency scores assigned by the agent to each image frame, representing the attention across 100 consecutive steps sampled from four consecutive frames input into the DQN. The horizontal lines in the plot represent the mean values, while the yellow dots indicate the attention scores for individual frames.

Nevertheless, certain limitations remain. The effectiveness of the reset strategy may vary according to the task or environment. Hence, further research is needed to determine the optimal reset conditions and frequencies for different scenarios. Additionally, exploring alternative visual attention mechanism designs and training strategies is necessary to enhance the accuracy and efficiency of learning attention maps. In particular, future work should investigate methods capable of learning from limited samples within a stacked-frame approach, aiming to extract task-relevant information effectively across both static and dynamic backgrounds.

CRediT authorship contribution statement

Jialin Ma: Resources, Methodology, Conceptualization. **Ce Li:** Resources, Methodology, Conceptualization. **Zhiqiang Feng:** Methodology. **Limei Xiao:** Supervision, Conceptualization. **Chengdan He:** Funding acquisition. **Yan Zhang:** Funding acquisition.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used CHATGPT and GRAMMARLY in order to improve language and readability. After using these tools/services, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The work of this thesis was partially funded by the National Natural Science Foundation of China under Grant (No. 62363025), the Education Industry Support Plan Project of Gansu Provincial Department, China (No. 2023CYZC-26), the Key R&D plan of Science and Technology Plan of Gansu Province - Social Development Field Project, China (No. 23YFFA0064), and the Science and Technology on Vacuum &Cryogenics Technology and Physics Laboratory, China (No. 61422072305).

Data availability

No data was used for the research described in the article.

References

[1] J. Degrave, F. Felici, J. Buchli, M. Neunert, B. Tracey, F. Carpanese, T. Ewalds, R. Hafner, A. Abdolmaleki, D. de Las Casas, Magnetic control of tokamak plasmas through deep reinforcement learning, Nature 602 (7897) (2022) 414–419.
[2] H. Ju, R. Juan, R. Gomez, K. Nakamura, G. Li, Transferring policy of deep reinforcement learning from simulation to reality for robotics, Nature Mach. Intell. 4 (12) (2022) 1077–1087.
[3] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, D. Scaramuzza, Champion-level drone racing using deep reinforcement learning, Nature 620 (7976) (2023) 982–987.
[4] N. Le, V.S. Rathour, K. Yamazaki, K. Luu, M. Savvides, Deep reinforcement learning in computer vision: a comprehensive survey, Artif. Intell. Rev. 55 (4) (2022) 2733–2819.
[5] B. Nikpour, D. Sinodinos, N. Armanfard, Deep reinforcement learning in human activity recognition: a survey and outlook, Ieee Trans. Neural Netw. Learn. Syst. (2024).
[6] I. Sorokin, A. Seleznev, M. Pavlov, A. Fedorov, A. Ignateva, Deep attention recurrent q-network, 2015.

[7] W. Shi, G. Huang, S. Song, Z. Wang, T. Lin, C. Wu, Self-supervised discovering of interpretable features for reinforcement learning, IEEE Trans. Pattern Anal. Mach. Intell. 44 (5) (2022) 2712–2724.

[8] U. Kaiser, M. Babaeizadeh, P. Mios, B. Osiski, R.H. Campbell, K. Czechowski, Model based reinforcement learning for atari, in: International Conference on Learning Representations, 2019.

[9] I. Kostrikov, D. Yarats, R. Fergus, Image augmentation is all you need: regularizing deep reinforcement learning from pixels, 2020, arXiv preprint arXiv:2004.13649.

[10] H.P. Van Hasselt, M. Hessel, J. Aslanides, When to use parametric models in reinforcement learning, in: Advances in Neural Information Processing Systems, vol. 32, 2019.

[11] R. Agarwal, M. Schwarzer, P.S. Castro, A.C. Courville, M. Bellemare, Deep reinforcement learning at the edge of the statistical precipice, in: Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 29304–29320.

[12] M. Schwarzer, A. Anand, R. Goel, R.D. Hjelm, A. Courville, P. Bachman, Data-efficient reinforcement learning with self-predictive representations, 2020, arXiv preprint arXiv:2007.05929.

[13] P. D'Oro, M. Schwarzer, E. Nikishin, P.-L. Bacon, M.G. Bellemare, A. Courville, Sample-efficient reinforcement learning by breaking the replay ratio barrier, in: Deep Reinforcement Learning Workshop NeurIPS 2022, 2022.

[14] M. Schwarzer, J.S.O. Ceron, A. Courville, M.G. Bellemare, R. Agarwal, P.S. Castro, Bigger, better, faster: Human-level atari with human-level efficiency, in: International Conference on Machine Learning, PMLR, 2023, pp. 30365–30380.

[15] M. Bellemare, Y. Naddaf, J. Veness, M. Bowling, The arcade learning environment: an evaluation platform for general agents, J. Artif. Intell. Res. 47 (2012).

[16] J. Robine, M. Höftmann, T. Uelwer, S. Harmeling, Transformer-based world models are happy with 100k interactions, 2023, arXiv preprint arXiv:2303.07109.

[17] W. Zhang, G. Wang, J. Sun, Y. Yuan, G. Huang, STORM: Efficient stochastic transformer based world models for reinforcement learning, Adv. Neural Inf. Process. Syst. 36 (2024).

[18] O.V. Cagatan, B. Akgun, BarlowRL: Barlow twins for data-efficient reinforcement learning, in: Asian Conference on Machine Learning, PMLR, 2024, pp. 201–216.

[19] K. Zheng, X. Zhang, C. Wang, Y. Li, J. Cui, L. Jiang, Adaptive collision avoidance decisions in autonomous ship encounter scenarios through rule-guided vision supervised learning, Ocean Eng. 297 (2024) 117096.

[20] K. Zheng, X. Zhang, C. Wang, M. Zhang, H. Cui, A partially observable multi-ship collision avoidance decision-making model based on deep reinforcement learning, Ocean Coast. Manag. 242 (2023) 106689.

[21] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.

[22] C.J. Watkins, P. Dayan, Q-learning, Mach. Learn. 8 (1992) 279–292.

[23] T. Schaul, J. Quan, I. Antonoglou, D. Silver, Prioritized Experience Replay, 2016.

[24] R. Sutton, Learning to predict by the method of temporal differences, Mach. Learn. 3 (1988) 9–44.

[25] M. Bellemare, W. Dabney, R. Munos, A distributional perspective on reinforcement learning, 2017.

[26] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double q-learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30, 2016.

[27] Z. Wang, N. Freitas, M. Lanctot, Dueling network architectures for deep reinforcement learning, 2015.

[28] M. Fortunato, M.G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, S. Legg, Noisy nnetworks for exploration, in: CoRR, 2023, pp. 10295–10304.

[29] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: combining improvements in deep reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, 2018.

[30] E. Nikishin, M. Schwarzer, P. D'Oro, P.-L. Bacon, A. Courville, The primacy bias in deep reinforcement learning, in: International Conference on Machine Learning, PMLR, 2022, pp. 16828–16847.

[31] W. Ye, S. Liu, T. Kurutach, P. Abbeel, Y. Gao, Mastering atari games with limited data, in: Advances in Neural Information Processing Systems, vol. 34, 2021, pp. 25476–25488.

[32] Y. Mei, J. Gao, W. Ye, S. Liu, Y. Gao, Y. Wu, Speedyzero: Mastering atari with limited data and time, in: The Eleventh International Conference on Learning Representations, 2023.

[33] V. Micheli, E. Alonso, F. Fleuret, Transformers are sample-efficient world models, in: International Conference on Learning Representations, 2022.

[34] L.X. Zhang, R.H. Zhang, Z.D. Liu, M.M. Hayhoe, D.H. Ballard, Learning attention model from human for visuomotor tasks, in: Thirty-Second Aaai Conference on Artificial Intelligence, 2018, pp. 8181–8182.

[35] R.H. Zhang, Z.D. Liu, L.X. Zhang, J.A. Whritner, K.S. Muller, M.M. Hayhoe, D.H. Ballard, AGIL: learning attention from human for visuomotor tasks, in: Computer Vision - eccv 2018, Pt Xi, vol. 11215, 2018, pp. 692–707.

[36] R.H. Zhang, C. Walshe, Z.D. Liu, L. Guan, K.S. Muller, J.A. Whritner, L.X. Zhang, M.M. Hayhoe, D.H. Ballard, Atari-HEAD: atari human eye-tracking and demonstration dataset, in: Thirty-Fourth Aaai Conference on Artificial Intelligence, vol. 34, 2020, pp. 6811–6820.

[37] C. Thammineni, H. Manjunatha, E.T. Esfahani, Selective eye-gaze augmentation to enhance imitation learning in atari games, Neural Comput. Appl. 35 (32) (2023) 23401–23410.

[38] M. Carrasco, Visual attention: The past 25 years, Vis. Res. 51 (13) (2011) 1484–1525.

[39] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[40] F. Beuth, T. Schlosser, M. Friedrich, D. Kowerko, Improving automated visual fault detection by combining a biologically plausible model of visual attention with deep learning, in: IECON 2020 the 46th Annual Conference of the IEEE Industrial Electronics Society, IEEE, 2020, pp. 5323–5330.

[41] T. Schlosser, M. Friedrich, F. Beuth, D. Kowerko, Improving automated visual fault inspection for semiconductor manufacturing using a hybrid multistage system of deep neural networks, J. Intell. Manuf. 33 (4) (2022) 1099–1123.

