

OFFLINE REINFORCEMENT LEARNING FOR INFRARED IMAGE-BASED MISSILE GUIDANCE AND CONTROL

Seongyeon Kim[1], Jongho Shin[1] & Hyeong-Geun Kim[2]

[1]Chungbuk National University
[2]Konkuk University

Abstract

Online reinforcement learning is difficult to apply to real missile systems because direct exploration involves high cost and safety risk. This study proposes an offline reinforcement learning framework for infrared image-based integrated missile guidance and control. A fixed dataset is generated using a sliding mode guidance and control law, and a TD3+BC algorithm is used to reduce distributional shift during offline policy learning. Expert actions are further introduced into the critic loss to improve Q-value estimation and keep the learned policy close to reliable data. The policy processes stacked infrared images and engagement-related state variables to generate fin commands. The proposed method is evaluated in a Unity-based engagement simulator under various initial conditions. Simulation results show stable interception behavior and smooth control responses. These results indicate that expert data-driven offline reinforcement learning can provide a practical approach to missile guidance and control without online exploration.

Keywords: guidance and control, offline reinforcement learning, TD3+BC, virtual environment simulation

1. Introduction

Missile guidance and control are key functions for accurate target engagement. Guidance generates commands that steer the missile toward the target, while control regulates the missile's attitude according to these commands [1]. Conventional methods often use fixed design parameters. Therefore, their performance can degrade under severe engagement conditions or rapid environmental changes. To improve adaptability, recent studies have investigated reinforcement learning (RL)-based integrated guidance and control methods.

RL can learn control policies without an explicit system model. It can also reflect nonlinear dynamics and sensor characteristics through simulation-based training. However, many RL methods require online interaction with the environment. This requirement is not suitable for real missile systems because exploration can be costly and unsafe.

Offline RL provides a practical alternative by learning a policy from a fixed dataset. In aerospace applications, such datasets can be generated using validated expert guidance and control methods. However, directly applying conventional actor–critic algorithms in an offline setting can lead to performance degradation. A main reason is inaccurate Q-value estimation for actions that are outside the data distribution.

This paper proposes a TD3+BC-based offline RL framework for infrared image-based integrated missile guidance and control [2]. Expert actions from a classical guidance and control law are used to improve critic training and to reduce divergence toward out-of-distribution actions. Infrared image observations are also used as policy inputs so that the learned controller can exploit target information in the image plane.

2. Offline RL-based integrated guidance and control

This section presents the two-dimensional engagement model and the proposed offline RL-based guidance and control method.

2.1 Two-dimensional dynamic model

Offline policy learning is conducted using the following two-dimensional missile model in [3]:

$$\begin{aligned}
\dot{x} &= V\cos\gamma, \\
\dot{z} &= -V\sin\gamma, \\
\dot{\gamma} &= \theta - \alpha, \\
\dot{a} &= q - \frac{L(\alpha,\delta)}{mV}, \\
\dot{q} &= \frac{M(\alpha,q,\delta)}{I}, \\
\dot{\theta} &= q, \\
\dot{\delta} &= -\frac{1}{\tau}\delta + \frac{1}{\tau}\delta_c,
\end{aligned} \tag{1}$$

where x and z denote the horizontal and vertical positions. V, γ, and α are the velocity, flight-path angle, and angle of attack, respectively. The variables m, q, I, θ, δ, δ_c and τ denote the missile mass, pitch rate, moment of inertia, pitch angle, fin angle, fin angle command and actuator time constant. L and M represent the lift force and pitch moment.

2.2 TD3+BC-based guidance and control

A TD3+BC-based offline RL architecture is designed to generate control commands from dynamic states and simulated infrared images.

2.2.1 Observation (state) configuration

The policy state consists of visual observations o_k^{img} and vector observations o_k^{vec}. A sequence of S consecutive frames is stacked as follows:

$$\begin{aligned}
o_k^{\text{img}} &= [\mathrm{I}_{k-S+1},\ldots,\mathrm{I}_k],\ \mathrm{I}_k \in \mathbb{R}^{W\times H}, \\
o_k^{\text{vec}} &= \left[\sin(\lambda-\gamma),\cos(\lambda-\gamma),\dot{\lambda},q,\delta,\dot{a},\dot{\theta},\dot{q},\dot{\gamma}\right]^{(S)}, \\
s_k &= \left(o_k^{\text{img}},\, o_k^{\text{vec}}\right), \\
\mathrm{a}_k &= \delta_c,
\end{aligned} \tag{2}$$

where I denotes the infrared image, and W and H are the image width and height. The terms $\sin(\lambda-\gamma)$ and $\cos(\lambda-\gamma)$ describe the missile velocity direction relative to the line-of-sight (LOS). The action a is the fin angle command and is bounded within $\pm 30^\circ$.
In TD3+BC, the policy is trained using only a fixed dataset $\mathscr{D} = \{(s_i,\mathrm{a}_i,r_i,s_i')\}_{i=1}^N$.

