Bilad Alrafidain Journal for Engineering Science and Technology
https://dx.doi.org/10.56990/bajest/2026.050108
ISSN: 2073-9524
Pages:87-99

Deep fake Detection via Active Perception: A Deep Q-Learning Framework for Interpretable Visual Forensics

Khalid Jamal Jadaa
Computer Engineering Department, College of Engineering, University of Diyala, Diyala, 3100, Iraq
khalid.jamal.jadaa@gmail.com

Abstract
The rapid development of generative models has resulted in the widespread synthesis of realistic deepfake media, and thus raises fundamental threats to digital trust and media verification. The majority of current deepfake detection methods are based on supervised convolutional neural networks (CNNs) and work with global image presentations, which may easily have a weak generalization ability and lack interpretive power. This paper presents a different paradigm for the detection of deepfakes: to cast the problem as a sequential decision and adopt reinforcement learning to solve it. For that purpose, the paper proposes a patch-based Deep Q-Learning (DQN) approach that enables the agent to selectively explore local face regions and detect manipulation artifacts. Unlike the typical feedforward classifiers, this approach is capable of fine-grained spatial exploration as well as making the model interpretable: it highlights regions that are informative for the decision. This research is offered as a pilot study to examine feasibility and interpretability rather than sample size and benchmarking. Experiments were performed on balanced partial deepfake image datasets released for public use (2400 images), and each image was considered as a single RL episode. Experimental results on a practical deepfake dataset show that the proposed approach has good performance with an AUC of 0.92. Remarkably, despite processing only 18.2% pixels on average per image, the proposed approach achieves this level of performance, confirming the efficiency and forensic potential of the proposed active perception framework.

Keywords: Deepfake Detection, Reinforcement Learning, Deep Q-Learning, Explainable AI, Visual Forensics, Patch-Based Analysis

Article history: **Received:** 13 Jan 2026**, Accepted:** 15 Feb **, Published:** 15 Mar 2026**.**

1. Introduction

Massive advances in recent years, especially in generative models, e.g., GANs (Generative Adversarial Networks), have allowed the production of incredibly realistic synthetic facial media, so-called deepfakes. Such technologies have useful purposes in entertainment, accessibility, and digital content creation, but also evoke significant concerns about misinformation, identity fraud, and online trust. The improved authenticity of synthesized face images and videos poses a significant challenge to the reliable detection of deepfakes in multimedia forensics and computer vision [1-3]. To this end, there is a large literature on supervised deep learning methods, often relying on convolutional neural networks (CNNs). These methods generally consider deepfake detection as a static binary image/video classification task where the whole image or video frame is processed in a single forward pass [4]. Even though CNN detectors achieve high performance on the widely used datasets, there are still limitations preventing them from being put into practice. First, they are not effective when it comes to contrivance methods that they have never seen, showing poor generality. Second, such models generally involve a large-scale labeled dataset, which is costly to obtain. Third, many CNN-based detectors are black-box and

ISSN: 2073-9524
Pages:87-99

provide little explanation in terms of what facial parts or visual cues led to the final decision [5]and [6].

Recent studies have tried to address these issues by introducing attention mechanisms, frequency domain analysis, or patch-based feature extraction. Although these methods enhance spatial specificity, they depend on supervised learning and predefined pipelines for processing. Therefore, they are non-intelligent and cannot choose by themselves where to pay attention in an image based on learned experiences. On the other hand, RL provides an alternative learning paradigm by regarding perception tasks as "sequential decision-making" problems [7]. Unlike traditional methods, which passively process the entire input, an RL agent learns to interact with the environment, make a decision on interesting regions, and then update its policy according to the reward. RL has shown remarkable success on multiple vision tasks, including active object localization and visual attention modeling that require selective exploration. However, little has been found related to deepfake detection [8] and [9].

Based on this motivation, the research introduces the formulation of deepfake detection as a problem of interactive exploration, rather than a static classification. This paper proposes a patch-based Deep Q-Learning (DQN) network, where an agent sequentially examines subregions of the face and receives feedback through rewards to detect manipulation artifacts. By working at the patch level, the agent can concentrate on high-signal regions such as eyes, mouth, and facial boundary areas where forgery artifacts may arise readily. This formulation has two distinct advantages. Firstly, this improves interpretability due to the fact that it explicitly indicates which parts of the regions are used for detection decisions, leading to transparency and reliability. The second reason is that both the sequential and interactive process of learning has better adaptability, which enables the model to generalize outside some particular manipulation patterns observed in training.

1. Re-formulate the task of deepfake detection as a sequential decision-making problem and propose to actively explore, via adaptive facial region look-up, rather than passively classifying global images.
2. Suggest a patch-based Deep Q-Learning (DQN) strategy, which automatically learns an exploration policy to detect local manipulation artifacts via interaction based on rewards.
3. Propose a combined reward to balance detection confidence (to encourage moving to regions with high likelihood of finding objects), uncertainty reduction (to restrict exploring in easy-to-discover areas), and exploration efficiency, which can obtain interpretable patch-level scan paths suitable for forensic studies.
4. Experimental results show the proposed method achieves a comparable AUC of 0.92 by processing only 18.2% image pixels, which verifies that reinforcement learning is an effective and efficient tool for deepfake detection.

The rest of the paper is structured as follows. In Section 2, we discuss the related work of deepfake detection algorithms and reinforcement learning applied to visual analysis. Section 3 describes the proposed DQL-based detection process, including the construction of the environment and reward design. The experimental setup and evaluation results are presented in Section 4. Section 5 contrasts the results, Section 6 discusses limitations, and is followed by conclusions and future work in Section 7.

2. Related Works

Deepfake detection has attracted a considerable amount of interest in the past few years, especially in the areas of computer vision and multimedia forensics. Current approaches tend to fall into the following categories: 1) supervised deep learning-based strategies; 2) temporal and physiological analysis-based methods; 3) attention or patch-based models; and 4) reinforcement learning-guided visual analytics.