[42] M. Jalal, I.U. Khalil, A. ul Haq, Deep learning approaches for visual faults diagnosis of photovoltaic systems: State-of-the-art review, Results Eng. (2024) 102622.

[43] S.N. Venkatesh, B.R. Jeyavadhanam, A.M. Sizkouhi, S.M. Esmailifar, M. Aghaei, V. Sugumaran, Automatic detection of visual faults on photovoltaic modules using deep ensemble learning network, Energy Rep. 8 (2022) 14382–14395.

[44] H. Du, L. Li, Z. Huang, X. Yu, Object-goal visual navigation via effective exploration of relations among historical navigation states, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2563–2573.

[45] Q. Liu, J. Zhai, S. Zhong, Z. Zhang, Q. Zhou, P. Zhang, A deep recurrent q-network based on visual attention mechanism, Chin. J. Comput. 40 (6) (2017) 1353–1366.

[46] X. Ling, J. Li, F. Zhu, Q. Liu, Y. Fu, Asynchronous advantage actor-critic with double attention mechanisms, Chin. J. Comput. 43 (1) (2020) 93–106.

[47] E. Gilmour, N. Plotkin, L.N. Smith, An approach to partial observability in games: learning to both act and observe, in: 2021 Ieee conference on games, 2021, pp. 971–975.

[48] F. Guo, Q. Wei, M. Wang, Z. Guo, S.W. Wallace, Deep attention models with dimension-reduction and gate mechanisms for solving practical time-dependent vehicle routing problems, Transp. Res. Part E: Logist. Transp. Rev. 173 (2023) 103095.

[49] H. Itaya, T. Hirakawa, T. Yamashita, H. Fujiyoshi, K. Sugiura, Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning, in: 2021 International Joint Conference on Neural Networks, 2021.

[50] J.H. Shang, K. Kahatapitiya, X. Li, M.S. Ryoo, StARformer: tansformer with state-action-reward representations for visual reinforcement learning, in: Computer vision, eccv 2022, Pt Xxxix, vol. 13699, 2022, pp. 462–479.

[51] H. Liu, P. Liu, C. Bai, Research on a fusion method of spatial relationship and memory in deep reinforcement llearning, Chin. J. Comput. 46 (4) (2023) 814–826.

[52] M. Ramicic, A. Bonarini, Uncertainty maximization in partially observable domains: a cognitive perspective, Neural Netw. 162 (2023) 456–471.

Applied Intelligence (2025) 55:565
https://doi.org/10.1007/s10489-024-05850-y

Vision-based attention deep q-network with prior-based knowledge

Jialin Ma[1] · Ce Li[1] · Liang Hong[1] · Kailun Wei[1] · Shutian Zhao[1] · Hangfei Jiang[1] · Yanyun Qu[2]

Accepted: 7 November 2024

Abstract
Vision-based reinforcement learning (RL) is a potent algorithm for addressing tasks related to visual behavioural decision-making; nevertheless, it operates as a black-box, directly training models with images as input in the end-to-end fashion. Therefore, to elucidate the underlying mechanisms of the model and the agent's focus on different features during the decision-making process, a vision-based attention (VA) mechanism is introduced into vision-based RL in this paper. A prior-based mechanism is introduced to address the issue of instability in the attention maps observed by the agent when attention mechanisms are directly integrated into network updates that results in an increase in single-step errors and larger cumulative errors. Thus, a vision-based attention deep Q-network (VADQN) method with a prior-based mechanism is proposed. Specifically, prior attention maps are obtained using a learnable Gaussian filtering and a spectral residual method. Next, the attention maps are fine-tuned using a self-attention (SA) mechanism to enhance their performance. During training, both the attention maps and the parameters of the policy network are concurrently trained to ensure explanations of the regions of interest during online training. Finally, a series of ablation experiments are conducted on Atari games to compare the proposed method with humans, convolutional neural networks, and other approaches. The results demonstrate that the proposed method not only reveals the regions of interest attended to by DRL during the decision-making process but also enhances DRL performance in certain scenarios. This approach provides valuable insights for understanding and improving the performance of DRL in visual decision-making tasks.

Keywords VADQN · Vision-based RL · Self-attention · Learnable gaussian filtering · Spectral residual

1 Introduction

Vision-based reinforcement learning (RL) [1, 2] has emerged as a crucial tool in deep reinforcement learning (DRL) for simulating human-like perception and decision-making processes. While tabular Q-learning has achieved some success in RL [3], its effectiveness is limited when addressing complex and nonlinear problems, leading to the development of deep Q-networks (DQN) [4]. DQN leverages deep neural networks [5–7] to model nonlinear state spaces, resulting in significant advancements in DRL and even achieving human-level performance in Atari games [4, 8, 9]. However, the end-to-end deep approximators in DRL, which function as black-box models, present challenges in terms of accuracy [10–12].

By incorporating vision-based attention (VA) mechanisms, we can gain deeper insights into the focal points of agents during the training of vision-based DRL [13–15]. In computer vision, common attention mechanisms include spatial attention [16, 17], channel attention [18–20], and self-attention (SA) [21–23]. Sorokin et al. [24] integrated attention mechanisms into DQN training to highlight critical regions on the game screen. However, this approach led to a decline in the DQN performance across most games. Shi et al. [8] proposed a self-supervised interpretable framework that uses attention masks to explain agent decision-making. Xing et al. [13] applied attention mechanisms within a distillation approach to produce policies that are both highly

Ce Li contributed equally to this work

✉ Ce Li
xjtulice@gmail.com

Jialin Ma
jialinm@lut.edu.cn

[1] School of Information Science and Technology, Lanzhou University of Technology, 36 Pengjiaping Road, Lanzhou, Gansu 730050, China

[2] School of Informatics, Xiamen University, 422 Siming South Road, Xiamen, Fujian 361005, China

Published online: 24 March 2025

interpretable and computationally efficient. Liu et al. [25] introduced an end-to-end interpretability method based on an adaptive region scoring mechanism. Although these methods have improved DRL performance in certain cases, they often require increased training iterations or interactions with the environment. Importantly, DRL performance is typically evaluated based on a fixed number of environment interactions; further training or reintegration of VA before or after training may enhance performance, but this does not align with the way humans learn and make decisions through visual perception. The performance degradation is attributed mainly to the accumulation of errors in DRL. When attention mechanisms are directly integrated into network updates, the attention maps observed by the agent can become unstable, leading to increased single-step decision errors. This error accumulation ultimately results in compounding inaccuracies, diminishing overall performance.

In deep learning, VA mechanisms are generally implemented by learning network weights. For example, spatial and channel attention allocate weights across the spatial and channel dimensions of an image, respectively, to identify important regions and key features [26, 27]. In some instances, these two attention mechanisms are combined to optimize the attention distribution. SA mechanisms capture internal correlations within the data, reducing dependency on external information. Additionally, Transformers employ multirhead SA mechanisms, enabling models to capture semantic information across different spatial dimensions. Since Transformers are applied to vision tasks, researchers have further explored how to effectively utilize attention mechanisms in these tasks [28, 29]. Early studies focused primarily on incorporating attention into deep model black boxes to reveal the areas of focus for agents [30, 31]. Although these methods have shown better results in deep learning, when applied to DRL, particularly in visual feature extraction, the integration of attention mechanisms often requires large amounts of data, potentially hindering the exploration efficiency of DRL; this contradicts the objective of DRL, which trains agents to reach decision-making proficiency with limited environment interactions.

Because DRL is a sequential decision-making model, its accuracy typically decreases with increasing number of decision steps. To address this issue, researchers have proposed various mechanisms to gain insights into the perception process of DRL. Inspired by the biological nervous system and early attention mechanisms, researchers have begun integrating attention mechanisms into deep models during online training to interpret model behaviour during the training process. However, direct integration of VA into network updates in this online attention approach often significantly reduces the performance of vision-based DRL; this is primarily due to error accumulation. When attention mechanisms are directly embedded into network updates, the attention maps observed by the agent can become unstable, leading to an increase in single-step decision errors. This accumulation of errors ultimately results in more severe error accumulation. Moreover, the integration of attention mechanisms generally requires large amounts of data, further impeding the exploration efficiency of DRL.

To address these challenges and better elucidate the internal mechanisms of vision-based DRL while enhancing its performance, this paper proposes a vision-based attention deep q-network (VADQN) method. Inspired by the innate visual understanding capabilities of infants, we introduce a prior-based mechanism to address the instability of attention maps and the error accumulation that arises when attention is directly integrated into network updates. To reduce noise in the amplitude spectrum, we employ a learnable Gaussian filter (LGF). Subsequently, we propose a spectral residual (SR) method to identify differences between attention maps based on local amplitude variations. Building on this, we incorporate an SA mechanism to further optimize attention maps and minimize interference from high-level attention tasks. Additionally, we present a learning method for vision-based DRL tasks, aiming to interpret which parts of an image are attended to under a given policy. The main contributions of this research can be summarized as follows.