2.2.2 Critic training

The critic loss is defined as follows:

$$\mathscr{L}_Q(\phi_i) = \mathbb{E}_{(s,\mathrm{a},r,s')\sim\mathscr{D}}\left[\left(Q_{\phi_i}(s,\mathrm{a})-y\right)^2\right], \quad i\in\{1,2\}. \tag{3}$$

where y is the Bellman target. An additional loss term is introduced to reduce Q-value underestimation:

$$\mathscr{L}_{exp} = \mathbb{E}_{(s,\mathrm{a}_{exp})\sim\mathscr{D}}\left[\mathrm{ReLU}\left(Q\left(s,\mathrm{a}_{exp}\right)-M-Q\left(s,\pi(s)\right)\right)\right], \tag{4}$$

where a_{exp} denotes the expert action, and M represents the margin. This term penalizes the critic when the policy action has a lower Q-value than the expert action by more than the margin. It encourages the learned action to obtain a value comparable to that of the expert action. The total critic loss is then given by:

$$\mathscr{L}_{total} = \mathscr{L}_Q + \mathscr{L}_{exp}. \tag{5}$$

2.2.3 Actor training

To suppress distributional shift in offline learning, the actor loss combines Q-value maximization with behavior cloning (BC):

$$\mathscr{L}_\pi(\theta) = \mathbb{E}_{(s,\mathrm{a})\sim\mathscr{D}}\left[-\lambda\, Q_{\phi_1}(s,\pi_\theta(s)) + \|\pi_\theta(s)-\mathrm{a}\|_2^2\right], \tag{6}$$

where λ is adaptively adjusted based on the Q-value scale and is defined as follows:

$$\lambda = \frac{\alpha_Q}{\mathbb{E}_{s\sim\mathscr{D}}\left[\left|Q_{\phi_1}(s,\pi_\theta(s))\right|\right]+\varepsilon}, \tag{7}$$

where $\alpha_Q > 0$ is a hyperparameter, and ε is a small constant (1×10^{-6}) for numerical stability. This formulation encourages actions with high predicted value while keeping the policy close to the expert data distribution.

2.3 Network architecture

The proposed policy uses a multimodal actor–critic architecture that processes infrared images and engagement-related vector states. Four grayscale frames are stacked as $o_k^{\mathrm{img}} \in \mathbb{R}^{64\times64\times4}$. Nine vector variables are stacked over the same time window as $o_k^{\mathrm{vec}} \in \mathbb{R}^{36}$. The image branch uses a convolutional encoder, and the vector branch uses a two-layer MLP. Their outputs are fused into a 384-dimensional state feature. The feature extraction process is written as follows:

$$z_k^{\mathrm{img}} = f_{\mathrm{img}}(o_k^{\mathrm{img}};\psi_{\mathrm{img}}), \qquad z_k^{\mathrm{vec}} = f_{\mathrm{vec}}(o_k^{\mathrm{vec}};\psi_{\mathrm{vec}}), \qquad z_k = [z_k^{\mathrm{img}}, z_k^{\mathrm{vec}}], \tag{8}$$

where f_{img} and f_{vec} denote the image and vector encoders. In this study, $z_k^{\mathrm{img}} \in \mathbb{R}^{256}$, $z_k^{\mathrm{vec}} \in \mathbb{R}^{128}$, and $z_k \in \mathbb{R}^{384}$.

The actor maps the fused feature to the fin command, which is bounded by $\delta_{\max} = 30^\circ$:

$$\mathrm{a}_k = \delta_{\max}\tanh(g_\pi(z_k;\theta)), \tag{9}$$

where g_π is the actor head and θ is the actor parameter.

Following TD3, two independent critics estimate the Q-value from the fused feature and action:

$$\tilde{z}_k = [z_k, \mathrm{a}_k], \qquad Q_{\phi_i}(s_k,\mathrm{a}_k) = g_{Q_i}(\tilde{z}_k;\phi_i), \qquad i\in\{1,2\}. \tag{10}$$

The expert-action critic augmentation in Eq. 4 is used to improve offline learning stability. In the implementation, the image encoder consists of Conv32–Conv64–Conv64 and FC256 layers. The vector encoder uses FC128–FC128 layers, and the actor and critic heads use FC256–FC256 layers. This structure allows the policy to use both image-plane target information and engagement dynamics for infrared image-based guidance and control.