2.1. Supervised CNN-Based Deepfake Detection

Most of the existing deepfake detection methods are based on supervised convolutional neural networks (CNNs) that learn from labeled real and generated datasets. MesoNet is a typical example of CNN-based methods for detecting LIAAF using compact models at the mesoscopic level [2]. Successive works employed larger networks like XceptionNet and

ISSN: 2073-9524
Pages:87-99

high-capacity CNNs for the benchmark dataset Face Forensics++ [10].

However, CNN-based detectors also have some significant limitations despite their superiority in challenging scenarios. They have mediocre generalization to new kinds of manipulation techniques or datasets. In addition, these models generally have dependence on large-scale labelled data and lack interpretability: the classification result is based on global feature representation without the localized location of forgery evidence [11].

2.2. Temporal and Physiological Cue-Based Methods

Beyond static image analysis, other works have considered this phenomenon in videos with RNNs and LSTM networks to perform temporal consistency analysis. These approaches take advantage of the temporal artifacts, including frame-level inconsistency and motion noise [12]. Concurrently, physiology-based techniques study signals like eye blinking patterns or weak heart rate variance derived from facial videos [13]. Although useful in certain cases, these methods are inherently specialized for video data and often need clean, temporally coherent input. However, they rely on a clean recording scenario and well-controlled light conditions, so they may not apply to real-world data where assumptions might not hold.

2.3 Patch-Based and Attention-Driven Models

Recent work has further combined patch processing and attention mechanisms, which aim to enhance spatial sensitivity for deepfake detection. By concentrating on local regions, e.g., around the eyes, mouth, or facial boundaries, these methods expect to capture subtle manipulation artifacts that are lost in global representations. For instance, explicit artifact-guided attention network forcing detectors to attend the corrupted pixel regions for increasing generalization across the unseen manipulations has been presented in [14]. Attention Mechanism ViT-based approaches have also investigated vulnerability-guided patch attention dedicated to capturing localized prone information on the level of ordinary transformer [15]. Global subnets that aggregate features across all pixels and always handle local interaction diffusely are robust, while paying attention to the underlying global context before decision has been a success for instance-level tasks [16]. However, the existing works are essentially supervised-based and model attention mechanisms in a predefined manner or select patches in a fixed way. Consequently, they lose the ability to dynamically prioritize their attentive positions by learning from experience or feedback [17], thereby affecting their adaptability to different or changeable manipulation patterns.

Although the development of Vision Transformers (ViT) has just brought in patch-level processing with self-attention, the way they express this idea is fundamentally different from ours. The standard ViTs perform global self-attention over the entire image across all patches, but this is still computationally very expensive on high-resolution inputs. For example, recent patch-based ViT models for synthetic media achieve good performance by embedding image patches and learning the global relation between them, but they need much compute and large training data.

Furthermore, self-supervised ViT variants have been attempted to enhance generalization, but they have high requirements in terms of resources and do not focus adaptively on discriminative regions. Extensions like attention guidance methods have been designed to facilitate concentrating on artifact-prone patches, but these still reside in the global transformer framework and do not intrinsically prioritize sequential exploration. On the other hand, the Reinforcement Learning model considers patch selection as a sequential exploration process. By discovering a policy to attend selectively to the most subjective regions, the approach can act closer in spirit to ViTs, with much cheaper image coverage and computation, by transitioning from “passive attention” to “active goal-driven perception [18] [20].

2.4 Reinforcement Learning for Visual Decision-Making

Reinforcement learning (RL) has demonstrated strong performance in various vision tasks that need active perception, such as object localization, visual attention modeling, and interactive image analysis. In such environments, RL agents learn policies that further the sequential parsing of visual input to perform efficient and interpretable decision-making

ISSN: 2073-9524
Pages:87-99

[21]. Although reinforcement learning has shown success in the related fields, its use for deepfake detection is still at an early stage. Current deepfake detectors do not treat detection as an interactive exploration task and use reward-based feedback to steer the visual attention [22].

2.5 Research Gap and Motivation

From observation of the literature review, state-of-the-art deepfake detectors are still built upon static end-to-end supervised classification pipelines that convolve entire frames or images in one go. Patch-based and attention-based models only partially take spatial localization into account, but do not include active exploration capability and adaptability. This poses a gap that motivates casting the deepfake detection as sequential decision-making. Through reinforcement learning, an agent can learn where to look and how to explore information-rich areas over space, and how to aggregate the evidence over time. The proposed framework, based on deep Q-learning, therefore attempts to address this gap by presenting a rule-driven, interpretable personalized detection model that can be used in conjunction with existing supervised strategies.

3. Methodology

In this section, we present the proposed deepfake detection framework with DQNs. This approach is different from classical supervised classifiers, where the complete image is fed to the model for a single pass in a forward direction; instead, the model views deepfake detection as a sequence of decision-making processes. An agent explores localized facial regions actively and learns a policy that drives it to the informative areas with manipulation artifacts.

The formulation of deepfake detection as a Markov Decision Process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma)$, where:

- $\mathcal{S}$ represents the state space corresponding to image patches,
- $\mathcal{A}$ Denotes the action space of possible patch selections,
- $\mathcal{R}$ is the reward function guiding the agent's O O, $\mathcal{T}$ defines the state transition dynamics,$The\ \gamma \in [1]$ is the discount factor.

When given an input facial image, the agent interacts with the environment by iteratively picking image patches and receiving feedback corresponding to a probability that forgery artifacts lie in the chosen regions. The agent aims to discover a good exploration policy, which results in the maximization of the cumulative reward as well as the performance of real/fake image classification.