(1) **Introduction of attention mechanisms in DRL:** We integrate a vision-based attention mechanism into the DRL model, enhancing the model's ability to focus on relevant information. Our emphasis is on understanding how the model perceives and processes information during decision-making.
(2) **Addressing Attention Challenges in Online Learning:** The proposed method effectively addresses the challenges associated with attention in online reinforcement learning, significantly reducing excessive fluctuations in the feature maps used for policy mapping.
(3) **Online training of attention and policy networks:** We propose a method for simultaneously training attention maps and policy networks in online reinforcement learning. This approach provides insights into how humans focus during decision-making, outperforming conventional approaches that use attention mechanisms to interpret pretrained policy black boxes.

2 Related work

2.1 Atari Games

Atari 2600 games, as implemented in the Arcade Learning Environment (ALE), serve as a challenging testbed for RL. These games present agents with high-dimensional visual input and a diverse, intriguing set of tasks that were originally

designed to challenge human players. The Markov decision process (MDP) model for Atari games can be defined as follows [32]:

State: The environment's state is represented by the screen image.

Action: An agent's actions correspond to joystick movements.

Reward: The reward is the score achieved by the agent.

Transition Probability Function: This function represents the probability of transitioning from one state to another given a specific action.

Discount Factor: The discount factor is utilized to discount future rewards.

In the Nature DQN study [4], the images were grayscale and had the dimensions of 84x84 pixels. These preprocessed images are stacked to form a 4-frame sequence, which serves as the input to the DRL model, as shown in Fig. 1.

This 4-frame sequence is employed to capture the game's motion information. We use the same definitions in this paper.

2.2 Deep reinforcement learning

DRL is a specialized branch of RL that synergistically combines deep learning with RL, enabling the solution of complex decision-making and control tasks. In DRL, deep neural networks serve as function approximators, directly estimating the value function or policy from high-dimensional sensory inputs, such as raw pixels in images or audio signals.

DQN is a pioneering DRL model that uses a convolutional neural network (CNN) to learn control policies directly from raw pixel inputs. The DQN extends Q-learning, and its core equation is presented in (1).

$$Q(s,a) \leftarrow Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \quad (1)$$

where s represents the current state, a denotes the action taken, r represents the immediate reward, s' represents the subsequent state, a' represents the subsequent action, α denotes the learning rate, and γ represents the discount factor.

The DQN has undergone numerous improvements and extensions, one of which is the Double DQN (DDQN) [33]. This variant addresses the overestimation bias in the action values that is inherent in the DQN by using two distinct Q-functions to decouple the processes of action selection and action evaluation. The update rule for the Q value in the double DQN algorithm is presented in (2).

$$\begin{aligned} Q(s,a) \leftarrow\ & Q(s,a) + \\ & \alpha(r + \gamma Q(s', \arg\max_{a'} Q(s',a')) - Q(s,a)) \end{aligned} \quad (2)$$

The landscape of the DQN also includes other variations, such as the Q-rainbow [34], Data-Regularized Q (DrQ) [35], and self-predictive representations (SPR) [36]. These algorithms have been applied to a wide range of tasks, from playing Atari games to addressing continuous control benchmark tasks and even venturing into the domain of quantum RL. Impressively, they have shown promising results in learning complex policies, often achieving state-of-the-art performance.

These endeavours involve black-box mapping from state to policy, where the process of how agents derive cognition from images remains a fascinating, yet elusive, research topic.

2.3 Parameter-free salient object detection

In the context of parameter-free salient object detection, two key techniques are often employed: spectral residual (SR) [37] and Gaussian filtering (GF).

SR is utilized to detect the salient regions within an image. This model is based on the spectral analysis of the image, which involves the calculation of the log spectrum of the image, followed by determining the SR. The SR is obtained by subtracting the smoothed log spectrum (achieved through averaging within the local neighbourhood of the log spectrum) from the original log spectrum. The saliency map is generated through the inverse Fourier transform (FT) of the SR. The mathematical representation of the SR model can be expressed by (3).

$$\begin{aligned} L(x,y) &= \log(|F(I(x,y))|) \\ R(x,y) &= L(x,y) - h_{av}(L(x,y)) \\ S(x,y) &= F^{-1}(e^{R(x,y)+i\Theta(x,y)}) \end{aligned} \quad (3)$$

where $I(x,y)$ is the input image, F and F^{-1} are the FT and inverse FT, $L(x,y)$ is the log spectrum, h_{av} is the average

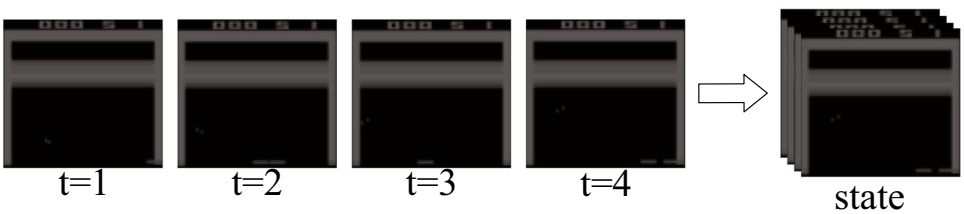


Fig. 1 Preprocessed grayscale images are stacked to form a 4-frame sequence. Each image is 84x84 pixels, and the resulting composite serves as the input state for DRL

filter, $R(x, y)$ is the spectral residual, $S(x, y)$ is the saliency map, and $\Theta(x, y)$ is the phase spectrum of the image.

GF is a technique used to blur an image or reduce its noise; it uses a Gaussian function, which is a function of the form in (4).

$$G(x, y) = \frac{1}{2\pi\sigma^2} \times \exp(-\frac{x^2 + y^2}{2\sigma^2}) \tag{4}$$

where x and y are the distances from the origin in the horizontal and vertical axes, respectively, and where σ is the standard deviation of the Gaussian distribution. GF is a convolution filter that uses this function as its impulse response. The convolution of an image $I(x, y)$ with a filter $G(x, y)$ is given by (5).

$$I'(x, y) = I(x, y) * G(x, y) \tag{5}$$

where $I'(x, y)$ is the filtered image and $*$ denotes the convolution operation. GF is often used in the preprocessing or postprocessing stages of salient object detection to smooth the saliency map and suppress noise.

2.4 Salient object detection for DRL

In the context of DRL, object-sensitive techniques have been proposed to incorporate object recognition processing into deep reinforcement learning models [38]. One such technique, called motion-oriented RL, automatically detects moving objects and uses the relevant information for action selection [39]. The detection of moving objects is performed in an unsupervised manner by exploiting the structure changes due to motion.

The object-sensitive representation learning technique can be combined with most value-based, policy optimization, and actor-critic algorithms. The technique is demonstrated on a suite of Atari games where the ability to detect moving objects reduces the amount of interactions with the environment necessary to obtain a good policy.

Shi et al. [8] argue that the agent provides the same evaluation for observations with saliency maps added as for previous observations, as follows in (6).

$$Q(f_{SOD}(s) \cdot s, a) = Q(s, a) \tag{6}$$

where $f_{SOD}(s)$ represents the saliency map regarding s. This method achieves the extraction of saliency maps during the unsupervised RL process. However, since this method implements saliency map extraction after training the policy, there is still a gap in understanding saliency cognition during the reinforcement learning training process.

3 Vision-based attention deep q-network (VADQN) with a prior-based knowledge mechanism

In this section, we introduce our novel approach, VADQN, which is built upon the attention mechanism to address the perception problem in online DRL decision-making processes. The primary focus of this section is on the key innovations introduced in our proposed model, as illustrated in Fig. 2.

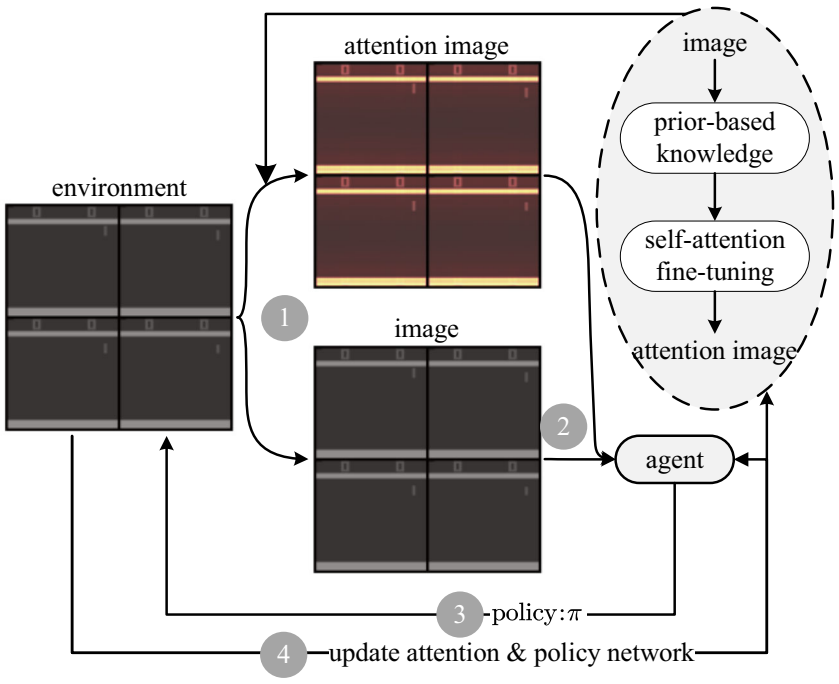


Fig. 2 In VADQN, the process begins with the extraction of attention maps through fine-tuning, which involves prior knowledge and SA. The agent subsequently cross-observes both the original image and the attention maps, utilizing policy π to facilitate exploration and exploitation. Finally, during the online training process, fine-tuning is conducted on the policy network, as well as the networks with prior knowledge and SA