3. Validation

This section describes the virtual simulation framework used to generate infrared images and engagement states for policy training and validation. Then, simulations are conducted to verify the validity of the proposed method, and the results are analyzed.

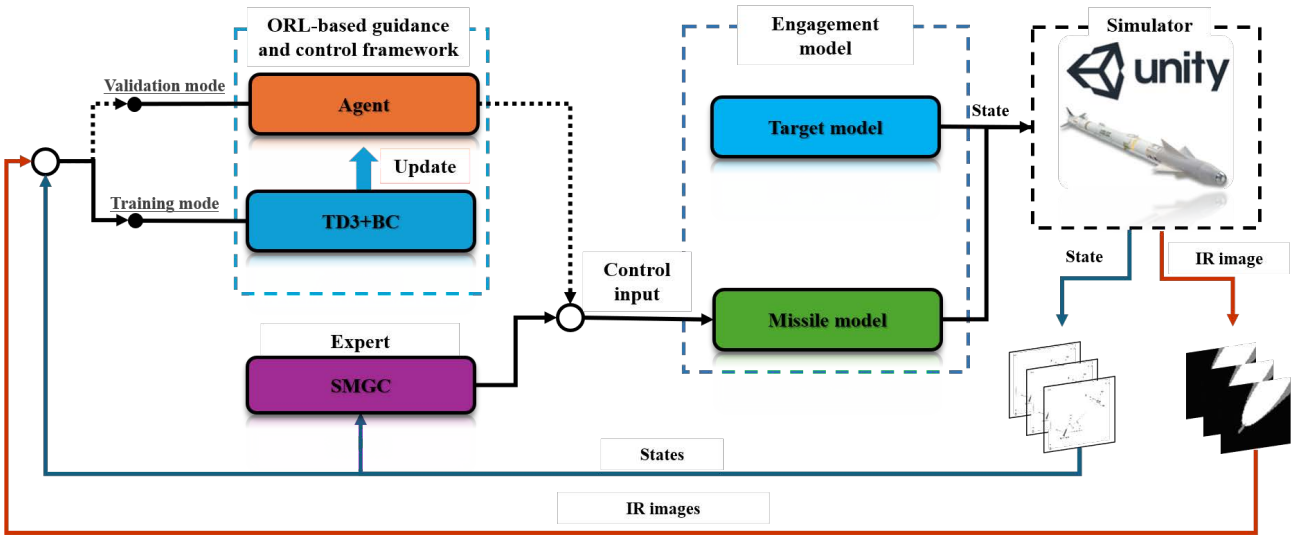


Figure 1 – Conceptual diagram of the Unity-based simulator

3.1 The Unity-based engagement simulator

Figure 1 shows the conceptual structure of the Unity-based simulation framework for offline reinforcement learning-based missile guidance and control. The simulator consists of target and missile engagement models, which generate engagement states and infrared images during the interception scenario. In the training mode, expert control commands generated by the SMGC law are used to construct the offline dataset and train the TD3+BC-based agent. In the validation mode, the trained agent directly receives stacked infrared images and state information from the simulator and generates fin control commands for missile guidance and control.

3.2 Virtual infrared image generation

The Unity-based engagement simulator generates 64×64 grayscale infrared observations from the missile viewpoint. At each step, the two-dimensional engagement model provides the relative distance R_k, LOS angle λ_k, and missile pitch angle θ_k. The seeker boresight error and image-plane target position are defined as follows:

$$\eta_k = \lambda_k - \theta_k, \qquad u_k = \frac{W}{2}\left(1 + \frac{\tan\eta_k}{\tan(\mathrm{FOV}/2)}\right), \qquad v_k = \frac{H}{2}, \tag{11}$$

where u_k and v_k denote the horizontal and vertical pixel coordinates of the target center in the infrared image at time step k, respectively. $W = H = 64$ and $\mathrm{FOV} = 60^\circ$. If $|\eta_k| > \mathrm{FOV}/2$, the target is outside the seeker field of view, and only the background image is generated.
The target spot is modeled as a Gaussian intensity distribution as follows:

$$I_k(u,v) = I_{\mathrm{bg}} + A_k \exp\left(-\frac{(u-u_k)^2 + (v-v_k)^2}{2\sigma_k^2}\right) + \varepsilon_k, \tag{12}$$

where $I_{\mathrm{bg}} = 0.05$ and $\varepsilon_k \sim \mathcal{N}(0, 0.01^2)$. The intensity scale A_k and apparent size σ_k vary with relative distance as follows:

$$A_k = \mathrm{clip}\left(\frac{5000}{R_k}, 0.3, 1.0\right), \qquad \sigma_k = \mathrm{clip}\left(2.0\frac{5000}{R_k}, 1.5, 6.0\right). \tag{13}$$

Pixel values are clipped to $[0,1]$. Thus, the target becomes brighter and larger as the missile approaches.