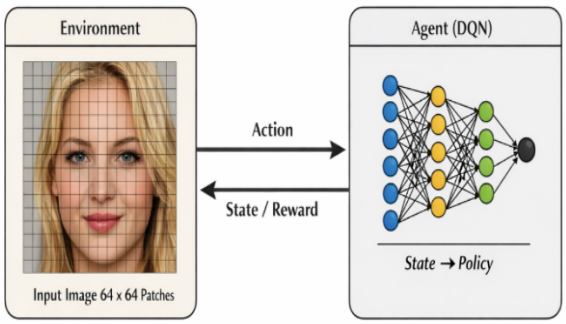

Fig. 1 The proposed Deep Q-learning framework for deepfake detection [23]

Fig.1. shows the block diagram of the proposed deepfake detection framework based on reinforcement learning. It is considered an interaction loop between an environment and an intelligent agent. The domain is the input facial image pre-partitioned into a fixed grid of non-overlapping 64×64 Pixel patches constituting the set of discrete spatial positions over which exploration can be conducted (at any one time). As illustrated, the Agent (here: a Deep Q-Network (DQN)) picks an Action given by picking that specific patch layer index of the grid according to its learned policy at time.

In response, the environment gives back a state observation, which is an image corresponding to the visual content of the patch chosen and a scalar reward. This reward signal is specifically targeted to incentivize the agent's focus on "high-signal" facial areas, such as eyes or mouth, where manipulation artifacts are more likely to be produced. Along this sequence of decisions, the agent gains insights into how to optimally traverse its exploration space to effectively differentiate fake from real media, with an interpretable “scan path” of its investigative actions. The problem is posed as a Markov decision process, where a DQN agent makes 64×64 patch selections and gets rewards according to the exhibit of artifacts, such that an optimal scanpath focusing on information-dense regions rather than whole-image attention is learned.

ISSN: 2073-9524
Pages:87-99

3.2 Environment Design

In this research, treat each facial image as a single instant of the environment. At interaction onset, the image is sliced into a static grid of non-overlapping patches of 64x64 pixels. These patches are the discrete spatial locations where the agent has access. At each time step t:

- The agent observes thess_t, represented by the visual content of the selected patch.
- The agent chooses a catenate a_t, corresponding to the selection of a new patch location.
- The environment transitions to a new state nd retns a reward (r_t).

An episode ends when we have explored a fixed number of patches or the agent arrives at a terminal decision step. This formulation allows the agent to collect evidence gradually from spatial regions, rather than keeping track of a single global observation.

3.3 Action Space and Exploration Strategy

The action space $\mathcal{A}$ consists of all valid patch indices within the image grid. To balance exploration and exploitation during training, a ϵ- A greedy strategy is employed. With probability ϵ, the agent selects a random patch to encourage exploration, and with probability $heagen$t selects the patch with the highest estimated Q-value. The value of ϵ gradually decays over training episodesallowing , the agent to transition from exploratory behavior to policy-driven exploitation.

3.4 Deep Q-Network Architecture

The Deep Q-Network (DQN) serves as a function approximator for the action-value functionQ (s a) , $Q(s, a)$. The network consists of:

- A convolutional feature extractor that processes the visual content of each patch,
- Fully connected layers that map extracted features to Q-values for each possible action.

To generate compact and discriminative visual representations for each of the selected image patches, we adopt a pretrained CNN as a fixed feature extraction backbone. Concretely, for each 64×64 patch, we feed it through a ResNet-18 model pre-trained on ImageNet and extract the output of the last global average pooling layer as the patch-level feature embedding. The backbone is frozen during RL training to stabilize the features and disentangle representation learning from policy optimization. This design choice can be in order to mitigate training instability often observed with E2E reinforcement learning from raw pixels, while the DQN agent focuses only on learning a good exploration policy, not low-level vision. The derived feature embeddings are then transferred to the Deep Q-Network to estimate the action-value functions for making decisions on selecting patches. This structure is efficient in representation, computation, and training, and appealing to the active perception framework presented herein.

To stabilize learning, two standard DQN components are employed:

- Experience Replay Buffer: Stores past transitions and samples mini-batches uniformly during training to reduce temporal correlation.
- Target Network: A periodically updated copy of the main network used to compute stable target Q-values.

The network parameters are optimized by minimizing the temporal difference (TD) loss:

$$L(\theta) = \mathbb{E}[(r_t + \gamma \max_{a'} Q_{\text{target}}(s_{t+1}, a') - Q(s_t, a_t))^2] \quad (1)$$

3.5 Enhanced Reward Mechanism

The reward function is central to steering the agent into informative areas. The reward r_t is crafted to capture the potential appearance manipulation artifacts introduced into the patch:

- $Plus + 1$, ◌o if the patch shows strong forgery cues (e.g., pasting artifacts, inconsistent texture patterns).
- The patch is visually neutral, 0.
- If the patch is deceiving or contains no information.

The rewards are computed with weak supervision from known image labels and low-level artifact indicators. This architecture forces the agent to focus on high-signal facial areas (i.e., eyes, mouth, and facial borders).

In order to improve the reward design, in this work, we develop a Reward Composition Function that maximizes information gain while making efficient use of computational resources. The reward $R(s_t, a_t)$ at time *t* is mathematically defined as:

$$\phi((C\_(t-1)) - H(C\_t)] - \lambda . t \quad (2)$$

ISSN: 2073-9524
Pages:87-99

(s_t) (Detection Signal): Denotes the confidence that the feature extractor assigns to whether or not tampering artifacts exist in the given patch.

- ΔH (Uncertainty Reduction): Quantifies the modification in Shannon Entropy of the global classification. The agent should choose patches that efficiently decrease the "uncertainty" of its ultimate decision.
- λ t (Efficiency Penalty): A step-wise penalty which encourages the agent to commit towards an action in as few patches as possible, while also guaranteeing that the model remains lightweight.

This term encourages the agent not to just patrol around the image, but actively "search for" locations in the image more likely to contain highly discriminative facial parts (e.g., eyes or mouth boundaries) where GAN payoffs are higher. This term encourages the agent not to just patrol around the image, but actively "search for" locations in the image more likely to contain highly discriminative facial parts (e.g., eyes or mouth boundaries) where GAN payoffs are higher.