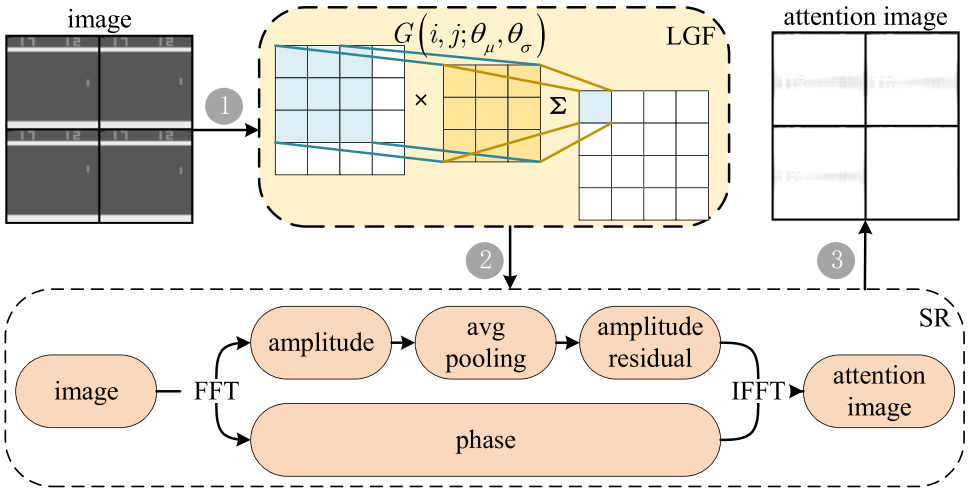


Fig. 3 Prior knowledge attention module diagram. An LGF is applied to the original image to improve its quality. Then, the frequency domain information is subjected to average pooling to extract neighbourhood residuals, thereby capturing the regions of prior attention in the image

3.1 Visual attention maps based on prior knowledge

Before humans receive any annotated information, they autonomously focus on key visual information obtained through vision, which is a result of the contrast between the salient information in the image and its surrounding environment. In light of this phenomenon, our work proposes a method based on learnable Gaussian filtering (LGF) and SR to extract the prior attention map, as shown in Fig. 3.

GF is a commonly used image processing technique that can blur an image. It is primarily used for smoothing images, denoising, and preprocessing before edge detection. In our work, we define an LGF module.

Given an input image $I_f(x, y)$, the module serves as a preprocessing step for the attention map to reduce the impact of image noise on the performance. The Gaussian kernel is defined as follows in (7).

$$G(x, y; \theta_\mu, \theta_\sigma) = \frac{1}{2\pi\theta_\sigma} \times \exp\left(-\frac{(x-\theta_\mu)^2}{2\theta_\sigma}\right) \times \exp\left(-\frac{(y-\theta_\mu)^2}{2\theta_\sigma}\right) \tag{7}$$

where the learnable parameters θ_μ and θ_σ represent the mean and variance, respectively, and $|x|$, $|y|$ denotes the size of the GF, which is set as B_k. $G(x, y)$ serves as the weight for the convolution operation, and the padding is set to $\lfloor \frac{B_k}{2} \rfloor$. B_k represents the size of the GF's kernel. The LGF module is defined as follows in (8).

$$\mathrm{GF}(I_{x,y}; \theta_\mu, \theta_\sigma) = \sum_{m=0}^{B_k-1} \sum_{n=0}^{B_k-1} G(m, n; \theta_\mu, \theta_\sigma) \times I(x+m, y+n) \tag{8}$$

In the LGF module, the mean and variance are learnable parameters, enabling adaptive GF for various input images. This provides better input for the SR module.

Defining the amplitude $A(u, v)$ and phase $\phi(u, v)$, inspired by (SR) [37], we obtain the frequency domain representation of the image using the fast Fourier transform (FFT) as $A(u, v), \phi(u, v) = F(I_f(x, y))$. We calculate the amplitude residual as shown in (9):

$$R(u, v) = \exp(\log(1 + A(u, v)) - M(\log(1 + A(u, v)), M_k)) \tag{9}$$

where $M(X, M_k)$ represents average pooling via a filter of size M_k, where the parameter M_k depends on the image size. Finally, we obtain the spatial domain attention map using the amplitude residual $R(u, v)$ and the phase $\phi(u, v)$ through the inverse fast Fourier transform (IFFT) as $I_f^a(x, y) = F^{-1}(R(u, v), \phi(u, v))$.

Because the attention maps generated by the LGF and SR modules do not yield satisfactory results, we fine-tune them through SA, utilizing attention maps on the basis of prior knowledge.

3.2 SA Fine-Tuning with prior knowledge

The attention maps obtained from the LGF and SR models still differ significantly from the attention maps generated by the deep model. Therefore, we introduce an SA mechanism based on prior attention maps to fine-tune the attention maps and obtain high-quality results, as shown in Fig. 4.

The spatial SA module maps the original image to three feature spaces, namely, K, Q, and V, through 2D convolutions, as shown in (10) [40].

$$K = f_K(I; \theta_K), Q = f_Q(I; \theta_Q), V = f_V(I; \theta_V) \tag{10}$$

where $f_K(\cdot)$, $f_Q(\cdot)$, and $f_V(\cdot)$ represent the convolutional mappings and θ_K, θ_Q, and θ_V denote the corresponding weights of the mappings. The attention weights are calcu-

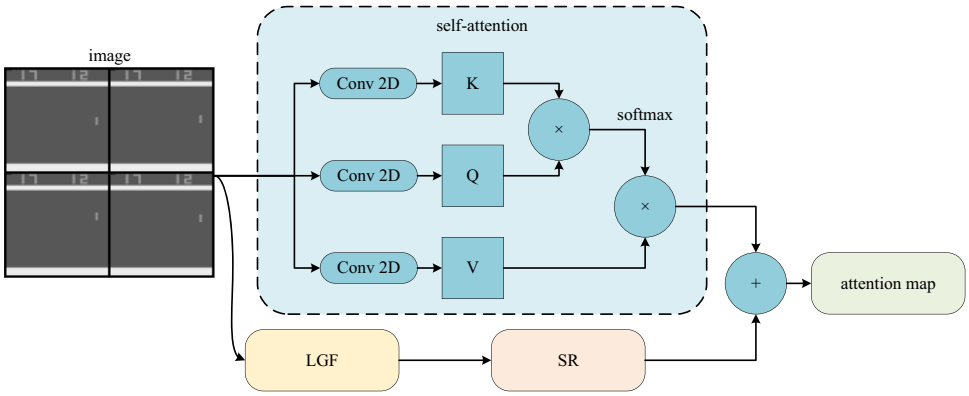


Fig. 4 SA fine-tuning with prior knowledge diagram. Prior attention maps are obtained through the LGF and SR modules. Subsequently, an attention map is extracted from the original image via the SA mechanism. These two attention maps are then fused through a weighted summation to obtain the final attention map

lated by taking the inner product of K and Q. Then, the attention weights are applied to the values V to obtain the final attention map O_s, as shown in (11).

$$O_s(I) = \text{softmax}(K\,Q^T)V \tag{11}$$

This allows the model to focus more strongly on the important parts of the input data, thereby improving the performance and representational capacity of the model. In summary, the spatial SA module with prior knowledge is composed as shown in (12).

$$\tilde{I} = \text{SR}(\text{GF}(I;\theta)) * O_s(I;\theta) \tag{12}$$

where $\tilde{I}$ represents the final attention map obtained. The parameter set θ consists of θ_μ, θ_σ, θ_K, θ_Q, and θ_V, all of which are jointly trained during the training process. Therefore, our proposed spatial SA module with prior knowledge can fine-tune the LGF and SR models by applying the SA mechanism. The SR model extracts attention maps from the image processed by the LGF, whereas the spatial SA module adaptively extracts attention maps from the original image. The two attention maps are then summed to achieve comprehensive attention map extraction. Therefore, we ultimately refer to this approach as LGFSRSA.

3.3 DQN framework with vision attention mechanism

DQN learns the parameters of the nonlinear approximator network through the temporal difference (TD) target during training, as shown in (13).

$$\mathcal{L}_{DQN}(\theta_d) = ||r + \gamma \max_{a' \in \{A\}} Q(s', a'; \theta_t) - Q(s, a; \theta_d)||_2 \tag{13}$$