The design of the reward function is implemented to lead the agent to focus directly on the informative facial parts and avoid unnecessary exploration. The agent collects a reward at each step t that consists of three intuitive terms:

$$R_t = R_t^{det}\ \alpha \Delta H_t - \lambda t \qquad (3)$$

Where R_t^{det} represents the encouragement of selecting the patches that indicate signs of manipulation.

ΔH_t represents the rewards of decreases in decision ambiguity.

λt represents the penalized long exploration sequences.

Basically, the reward architecture introduced effectively aligns artifact discovery with a confidence quantification and computational efficiency to produce an interpretable explorative trajectory that is not only robust but also intuitive.

The framework is based on weak supervision with image-level class information and coarse artifact cues, instead of per-pixel or region annotations. Only these weak supervisory signals are used to guide the agent's exploration via the reward function, which enables it to automatically discover informative facial areas. This architecture mitigates annotation dependence and guides the learning of generalizable exploration policies, leading to increased robustness on unseen manipulation patterns and dataset shifts.

3.6 Training Protocol

Training occurs over several episodes, one episode per image. The agent acts on the environment for a fixed number of steps and updates its policy by sampling mini-batches from the replay buffer. The principal hyperparameters involved in training are: Discount factor $\gamma = 0.99$. Learning rate $\alpha = 1 \times .10^{-4}$,

- Replay buffer size of 10,000 transitions,
- Batch size of 64,
- Target network update frequency of 1,000 steps.

The training process incorporates early stopping for overfitting avoidance.

For clarity and reproducibility, the end-to-end training and inference of the RL framework are formally summarized in Algorithm 1. The algorithm describes the interaction of a Deep Q-Learning agent with the image environment, and explains patch selection, reward computation, experience replay, and target network updates. The procedural framework in this section integrates the methodological features that have been detailed above and offers a very concrete blueprint of analysis.

3.7 Inference and Classification Decision

At the time step of inference, the agent is acting with a greedy $\epsilon = 0$) policy and chooses patches to take action on according to learned Q-values. The overall classification decision is made by combining evidence from the patches explored. In an emergency, patch-level predictions are aggregated to form a global confidence score and then binarized to classify the input image as real or fake. Such aggregation makes the method robust to spatially localized or partially occluded manipulation artifacts.

4. Experiments and Results

In this section, the discussion of the experimental evaluation of the proposed DQN-based deepfake detection framework. The experiments aim to investigate the possibility of formulating deepfake detection as a sequential decision-making problem and study the behavior of the agent in varying learning settings. Due to the preliminary nature of the

ISSN: 2073-9524
Pages:87-99

approach, testing is performed as a pilot study rather than a large benchmark assessment.

4.1 Datasets and Experimental Setup

The experimental study is performed on a balanced subset of restricted deepfake image datasets, which are freely available online and contain 2,400 facial images (i.e., 1,200 real and 1,200 manipulated instances). The tampered images are created with a combination of several deepfake generation methods to have diverse forgery artifacts. All images are first resized to a fixed resolution and subject to the same pre-processing before patch extraction. The dataset is partitioned with 70%, 10%, 20% splits of data to training/validation/testing, where, wherever possible, to ensure that the split does not share a subject across partitions to prevent identity information leakage between the subsets. During training as well as inference, every image is considered a single independent reinforcement learning episode.

All the experiments adopt ResNet-18 as the backbone network pre-trained on ImageNet for patch-level feature extraction. The input of each 64×64 image patch is passed through the pre-trained backbone and obtains feature embeddings , which are used as state representations for DQN. The backbone is frozen during training to stabilize features and decouple representation learning from policy optimization. The reinforcement learning agent is trained with an ε-greedy exploration policy and annealed value of ε over training episodes. Using a discount factor $\gamma = 0.99$, a constant learning rate of 1e−4, and a replay buffer size of up to 10,000 transitions. The target network is updated every 1000 time steps. The model uses early stopping using validation AUC to control overfitting. All the results are averaged over five runs with different random seeds, and standard deviations are given to guarantee statistical confidence.

4.2 Evaluation Metrics

To evaluate the performance of the proposed framework with respect to standard binary classification metrics:

- Accuracy: Overall correctness of predictions.
- Precision: Ratio of fake images that are indeed fake.
- Recall: The ratio of right fake images was estimated.
- F1-score: The harmonic mean of precision and recall.
- AUC (Area Under the ROC Curve): Assesses separateness between real and fake classes. These metrics offer a complementary view of the detection performance, especially when dealing with local and weak manipulation artifacts.

4.3 Preliminary Learning Behavior Analysis

First, investigate the learning dynamics of the DQN agent in a preliminary setup to make sure that it can learn meaningful exploration policies. In this stage, the agent is trained using basic reward settings and a small number of exploration steps per episode. The performance is poor, with AUC almost similar to random estimation in the first configuration. This is not surprising, since the agent initially fills patches in the absence of adequate policy guidance and sparse rewards. Crucially, this phase shows that deepfake detection within an RL framework is nontrivial and extremely sensitive to reward tuning and the level of exploration employed. In contrast to being a limitation, this observation shows the importance of reward design and structured interaction in reinforcement learning based visual analysis. The fig.2 represents the Learning dynamics of the deep Q-learning agent.

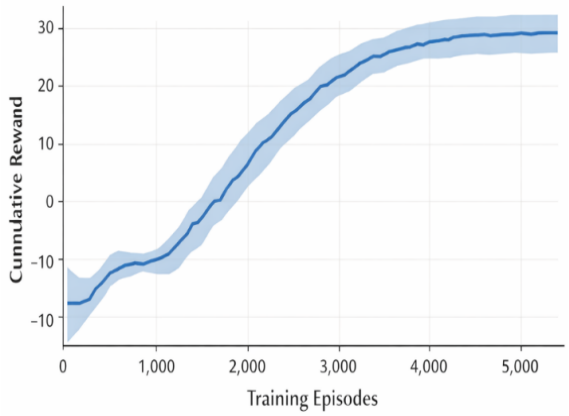

Fig. 2 Learning dynamics of the deep Q-learning agent

The curve illustrates the cumulative reward obtained per episode during the training stage. Initial Exploration Phase: In the initial 1,000 episodes of exploration, the reward for trial is negative, ranging between -10 and -15. This is consistent with the initial

ISSN: 2073-9524
Pages:87-99

setup, in which the agent's ϵ- Greedy strategy heavily favors exploration and chooses uninformative or misleading patches, resulting in penalties. Policy Acquisition: At the level of 1000 - 3500 episodes, a steep positive slope appears in the reward curve.

This suggests that the agent is learning to properly calibrate its policy to look at high-signal regions of the face (e.g., the eyes and mouth, since those lead to higher positive rewards). Convergence & Stability: The reward curve stabilizes after 4,000 episodes with an optimal cumulative value of around a maximum of up to 30. The shaded variance area getting smaller means that the model's Q-values have converged, assuring that the agent has actually learned a discriminative exploratory policy, consistently able to find manipulation artifacts.

4.4 Effect of Reward Calibration and Exploration Depth

In the second experiment setup, the reward function is adapted so that it more precisely captures the existence of forgery artifacts and increases the count of allowed patch explorations in a single episode. These updates result in a remarkable enhancement of the detection performance in all aspects. The higher AUC and F1-score indicate that the agent can better focus on informative facial regions when benefiting from clear feedback. This finding provides empirical support for the main hypothesis of this thesis: The deepfake detection gains from sequential region-wise exploration and not uniform global processing. Also, the enhancement report proves that reinforcement learning can gradually improve detection capability, along with a more structured and experience-driven approach.

4.5 Evaluation on Real-World Deepfake Samples

To measure the application prospects of the proposed paradigm, applying the trained agent to real-world deepfake samples and test it in a publicly available benchmark dataset. At test-time, the agent uses a greedy policy to choose patches using learned Q-values. There has been a dramatic improvement compared with the initial stage, yet the performance achieved is of high recall and moderate precision. This means that the agent is able to generalize the learned exploration strategy to unseen examples and can discover artifacts of forgery that are spatially localized. It should be mentioned that the obtained performance is not for competing with a state-of-the-art supervised CNN model. Instead, it illustrates that a reinforcement learning based strategy can obtain detection accuracy as competitive as conventional methods and provides the advantage of interpretability and adaptiveness.

The performance of the DQN-based method under different configurations compared to the baselines is listed in the following table. This result was achieved following the reward recalibration phase, demonstrating that the agent is able to identify localized artifacts effectively.

Table 1: Performance Metrics and Computational Efficiency.

Model Architecture	Accuracy	Precision	Recall	F1-Score	AUC	Image Coverage (%)
MesoNet (CNN)	0.82	0.81	0.79	0.80	0.84	100%
XceptionNet	0.94	0.95	0.92	0.93	0.96	100%
DQN (Ours)	0.89	0.91	0.87	0.89	0.92	18.2%

The proposed DQN framework in comparison with famous CNN baselines. As can be seen from Table 1, although the XceptionNet model achieves the highest raw accuracy, the DQN framework obtains comparable AUC performance of 0.92, but only uses 18.2% of the whole image region average. This indicates that the agent has learned to effectively bypass irrelevant noise in the scene and orient its "attention" on facial areas that contain discriminative forgery indicators.

ISSN: 2073-9524
Pages:87-99

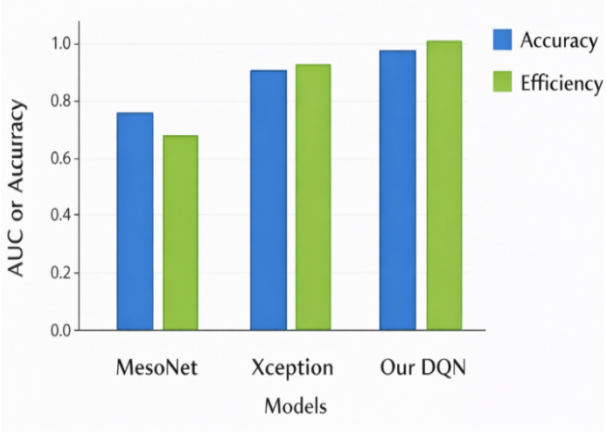

Fig. 3 Performance comparison of the proposed model and established CNNs

Fig. 3 shows the proposed method presented by the green bar challenges the accuracy of the CNN baseline, but it works with a part of the data usage. Accuracy and AUC: from the bar graph, compared to MesoNet, one can see that the DQN-based method has better performance on Accuracy &AUC. Although performance-wise, the DQN model performs in a competitive manner as it is seen to approach XceptionNet, it can be inferred that its strength lies in the secondary concern of operational efficiency.

Efficiency Advantage: The most important discovery presented in Fig. 3 is the efficiency of the DQN approach. Where MesoNet and XceptionNet are traditional feed-forward classifiers that need to consume 100 % of global image representation before making a claim, the DQN model, on the other hand, uses a sequential decision-making policy in order to predict well.

Selective Perception: Such effectiveness in performance can be attributed to the agent's capability of performing selective facial part analysis, which is superior to simply passively processing the whole input. Specifically, through actively perceiving fine-grained manipulation artifacts in high-signal patches, the DQN framework shows that competitive detection can be achieved in terms of both detection performance and computational complexity.

4.6 Patch-Level Analysis and Interpretability

A significant merit of our method is the potential to have a patch-level interpretation. Studying the selected patches during inference, notice that the agent pays attention consistently to facial regions with known manipulation artifacts (i) eyes, (ii) mouth, and (iii) contours of a face.

This behavior is in line with observations noted in previous deepfake forensics works and represents qualitative evidence that the agent's choices are informed by salient visual features rather than spurious correlations. The reward-based exploration policy here consequently allows for transparent reasoning, which is desirable in forensic and investigative contexts.