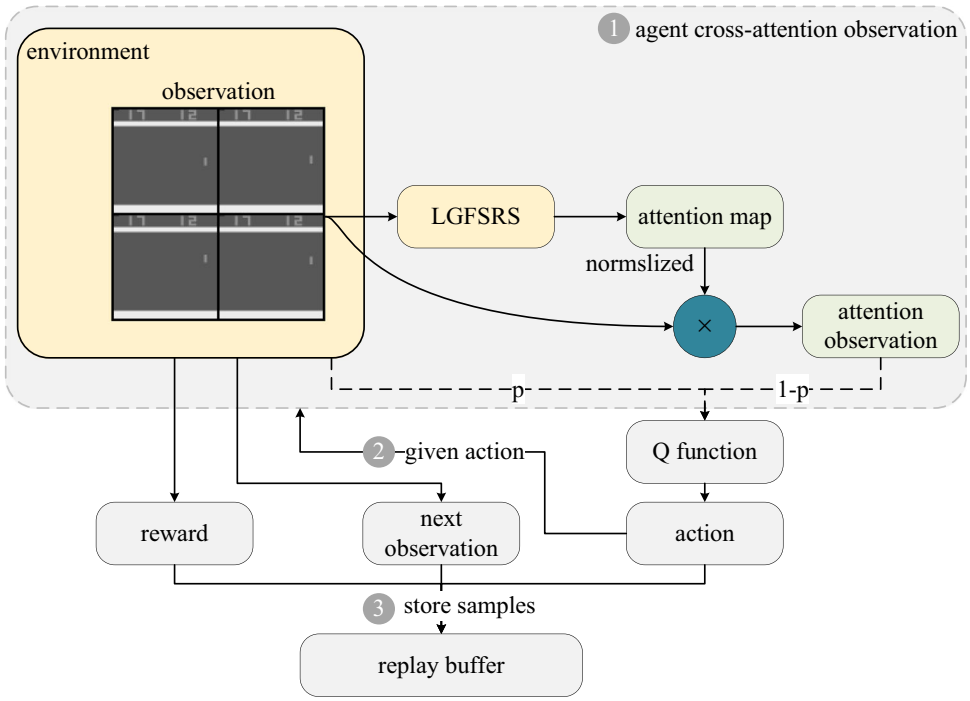


Fig. 5 The training process diagram of the VADQN algorithm is based on the DQN; it involves generating an attention visual map on the original image by using LGFSRSA to obtain an attention map. In the VADQN algorithm, with probabilities of p and $1-p$, either the attention visual map or the original image is randomly chosen as input for further processing

where the Q function $Q(\cdot)$ represents the evaluation of the agent performing action a under state s, A denotes the complete action space, and s' represents the next state after taking action a in state s. We define the elementwise multiplication of an image and its corresponding attention map as $\tilde{I}$. When making decisions, humans use the same policy as long as the task-relevant information flow is present, disregarding the existence of irrelevant information flows. Therefore, during inference, we alternate between I and $\tilde{I}$, as shown in (14).

$$\begin{aligned} &P(Q(I) = Q(I)) = p, \\ &P(Q(I) = Q(\tilde{I})) = 1 - p, \quad p \in [0, 1] \end{aligned} \tag{14}$$

where p represents the probability that $Q(I)$ is inferred directly from the original image and $1 - p$ represents the probability that $Q(I)$ is inferred from the weighted image obtained by elementwise multiplication of the image and attention map, as shown in Fig. 5.

$\tilde{I}$ retains task-relevant information flows while de-emphasizing irrelevant information flows. Consequently, we posit that the agent should receive the same Q value for both the original image I and $\tilde{I}$, i.e., $Q(I) = Q(\tilde{I})$. However, directly setting $Q(I) = Q(\tilde{I})$ would cause $\tilde{I}$ to approach I infinitely, resulting in all of the values of the attention map equal to one.

Therefore, during training, we combine the TD target from the original DQN with the Q value approximation from $\tilde{I}$, which trains the weights of the Q network, as shown in (15).

$$\mathcal{L}_1(\theta_d) = ||r + \gamma \max_{a' \in \{A\}} Q(\tilde{s}', a'; \theta_t) - Q(s, a; \theta_d)||_2 \tag{15}$$

The Q value from $\tilde{I}$ approximates the TD target from I, which trains the weights of the Q network as well as the weights of the spatial SA module with prior knowledge, as shown in (16).

$$\mathcal{L}_2(\theta_d) = ||r + \gamma \max_{a' \in \{A\}} Q(s', a'; \theta_t) - Q(\tilde{s}, a; \theta_d)||_2 \tag{16}$$

Since approximating the former would cause the TD target of the original image to keep changing, it encourages $\tilde{I}$ to continue receiving training. The overall framework is shown in Fig. 6.

Therefore, we improve the DQN loss as follows:

$$\mathcal{L}_{\text{VADQN}}(\theta) = \mathcal{L}_1 + \mathcal{L}_2 + (Q(I) - Q(\tilde{I})) \tag{17}$$

Overall, the combination of GF and the SA mechanism in our LGFSRSA model provides a powerful framework for robust saliency region detection and accurate attention mapping. These innovations enable our VADQN to outperform existing methods in various DRL tasks, highlighting the potential of our proposed approach in advancing the field of DRL. In summary, the pseudocode of VADQN is shown in Algorithm 1.

Algorithm 1 VADQN

Initialize replay memeory $\mathcal{D}$ to capacity N
Initialize Q network, target Q network, and visual attention weights, using Xavier initialization
for episode $= 1, M$ do **do**
  Initialize sequence $s_1 = x_1$ and preprocessed sequenced $\phi_1 = \phi(s_1)$
  for $t = 1, T$ do **do**
    With probability ε select a random action a_t
    otherwise select $a_t = \max_a Q^*(\phi(I), a; \theta)$ *// Compute the attention image from s to $\tilde{s}$ with (12), and use s or $\tilde{s}$ with (14).*
    Execute action a_t in the emulator and observe reward r_t and image x_{t+1}
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and preprocess $\phi_1 = \phi(s_{t+1})$
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathcal{D}$
    Sample random minibatch of transitions $(\phi_t, a_t, r_t, \phi_{j+1})$ form $\mathcal{D}$
    Set $y_j = \begin{cases} r_j & \text{forterminal}\phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta) & \text{fornon-terminal}\phi_{j+1} \end{cases}$
    Perform a gradient descent step on $(\phi_t, a_t, r_t, \phi_{j+1})$ with (17). *// Use VADQN loss.*
  end for
end for

4 Experiment

To validate the effectiveness of our proposed method, we conducted experiments using the Atari 2600 scenario as the testing environment. The observations were transformed into 84 × 84 grayscale images, following the DQN setup, where

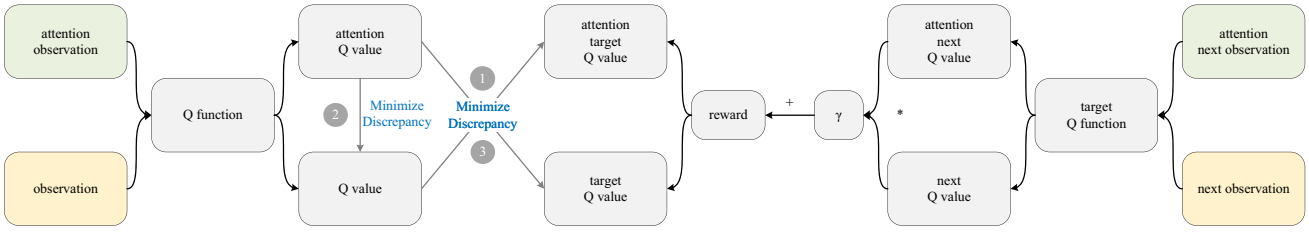


Fig. 6 The VADQN loss update method involves three steps: Update the attention Q value towards the target Q value according to (16). The Q value is updated towards the attention target Q value according to (15). Update the Q value towards the target Q value as $||Q(\tilde{I}) - Q(I)||_2$

Table 1 Experimental parameters

Parameter	Buffer size	Gamma	Learning rate	total step	batch size
Setting	1,000,000	0.99	1e-4	10,000,000	32

every 4 frames were stacked together. The feature extraction network used in our experiments consisted of three layers [4] of convolutional neural networks, and the Q network structure mirrored that of DQN. The agent interacted with the environment for 10 million exploration frames. The experiments were performed on two Xeon E5 2696v4 processors and an Nvidia GTX 3060. The buffer size was set to 1 million, gamma to 0.99, and the Adam optimizer with a learning rate of 1e-4 was employed. The exploration ratio began at 1.0 and gradually decreased to 0.05 during the initial 10% of the training process. In summary, to validate the effectiveness of our proposed method, we conducted experiments using the Atari 2600 scenario as the testing environment. We subsequently performed extensive ablation studies to assess the contributions of the introduced mechanisms. Furthermore, to better understand the impact of different mechanisms on training, we visualized the training processes involving these various mechanisms. Finally, to gain deeper insights into the agent's focus during decision-making, we visualized the attention maps. The experimental parameters are presented in Table 1.

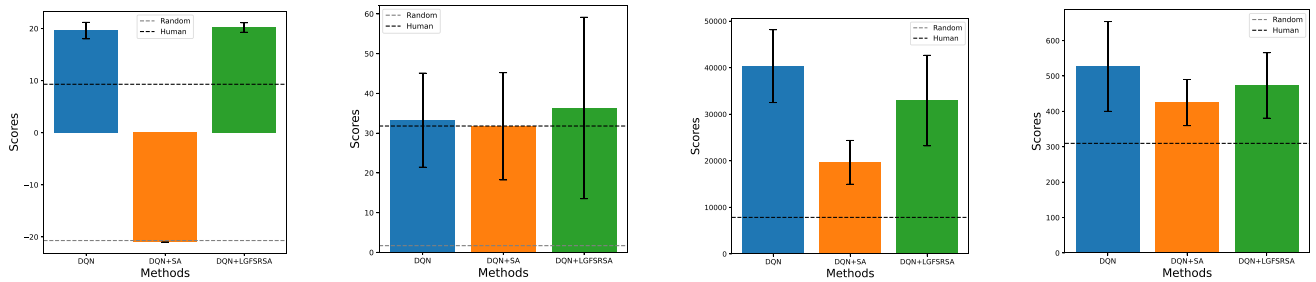
4.1 Performance comparison of the attention mechanisms

To validate the effects of priors on attention mechanisms, we introduced SA and prior-informed SA separately into both DQN and double DQN. We use the random score and the human score for validation. The experimental results are illustrated in Fig. 7.