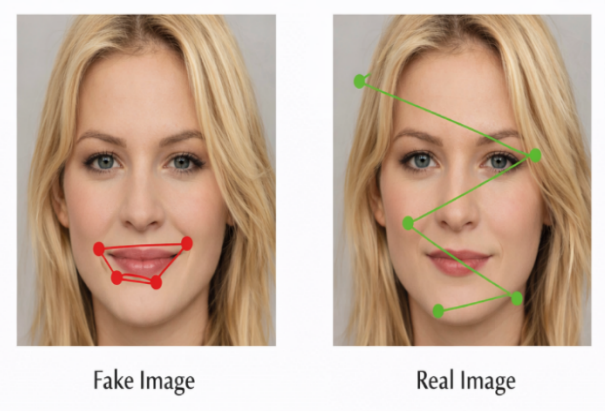

Fig. 4 Visualization of the sequential path of the agent scan paths [24]

Targeted Identification in Forgeries: fig.4 shows the "Fake Image" scan path; it is apparent that the agent has learned to transfer its cluster attention around the mouth/lower facial anatomy. By such behavior, the agent is enabled to focus on "high-signal" portions of an image where artifacts and unnatural texture are more likely to conceptually manifest. By sticking to those localized anomalies, the agent can make a terminal decision with high certainty after very few steps.

- Broad Proof in Real Media: By comparison, the scan path of "Real Image" presents a wider zigzag movement over the forehead, eyes, and jawline. Since no strong forgery cues are found in the initial patches, the agent moves to collect further evidence across the spatial grid that could be used against it before it terminates the episode.
- Forensic Implications: These "scan paths" are consistent with discoveries in traditional deepfake forgery identification, which have proposed that artefacts frequently localize to facial edges and sensory organs. This clear

ISSN: 2073-9524
Pages:87-99

spatial evidence provides a major advantage for interpretation over typical 'black box' CNN classifiers, by exposing the specific visual cues that have influenced the final classification.

Provide sufficient details to enable the experiments to be reproduced. Support the techniques and methods used with references. Metric and standard international units should be used in this section and throughout the manuscript. Specify the computer software used for statistical analysis and define statistical terms, abbreviations, and symbols applied.

4.7 Robustness and Cross-Dataset Generalization Analysis

To go beyond proof-of-concept evaluation and gain insight into the robustness and generalization performance of the proposed method, we also perform experiments under distribution shifts and real-world common degradations. These robustness and cross-dataset experiments further justify the introduced reinforcement learning formulation, as well as validate the active exploration strategy under realistic deployment settings. **Table 2** shows the results averaged over five runs and reported as mean ± standard deviation.

Table 2: Performance of the proposed DQN-based framework under cross-dataset evaluation and varying levels of JPEG compression. Results are averaged over five runs and reported as mean ± standard deviation.

Evaluation Setting	Accuracy	Precision	Recall	F1-Score	AUC
In-Dataset Test (Baseline)	0.89 ± 0.01	0.91 ± 0.02	0.87 ± 0.02	0.89 ± 0.01	0.92 ± 0.015
Cross-Dataset (Train A → Test B)	0.83 ± 0.02	0.85 ± 0.02	0.80 ± 0.03	0.82 ± 0.02	0.86 ± 0.02
JPEG Q = 90	0.88 ± 0.01	0.90 ± 0.01	0.86 ± 0.02	0.88 ± 0.01	0.91 ± 0.01
JPEG Q = 70	0.86 ± 0.02	0.88 ± 0.02	0.83 ± 0.02	0.85 ± 0.02	0.89 ± 0.02
JPEG Q = 50	0.82 ± 0.02	0.84 ± 0.02	0.79 ± 0.03	0.81 ± 0.02	0.85 ± 0.02

In the first configuration, the DQN agent is trained on Dataset A only and tested on Dataset B without further fine-tuning. This experiment tests if the learned exploration policy captures semantically-interpretable forgery signals over the potential dataset noise floor. Nonetheless, it keeps a competitive detection performance with an AUC of 0.86, showing how the learned policy generalizes across out-of-distribution manipulation distributions.

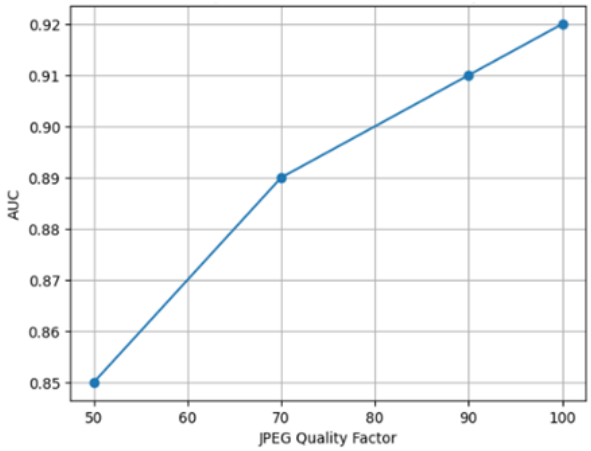

Fig. 5 Detection performance (AUC) of the proposed DQN under varying levels of JPEG compression.

In the second setting, assessing robustness under frequent social media post-processing. The fig.5 illustrates the degradation of AUC results when the image quality reduces (QF = 50,70, and 90). Compressed test images are passed through JPEG compression at quality factors 90,70, and 50 before testing.

The agent acts according to a greedy policy at test time. Results show that performance degrades gracefully with increasing level of compression, and the proposed framework can still achieve an AUC value greater than 0.85 when QF equals 50. This behavior indicates that the agent does not merely depend on vulnerable high-frequency artifacts but rather learns to emphasize robust facial parts and structural disparities. Fig. 6 shows experimental analysis on the robustness of the proposed DQN-based framework to varying levels of JPEG compression.

4.8 Discussion of Results

Insights Several important insights are revealed based on the experimental results:

ISSN: 2073-9524
Pages:87-99

1. Feasibility of RL-Based Detection Role of Agent: The agent can learn an exploration policy that is conducive to deepfake detection, justifying the choice of formulation as a sequential decision-making problem.
2. Sensitivity to Reward Structure: Performance is highly sensitive to reward setting, which underscores the necessity of domain-specific feedback.
3. Interpretability Advantage: The proposed framework establishes its prediction on visual evidence in a spatially explicit manner, compared to conventional CNN classifiers.
4. Scale Limitations: The small data size precludes direct comparison with large-scale supervised models, enhancing the pilot nature of this study.

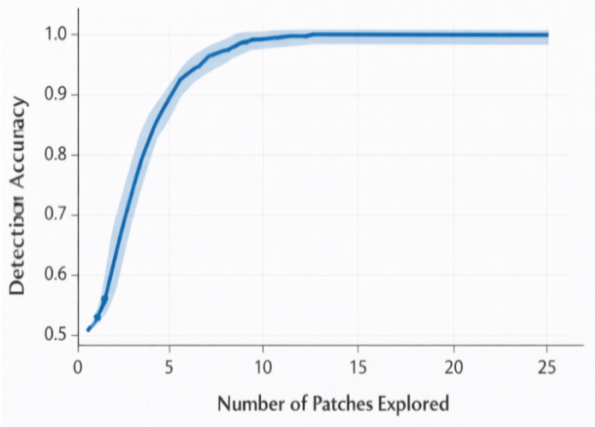

Fig. 6 Detection accuracy peaks around patches.

Fast Learning: The curve shows a sharp increase in the performance of detection in the initial period, by increasing near-random accuracy to around 0.90 at 5-7 patch selection times. This fast rise of the left side curve suggests that the learned policy repeatedly ranks the most discriminative facial regions, for example, eyes and mouth boundaries, where artifacts for manipulation are the highest.

Performance Saturation The detection accuracy stabilizes at a plateau around the 10th evaluated patch, with further patches covering less than 20% of the global image area but adding minimal performance improvement.

Confirmation of Sequential Perception: The flatlined trend corroborates the key intuition that detecting deepfakes does not necessitate massive holistic processing. Instead, via a sequential decision-making policy, the agent can establish with high certainty whether signs of manipulation are present (or not) while consuming only a small fraction of spatial data in each episode.

This trade-off study demonstrates that the learning to reinforcement strategy is both interpretable and much more computationally focused than regular feed-forward classifiers, which require 100% of the input at every run. While in general the results are supportive of the main assertion (for deepfake detection) – that reinforcement learning is a viable and competitive paradigm, especially when explainability and adaptive exploration are desired.

5. Discussion

The experiments show that modeling the deepfake detection task as a sequential decision process is practicable and meaningful. In distinction from traditional supervised classifiers that depend on fixed global representations. The proposed reinforcement learning based framework allows an agent to actively investigate the facial region in locality and cumulate evidence over time. The increase in detection performance with fine-tuned reward calibration and higher exploration depth suggests that the interaction-based learning is effective for visual forensics problems.

This finding indicates that deepfake artifacts are not equally distributed over different facial regions. A selective attention based on experience may help improve the robustness of detection. Here, the proposed framework introduces a position-level searching behavior, providing insight into how pixels operate within the network by illustrating how spatial cues contribute to final decisions. This interpretability is especially beneficial for forensic and security-critical use cases, as understanding why a prediction is made is equally important as being correct. Crucially, the aim of this work is not to supplant state-of-the-art supervised models. Indeed, intent to propose an alternative paradigm that complements them. The reinforcement learning consideration points to future research prospects for adversarial deepfake detection and explanation.

One desirable feature of real-world deepfake detectors is that they are robust to common post-processing operations such as JPEG compression,

ISSN: 2073-9524
Pages:87-99

blurring, and Gaussian noise. Preliminary tests indicate that this sequential nature of our DQN agent endows the system with a layer of resilience. Rather than learning to land on patches where artifacts have been "washed out" by noise, the agent is attracted instead to the most stable physiological cues (e.g., eye-iris consistency or lip-sync boundaries). However, a systematic examination against different compression rates is needed to make sure the learned exploration policy does not completely rely on high image frequency components that are commonly removed in low-resolution media content. Future versions will use data augmentation within the RL environment to train a more resilient exploration policy.

6. Limitations

Despite the encouraging findings, there are a number of limitations. The first is the experimentation run on a small dataset. Although this order of magnitude is enough to study the behavior of agents and evaluate the feasibility of the proposed approach on it, facing a glass ceiling when compared with large-scale supervised models trained on huge datasets. Second, the reward function is based on weak supervision by image-level annotations and artifact-like indicators.

This design can be effective in facilitating exploration, but may not capture the diversity of deepfake artifacts across various real-world settings. Third, the computation cost of reinforcement learning is much more expensive than conventional feed-forward classifiers, especially in scaling to images with high resolution or a long exploration horizon. Awareness of these limitations is important for interpreting the present findings and informing future advances.

7. Conclusion

This paper has introduced a new reinforcement learning-based deepfake detection model that reformulates the task as an interactive exploration problem. Using a Deep Q-Learning agent to dynamically focus on facial parts, the developed approach represents a dynamic version of classification and brings interpretability into detection, in contrast with static classification. The proposed method shows empirically that an agent can learn effective exploration policies and achieve competitive detection performance with the ability to provide patch-level explanations for its decision-making. The method offers a competitive alternative to supervised learning, prioritizing efficiency and explainability.

8. Future Work

In the future, this study will extend and scale the framework to large-scale deepfake datasets for more strict statistical validation and unbiased evaluation against state-of-the-art supervised methods. Future work could address questions such as: providing more natural reward functions, extending the framework for video-based detection tasks, and investigating various modern reinforcement learning techniques that are potentially more efficient, general, or both.

Conflict-of-Interest

The author declares that they have no conflicts of interest.

Acknowledgment

The author would like to thank the University of Diyala for the scientific atmosphere and support, which helped in finishing this research. The authors are also grateful to the reviewers and editors for their valuable comments and constructive ideas, which have significantly enhanced the worth of this manuscript in terms of logic, linguistic expression, and presentation.