From Fig. 7, it is evident that DQN+LGFSRSA generally outperforms the DQN and the DQN+SA across all games. This implies that the integration of priors into the attention mechanism can bolster the performance of DQN. However, the standard deviation for DQN+LGFSRSA is greater than those of the other methods, suggesting greater variability in its performance; this can be attributed to either the complexity of the games or the inherent variability of the attention mechanism. Comparison to human and random play shows that all DQN methods significantly outperform random play. However, performance relative to human play varies across games. For example, in 'Enduro', all DQN methods exceed human performance, whereas in 'RoadRunner', they fall short. In conclusion, the results underscore the potential of integrating attention mechanisms, particularly prior-informed self-attention, into DQN to increase its performance. However, further research is necessary to minimize the variability of the results and to boost the performance across all games.

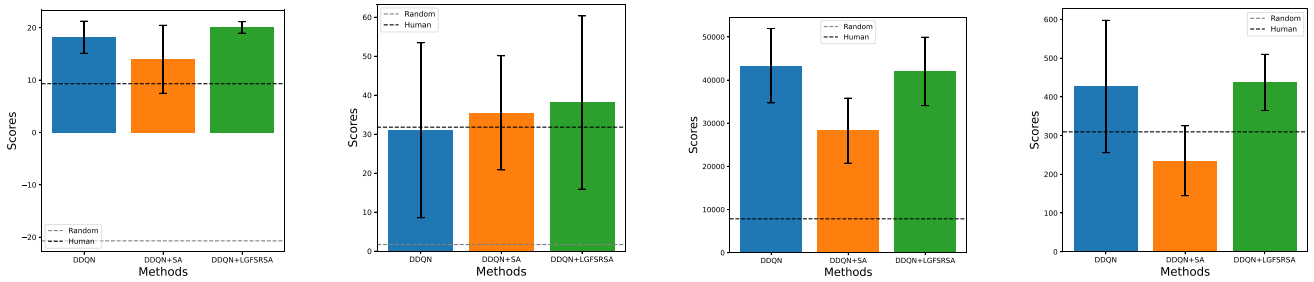
Figure 7 shows that the DQN+LGFSRSA method outperforms the other two methods in all games, with the highest score achieved in RoadRunner. The DQN+SA method performs better than the DQN method in Breakout and Pong but worse in Enduro and RoadRunner. These results suggest that prior-informed SA can enhance the performance of DQN.

Figure 8 shows that the DDQN+LGFSRSA method outperforms the other two methods in all games, with the highest score achieved in Pong. The DDQN+SA method performs better than the DDQN method in Breakout and Pong but worse in Enduro and RoadRunner. These results suggest that prior-informed SA can also enhance the performance of DDQN. In summary, the experimental results demonstrate that prior-informed SA can enhance the performance of both DQN and DDQN. The LGFSRSA method consistently outperforms the other two methods in all games, indicating that



(a) Pong (b) Breakout (c) RoadRunner (d) Enduro

Fig. 7 Performance comparison of attention mechanisms in the DQN. The graph depicts the mean scores for different methods, including DQN, DQN with SA (DQN+SA), and DQN with LGFSRSA (DQN+LGFSRSA). The error bars represent the standard deviations, providing insights into the stability of each method's performance. Additionally, dashed lines indicate the reference points for random, human and HMS performance in each game



(a) Pong (b) Breakout (c) RoadRunner (d) Enduro

Fig. 8 Performance Comparison of the Attention Mechanisms in DDQN. Same Setting as Fig. 7

it is a promising approach for incorporating priors into attention mechanisms.

To further validate the effectiveness of the proposed method, we conducted a comparison with SPR, using the HMS score as the evaluation metrics. The results are presented in Table 2.

The experimental results presented in Table 2 clearly indicate that the LGFSRSA method outperforms the SPR method across all tested environments. Specifically, in the Alien environment, LGFSRSA achieves a score of 822.3, which is slightly higher than the SPR score of 801.5. In Breakout, LGFSRSA improves the score to 18.2, whereas SPR's score is 17.1. The most significant improvement is observed in the Pong environment, where LGFSRSA not only improves on the negative score of SPR (-5.9) but also achieves a positive score of 2.2, indicating substantial performance enhancement. Additionally, in the RoadRunner environment, LGFSRSA demonstrates superior performance, with a score of 14410.1, compared with 14220.5 for SPR. Collectively, these results indicate that LGFSRSA provides more robust and effective learning capabilities, leading to better overall performance in vision-based DRL tasks.

4.2 Ablation study

In our VADQN method, multiple mechanisms are incorporated to enhance performance. To identify the individual impact of each mechanism, ablation experiments were conducted on the DQN. The results are summarized in Table 3.

An examination of the data presented in Table 3 shows that the introduction of SA alone leads to a decrease in performance, particularly in the Pong environment; this is due to the need for more identically distributed samples to train network weights, and the difficulty of making policy updates owing to the lack of a well-trained attention map arising from the constant changes in the policy and environmental content. When SR is introduced alone, there is a slight improvement in the performance in some games compared to SA. However, the performances declines compared to the method without the SA mechanism. This decline is attributed to the significant influence of the background and colour of the original image on the performance of the SR method's extracted feature maps. Combining SR with LGF (LGFSR) results in improved performance, indicating that LGF enhances the quality of image noise, making it effective. This improvement is evident in the comparison between LGFSRSA and SRSA. Compared to direct use of SA or SR, the SRSA method shows improvement in all environments, indicating that the method with the SR prior has significantly enhanced initial attention compared to SA. The fine-tuning of SA is also effective for the SR algorithm. Introducing the LGFSRSA mechanism leads to substantial performance improvement across multiple environments.

Table 2 Comparison of SPR with our method

method	Alien	Breakout	Pong	RoadRunner
SPR	801.5	17.1	-5.9	14220.5
LGFSRSA	822.3	18.2	2.2	14410.1

4.3 Training process and time complexity analysis

To further understand the impact of different mechanisms (LGFSRSA, LGFSR, SA, SR, SRSA, and SPR) on training, the training processes involving the introduction of various mechanisms are visualized in Fig. 9.

From Fig. 9, it is evident that LGFSRSA achieves higher scores than SA with the same number of steps, learning from fewer frames. Exploration rate reduction to 0 further emphasizes the improved performance of LGFSRSA compared to each attention mechanism.

$$\begin{aligned} &\text{SA} : O\left(C^2 \cdot H \cdot W + C \cdot HW \cdot HW\right) \\ &\text{LBF} : O\left(K^2 \cdot C \cdot H \cdot W\right) \\ &SR : O\left(C \cdot HW \log\left(HW\right) + C \cdot HWS\right) \end{aligned} \tag{18}$$

To further validate the effectiveness of the proposed method, we conducted an analysis of its time complexity and number of floating point operations (FLOPS). The

Table 3 The table presents the average variance of 100 test results after completing training

SA	SR	LGF	Breakout	Enduro	Pong	RoadRunner
✓			31.72 ± 13.50(100%)	424.87 ± 65.21(100%)	−21.0 ± 0(100%)	19675.0 ± 4697.67(100%)
	✓		31.29 ± 14.06(98.64%)	184.06 ± 45.78(43.32%)	13.07 ± 6.31(−%)	21755.0 ± 8383.23(110.57%)
✓	✓		33.70 ± 13.59(106.24%)	431.67 ± 65.91(101.60%)	19.2 ± 1.48(−%)	33334.0 ± 6218.86(169.42%)
	✓	✓	35.62 ± 14.40(112.30%)	271.31 ± 111.96(63.86%)	14.49 ± 5.44(−%)	33533.0 ± 15379.71(170.43%)
✓	✓	✓	36.34 ± 22.83(114.56%)	473.39 ± 92.34(111.42%)	20.2 ± 0.91(−%)	32936.0 ± 9727.61(167.40%)

The best and second-best results are highlighted in red and blue, respectively. The percentages denote the improvement compared to results obtained with the SA mechanism only

results, as shown in (18) and Table 4, where H represents the image height, W represents the image width, C represents the number of image channels, and K represents the size of the Gaussian convolution kernel, indicate that the LBFSR method incurs a minimal increase in the computational overhead, with 0.004 GFLOPs and 0.2 million parameters. LBFSRSA, which is slightly more complex, still maintains a relatively efficient computational profile with 0.024 GFLOPs and 6.2 million parameters. On the other hand, the SA method requires 0.02 GFLOPs and 6 million parameters. Notably, SR has no additional computational cost, as it does not introduce any new parameters or operations. These findings suggest that although LBFSRSA involves a slight increase in computational requirements, it remains competitive in terms of efficiency while delivering superior performance, as previously demonstrated in the experimental results. This balance between computational cost and performance further underscores the practical utility of the LGFSRSA method in real-world applications.