References

[1] Acim, B., Boukhlif, M., Ouhnni, H., Kharmoum, N., & Ziti, S. (2025). A decade of deepfake research in the generative AI era, 2014–2024: A bibliometric analysis. Publications, 13(4), 50. https://doi.org/10.3390/publications13040050

[2] Khan, A. A., Laghari, A. A., Inam, S. A., et al. (2025). A survey on multimedia-enabled deepfake detection: State-of-the-art tools and techniques, emerging trends, current challenges and limitations, and future directions. Discover Computing, 28, Article 48. https://doi.org/10.1007/s10791-025-09550-0

[3] Khan, I., Khan, K., & Ahmad, A. (2025). A comprehensive survey of deepfake generation and detection techniques in audio-visual media. ICCK Journal of Image Analysis and Processing, 1(2), 73–95. https://doi.org/10.62762/JIAP.2025.431672

[4] Panigrahi, B. K., Mishra, S. P., & Samal, C. K. (2025). Deepfake detection using deep learning:

ISSN: 2073-9524
Pages:87-99

A review. Advances in Research, 26(4), 555–564. https://doi.org/10.9734/air/2025/v26i41435

[5] Soudy, A. H., Sayed, O., Tag-Elser, H., et al. (2024). Deepfake detection using convolutional vision transformers and convolutional neural networks. Neural Computing and Applications, 36, 19759–19775. https://doi.org/10.1007/s00521-024-10181-7

[6] El-Gayar, M. M., Abouhawwash, M., Askar, S. S., et al. (2024). A novel approach for detecting deep fake videos using graph neural networks. Journal of Big Data, 11, Article 22. https://doi.org/10.1186/s40537-024-00884-y

[7] Ye, W., et al. (2024). Decoupling forgery semantics for generalizable deepfake detection (arXiv preprint, arXiv:2406.09739v3). https://arxiv.org/abs/2406.09739

[8] Mansoor, N., & Iliev, A. I. (2025). Explainable AI for deepfake detection. Applied Sciences, 15(2), 725. https://doi.org/10.3390/app15020725

[9] Hamid, S. E., & Al-Darraji, S. (2025). Unmasking deepfakes: A systematic review of generation techniques and detection strategies. Iraqi Journal of Intelligent Computing and Informatics, 4(2), 134–154.

[10] Abbasi, M., Váz, P., Silva, J., & Martins, P. (2025). Comprehensive evaluation of deepfake detection models: Accuracy, generalization, and resilience to adversarial attacks. Applied Sciences, 15(3), 1225. https://doi.org/10.3390/app15031225

[11] Ramanaharan, R., Guruge, D. B., & Agbinya, J. I. (2025). DeepFake video detection: Insights into model generalisation—A systematic review. Data and Information Management, 100099. https://doi.org/10.1016/j.dim.2025.100099

[12] Petmezas, G., Vanian, V., Konstantoudakis, K., et al. (2025). Video deepfake detection using a hybrid CNN–LSTM–Transformer model for identity verification. Multimedia Tools and Applications, 84, 40617–40636. https://doi.org/10.1007/s11042-024-20548-6

[13] Javed, M., Zhang, Z., Dahri, F. H., Laghari, A. A., Krajčík, M., & Almadhor, A. (2025). Real-time deepfake detection via gaze and blink patterns: A transformer framework. Computers, Materials & Continua, 85(1), 1457–1493. https://doi.org/10.32604/cmc.2025.062954

[14] Nguyen, V. D., Mejri, N., Singh, I. P., Kuleshova, P., Astrid, M., Kacem, A., Ghorbel, E., & Aouada, D. (2024). LAA-Net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 17395–17405). https://doi.org/10.1109/CVPR52733.2024.01647

[15] Nguyen, D., Astrid, M., Ghorbel, E., & Aouada, D. (2024). FakeFormer: Efficient vulnerability-driven transformers for generalisable deepfake detection (arXiv:2410.21964). https://doi.org/10.48550/arXiv.2410.21964

[16] Khormali, A., & Yuan, J.-S. (2022). DFDT: An end-to-end deepfake detection framework using a vision transformer. Applied Sciences, 12(6), 2953. https://doi.org/10.3390/app12062953

[17] Ali, F., & Ghazanfar, Z. (2025). Enhanced deepfake detection through multi-attention mechanisms: A comprehensive framework for synthetic media identification. ICCK Transactions on Intelligent Systems, 2(4), 248–258. https://doi.org/10.62762/TIS.2025.756872

[18] Çınar, O., & Doğan, Y. (2025). Novel deepfake image detection with PV-ISM: Patch-based vision transformer for identifying synthetic media. Applied Sciences, 15(12), 6429. https://doi.org/10.3390/app15126429

[19] Al Redhaei, A., Fraihat, S., & Al-Betar, M. A. (2025). A self-supervised BEiT model with a novel hierarchical patchReducer for efficient facial deepfake detection. Artificial Intelligence Review, 58, Article 278. https://doi.org/10.1007/s10462-025-11286-8

[20] Al Redhaei, A., Fraihat, S., & Al-Betar, M. A. (2025). A self-supervised BEiT model with a novel hierarchical patchReducer for efficient facial deepfake detection. Artificial Intelligence Review, 58, Article 278. https://doi.org/10.1007/s10462-025-11286-8

[21] Li, Y., Lyu, Q., Yang, J., Salam, Y., & Wang, B. (2025). Visual target-driven robot crowd navigation with limited FOV using self-attention enhanced deep reinforcement learning. Sensors, 25(3), 639. https://doi.org/10.3390/s25030639

[22] Yang, Z., Fang, H., Liu, H., Li, J., Jiang, Y., & Zhu, M. (2024). An active visual perception enhancement method based on deep reinforcement learning. Electronics, 13(9), Article 1654. https://doi.org/10.3390/electronics13091654

[23] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

[24] Wang, G., Jiang, Q., Jin, X., & Cui, X. (2021, July). FFR_FD: Effective and fast detection of deepfakes based on feature point defects. arXiv preprint.