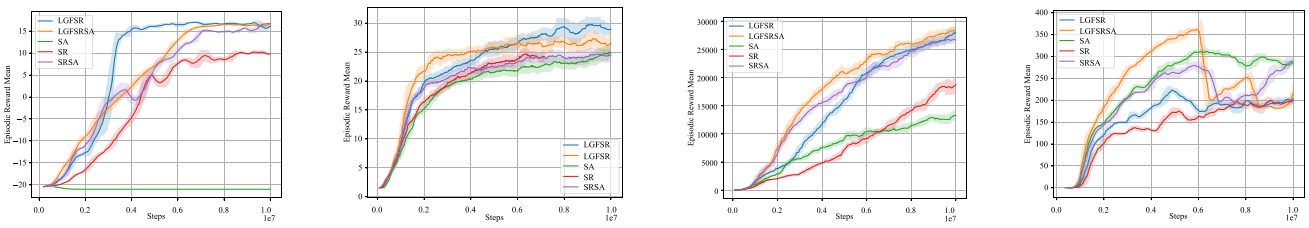
4.4 Visualization of attention results

To gain a deeper understanding of the agent's focus during the decision-making process, we visualized the attention maps, as shown in Fig. 10.

Figure 10 provides a comparative visualization of the attention maps generated by different mechanisms in the Pong and Breakout environments. In the Pong environment, the results indicate that the SA mechanism, when used in isolation, struggled to focus on the dynamic and relevant aspects of the environment. This limitation suggests that on its own, SA may not effectively capture the crucial elements necessary for optimal decision-making. However, a noticeable improvement was obtained when we incorporated the prior attention of the SR mechanism. The enhanced SA, informed by SR, was able to more accurately attend to the dynamic parts of the data, highlighting the value of integrating prior knowledge into the attention process.

In the Breakout environment, the attention maps generated by both SA and LGFSRSA were quite similar, suggesting that SA is inherently capable of learning attention maps that are relevant to the task at hand. Nevertheless, the LGFSRSA mechanism, which incorporates the LGF preprocessing step, produced slightly refined attention maps. This refinement underscores the effectiveness of the LGF mechanism in enhancing the saliency of critical regions in the images, which aids in the generation of more accurate attention maps.

Despite the similarity in the attention maps between SA and LGFSRSA in the Breakout environment, it is noteworthy that LGFSRSA and SRSA outperformed the SA alone. This improvement highlights the enhanced generalization ability of the SR mechanism. The SR mechanism contributes by delivering baseline attention maps that prevent significant errors in attention allocation, particularly in situations involving unknown or unfamiliar distributions. This is akin to human experience, where prior knowledge helps to avoid complete misjudgment when encountering novel stimuli.



(a) Pong (b) Breakout (c) RoadRunner (d) Enduro

Fig. 9 Training process visualization. The horizontal axis represents the number of frames, and the vertical axis represents the average reward. Lines represent the mean within a window step, while the shaded area represents the variance of the 95% confidence interval

Table 4 Number of floating point operations

method	Flops(G)	Params(M)
SR	0	0
LBFSR	0.004	0.2
LGFSRSA	0.024	6.2
SA	0.02	6

In summary, the analysis of attention maps reveals crucial insights into how different attention mechanisms impact the agent's decision-making process. The integration of prior-informed SA into the DQN framework not only enhances the accuracy of attention but also significantly improves generalization across varying environments. These findings demonstrate the potential of combining structured prior knowledge with DRL models to achieve better performance in complex and dynamic tasks.

5 Conclusion

In this study, we propose the VADQN algorithm to explore how DRL perceives and processes information during decision-making. Our approach aims to address the challenges of error accumulation and exploration difficulties caused by the instability of attention maps when they are integrated directly into network updates. Specifically, LGFSRSA leverages the LGF to filter out image noise, reducing the amplitude spectrum and details to enhance the salient regions in the images. Additionally, we implemented an SA mechanism to fine-tune the attention maps generated by SR, reducing interference from higher-level attention tasks and improving the generalization ability of DRL.

We validated the effectiveness of LGFSRSA through experiments in Atari games, where the results demonstrated that LGFSRSA not only outperforms SA with the same number of steps but also requires fewer frames for learning. Furthermore, comparisons with DQN and other attention mechanisms confirmed the robustness and efficiency of LGFSRSA.

Our findings significantly contribute to understanding and addressing the issue of unlearnability in DRL caused by changes in the exploration policies due to attention mechanisms. By revealing the intrinsic perceptual mechanisms of DRL models during decision-making, our research lays a solid foundation for advancing intelligent decision-making systems in complex, nonlinear environments. Future research can focus on the further optimization of the LGFSRSA attention mechanism by integrating more prior knowledge and advanced attention methods to enhance the performance and

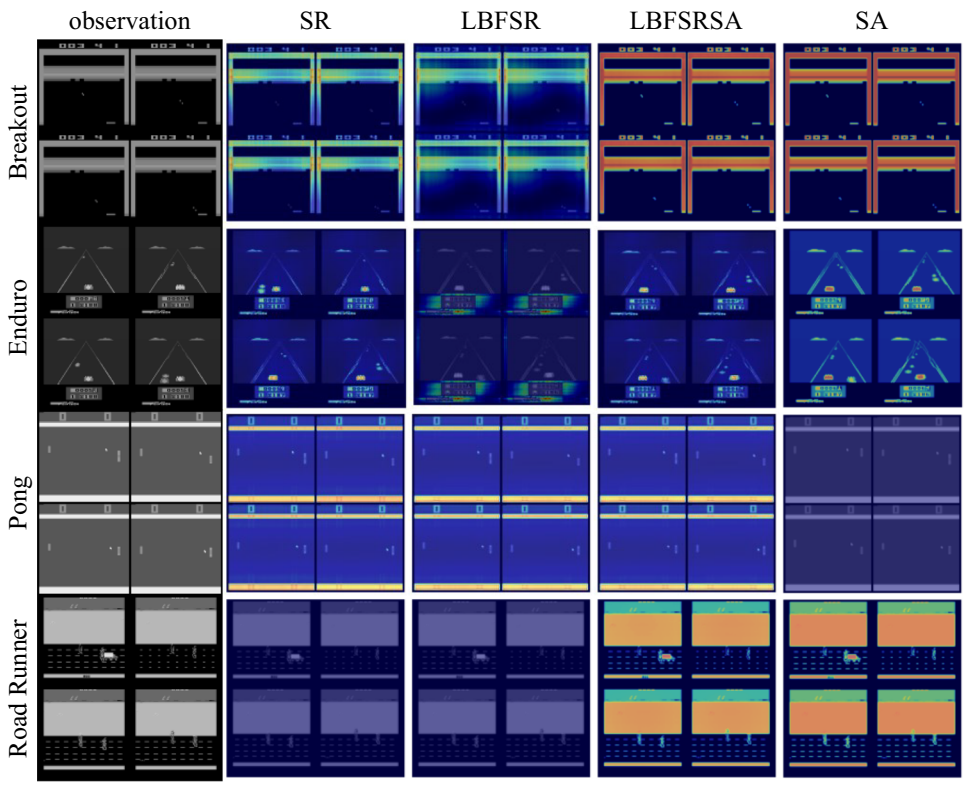


Fig. 10 Attention maps of various mechanisms in different environments. The attention maps in the figure represent different mechanisms in diverse environments. Transparent attention heatmaps are overlaid on the original image, where hotter regions indicate higher weights assigned by the attention mechanism to those areas

generalization of vision-based DRL in continuous decision-making environments.

Author Contributions All authors contributed to the study conception and design. Jialin Ma and Ce Li completed the data analysis and code writing. Jialin Ma, Liang Hong, Kailun Wei, Shutian Zhao, and Hangfei Jiang designed and drafted the manuscript, and Ce Li and Yanyun Qu revised the paper. All authors read and approved the manuscript.

Funding The work of this thesis was partially funded by the National Natural Science Foundation of China under Grant (No. 62363025), the Education Industry Support Plan Project of Gansu Provincial Department (No. 2023CYZC-26), the Key R&D plan of Science and Technology Plan of Gansu Province - Social Development Field Project (No. 23YFFA0064), the Gansu Education Department, China (No. 2023CXZX-468) and the Science and Technology on Vacuum & Cryogenics Technology and Physics Laboratory (No. 61422072305).

Availability of data and materials The environment data in the OpenAI Gym repository, accessible at https://www.gymlibrary.dev/environments/atari/complete_list/, and the materials associated with the present study are obtainable from the corresponding author upon reasonable request.

Declarations

Competing interests The authors have no relevant financial or nonfinancial interests to disclose.

References

1. Kim K, Ha J, Kim Y (2022) Self-predictive dynamics for generalization of vision-based reinforcement learning. In: Proceedings of the Thirty-First international joint conference on artificial intelligence, pp 3150–3156
2. de Morais GAP, Marcos LB, Bueno JNAD, de Resende NF, Terra MH, Grassi V Jr (2020) Vision-based robust control framework based on deep reinforcement learning applied to autonomous ground vehicles. Control Eng Pract 104:104630
3. Watkins CJCH, Dayan P (1992) Q-learning. Mach Learn 8(3–4):279–292
4. Mnih V, Kavukcuoglu K, Silver D (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533
5. Yuan J, Zhu A, Xu Q, Wattanachote K, Gong Y (2023) Ctif-net: A cnn-transformer iterative fusion network for salient object detection. IEEE Trans Circuits Syst Video Technol
6. Yan R, Yan L, Geng G, Cao Y, Zhou P, Meng Y (2024) Asnet: Adaptive semantic network based on transformer-cnn for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens
7. Gao L, Liu B, Fu P, Xu M (2023) Adaptive spatial tokenization transformer for salient object detection in optical remote sensing images. IEEE Trans Geosci Remote Sens 61:1–15
8. Shi WJ, Huang G, Song SJ (2022) Self-supervised discovering of interpretable features for reinforcement learning. IEEE Trans Pattern Anal Mach Intell 44(5):2712–2724
9. Huang W, Zhang C, Wu J, He X, Zhang J, Lv C (2023) Sampling efficient deep reinforcement learning through preference-guided stochastic exploration. IEEE Trans Neural Netw Learn Syst 1–12
10. Zhu JH, Xia YC, Wu LJ (2023) Masked contrastive representation learning for reinforcement learning. IEEE Trans Pattern Anal Mach Intell 45(3):3421–3433
11. Shang J, Kahatapitiya K, Li X (2022) Starformer: Transformer with state-action-reward representations for visual reinforcement learning. In: European conference on computer vision, pp 462–479
12. Tang J, Mihailovic A, Aghvami H (2022) Constructing a drl decision making scheme for multi-path routing in all-ip access network. In: IEEE Global communications conference, IEEE, pp 3623–3628
13. Xing JW, Nagata T, Zou XY (2023) Achieving efficient interpretability of reinforcement learning via policy distillation and selective input gradient regularization. Neural Netw 161:228–241
14. Pham NT, Dang DNM, Nguyen ND, Nguyen TT, Nguyen H, Manavalan B, Lim CP, Nguyen SD (2023) Hybrid data augmentation and deep attention-based dilated convolutional-recurrent neural networks for speech emotion recognition. Expert Syst Appl 230:120608
15. Fathnejat H, Ahmadi-Nedushan B, Hosseininejad S, Noori M, Altabey WA (2023) A data-driven structural damage identification approach using deep convolutional-attention-recurrent neural architecture under temperature variations. Eng Struct 276:115311
16. Yasutomi AY, Ichiwara H, Ito H, Mori H, Ogata T (2023) Visual spatial attention and proprioceptive data-driven reinforcement learning for robust peg-in-hole task under variable conditions. IEEE Robot Automat Lett 8(3):1834–1841
17. Zheng Y, Lin Y, Zhao L, Wu T, Jin D, Li Y (2023) Spatial planning of urban communities via deep reinforcement learning. Nat Comput Sci 3(9):748–762
18. Liu Y, Yang D, Zhang F, Xie Q, Zhang C (2024) Deep recurrent residual channel attention network for single image super-resolution. Visual Comput 40(5):3441–3456
19. Sun H, Li B, Dan Z, Hu W, Du B, Yang W, Wan J (2023) Multi-level feature interaction and efficient non-local information enhanced channel attention for image dehazing. Neural Netw 163:10–27
20. Hou Y-E, Yang K, Dang L, Liu Y (2023) Contextual spatial-channel attention network for remote sensing scene classification. IEEE Geosci Remote Sens Lett
21. Pan X, Ye T, Xia Z, Song S, Huang G (2023) Slide-transformer: Hierarchical vision transformer with local self-attention. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pp 2082–2091
22. Mehrani P, Tsotsos JK (2023) Self-attention in vision transformers performs perceptual grouping, not attention. Front Comput Sci 5:1178450
23. Bilal A, Liu X, Shafiq M, Ahmed Z, Long H (2024) Nimeq-sacnet: A novel self-attention precision medicine model for vision-threatening diabetic retinopathy using image data. Comput Biol Med 171:108099
24. Sorokin I, Seleznev A, Pavlov M (2015) Deep attention recurrent q-network. arXiv preprint arXiv:1512.01693, 1–7
25. Liu Y, Wang X, Chang Y (2022) Towards explainable reinforcement learning using scoring mechanism augmented agents. In: Knowledge science, engineering and management, proceedings, Part II, pp 547–558
26. Ma X, Zhang S, Wang Y, Li R, Chen X, Yu D (2023) Ascam-former: Blind image quality assessment based on adaptive spatial & channel attention merging transformer and image to patch weights sharing. Expert Syst Appl 215:119268
27. Zhang S, Liu Z, Chen Y, Jin Y, Bai G (2023) Selective kernel convolution deep residual network based on channel-spatial attention mechanism and feature fusion for mechanical fault diagnosis. ISA Trans 133:369–383
28. Wang Y, Shi K, Lu C, Liu Y, Zhang M, Qu H (2023) Spatial-temporal self-attention for asynchronous spiking neural networks. In: Thirty-Second international joint conference on artificial intelligence, vol. 8, pp 3085–3093
29. Zhao Y, Luo C, Tang C, Chen D, Codella N, Zha Z-J (2023) Streaming video model. In: IEEE/CVF Conference on computer vision and pattern recognition, pp 14602–14612

30. Feng W, Xu N, Zhang T, Zhang Y, Wu F (2024) Enhancing cross-task transferability of adversarial examples via spatial and channel attention. IEEE Trans Multimed
31. Hassanin M, Anwar S, Radwan I, Khan FS, Mian A (2024) Visual attention methods in deep learning: An in-depth survey. Inf Fusion 108:102417
32. Justesen N, Bontrager P, Togelius J, Risi S (2020) Deep learning for video game playing. IEEE Trans Games 12(1):1–20
33. Hasselt Hv, Guez A, Silver D (2016) Deep reinforcement learning with double q-learning. In: Thirtieth AAAI conference on artificial intelligence, pp 2094–2100
34. Hessel M, Modayil J, Hasselt H, Schaul T, Ostrovski G, Dabney W, Horgan D, Piot B, Azar M, Silver D (2018) Rainbow: Combining improvements in deep reinforcement learning. In: Thirty-Second AAAI Conference on artificial intelligence and thirtieth innovative applications of artificial intelligence conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, pp 11796–11804
35. Yarats D, Kostrikov I, Fergus R (2021) Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In: International conference on learning representations, pp 1–22
36. Schwarzer M, Anand A, Goel R, Hjelm RD, Courville A, Bachman P (2021) Data-efficient reinforcement learning with self-predictive representations. In: International conference on learning representations, pp 1–18
37. Hou X, Zhang L (2007) Saliency detection: A spectral residual approach. In: IEEE Conference on computer vision and pattern recognition, pp 1–8
38. Li Y, Sycara K, Iyer R (2017) Object-sensitive deep reinforcement learning. In: Benzmüller C, Lisetti C, Theobald M (eds.) GCAI 2017. 3rd Global conference on artificial intelligence, vol. 50, pp 20–35. EasyChair, Miami, USA. 18-22
39. Goel V, Weng J, Poupart P (2018) Unsupervised video object segmentation for deep reinforcement learning. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in Neural Information Processing Systems, vol 31. Curran Associates Inc, Montréal, Canada
40. Vaswani A, Shazeer N, Parmar N (2017) Attention is all you need. In: International conference on neural information processing systems, pp 6000–6010

Jialin Ma received the B.S. degree from the Bohai University, Jinzhou, China, in 2018, and the M.S. degree from the Lanzhou University of Technology, Lanzhou, China, in 2022, where he is currently pursuing the Ph.D. degree. The main research directions are reinforcement learning, computer vision and pattern recognition.

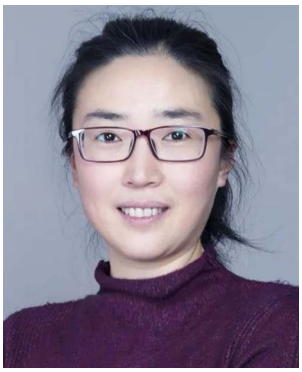
Ce Li received the Ph.D. degree in pattern recognition and intelligence system from Xi'an Jiaotong University, Xi'an, China, in 2013. He is currently a Professor with the College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, China. His research interests include computer vision and pattern recognition.

Liang Hong received the B.S. degree from the Zhoukou Normal University, Henan, China, in 2021, and is currently pursuing a master's degree at Lanzhou University of Technology. The main research directions are computer vision, salient object detection, and brain-like computing.

Kailun Wei received the B.S. degree from the Henan Institute of Technology, Henan, China, in 2021, and is currently pursuing a master's degree at Lanzhou University of Technology. The main research directions are gaze estimation, and computer vision.

Shutian Zhao received the B.S. degree from the Jiangsu University of Jingjiang College, Jiangsu, China, in 2021, and is currently pursuing a master's degree at Lanzhou University of Technology. The main research directions are computer vision, and image derain.

Hangfei Jiang received the B.S. degree from the Shenyang Normal University, Shenyang, China, in 2022, and is currently pursuing a master's degree at Lanzhou University of Technology. The main research directions are computer vision, and saliency object detection.



Yanyun Qu received the Ph.Ddegree in pattern recognition and intelligent systems from the Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China, in 2006. She is currently a Professor with the Department of Computer Science School of Informatics, Xiamen University, Xiamen, China. She has authored and coauthored over 130 papers in major international journals and conferences, including the International Journal of Computer Vision, the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CYBERNETICS, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, Pattern Recognition, the IEEE International Conference on Computer Vision, the IEEE Conference on Computer Vision and Pattern Recognition, European Conference on Computer Vision, National Conference on Artificial Intelligence (AAAI), International Joint Conferences on Artificial Intelligence, Association for Computing Machinery (ACM), International Conference on Multimedia, and International Conference on Acoustics, Speech and Signal Processing. Her current research interests include image processing, computer vision, machine learning, and pattern recognition. Dr. Qu is a member of ACM and the Secretary of the Technical Committeeof Hybrid Artificial Intelligence and Chinese Association of Automation.