Four consecutive infrared frames are stacked and combined with the vector observation to form the policy input:

$$o_k^{\text{img}} = [I_{k-3}, I_{k-2}, I_{k-1}, I_k], \qquad I_k \in \mathbb{R}^{64\times 64}, \qquad s_k = \left(o_k^{\text{img}}, o_k^{\text{vec}}\right). \tag{14}$$

This representation provides target-location and temporal motion information for image-based guidance and control.

3.3 Simulation and results

The expert dataset was generated using the sliding mode guidance and control (SMGC) method [4]. This method provides stable interception performance under diverse initial engagement conditions, enabling the collection of reliable reference data for offline learning.
The initial simulation settings are listed in Table 1.

Table 1 – Simulation settings

Type	$[x, z]$ (m)	V (m/s)	γ (deg)	τ
Missile	[0, 0]	500	0–40	0.02
Target	[5000, 0]	200	90–180	-

Training was conducted for 5×10^6 time steps. The learning rate, buffer size, and batch size were set to 0.0003, 500,000, and 256, respectively. At each training epoch, the initial flight-path angle γ was randomly sampled at 5° intervals within the range in Table 1.

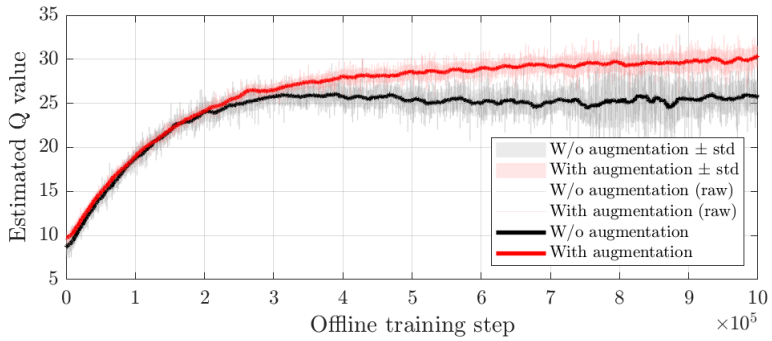


Figure 2 – Estimated Q-value during training

The reward function is defined as follows:

$$\begin{aligned}
r_{total} &= r_1 + r_2 + r_3 + r_4 + r_5 + r_6 + r_7 + r_8, \\
r_1 &= 100\exp\left(-ZEM^2\right), \\
r_2 &= \exp\left(-\frac{d^2}{5000^2}\right), \\
r_3 &= 5\exp\left(-\dot{\lambda}^2\right), \\
r_4 &= \exp\left(-\Delta_{action}^2\right), \\
r_5 &= \exp\left(-(\mathrm{a}_k - 2\mathrm{a}_{k-1} + \mathrm{a}_{k-2})^2\right), \\
r_6 &= 1000 \ (\text{if } \ d \le 3), \\
r_7 &= -1000 \ (\text{if } \ \dot{r} \ge 0), \\
r_8 &= -1,
\end{aligned} \tag{15}$$

Table 2 – Case study under 4,000 validation scenarios

Method	Successes	Rate	$d_{\min}$	$J_{\Delta a}$
SMGC	3661	91.5%	2.46	0.68
TD3+BC w/o image	3527	88.2%	2.72	0.44
TD3+BC	3632	90.8%	2.31	0.39
Proposed	3708	92.7%	2.17	0.33

where $r_1, \ldots, r_8$ denote the Zero-Effort-Miss (ZEM) reward [5], target distance reward, LOS angular rate reward, action magnitude reward, action smoothness reward, interception success reward, interception failure penalty, and time penalty, respectively. For validation, the missile and target flight-path angles in Table 1 were discretized at 1° intervals. A total of 4,000 validation runs were conducted with randomly sampled angles. An interception was considered successful when the target distance d was within 3 m. The proposed method achieved 3,708 successful interceptions, while SMGC achieved 3,661. These values correspond to success rates of 92.7% and 91.5%, respectively. The proposed method improved the success rate by 1.2 percentage points. The trajectories and control commands in Fig. 3 also show stable interception behavior.

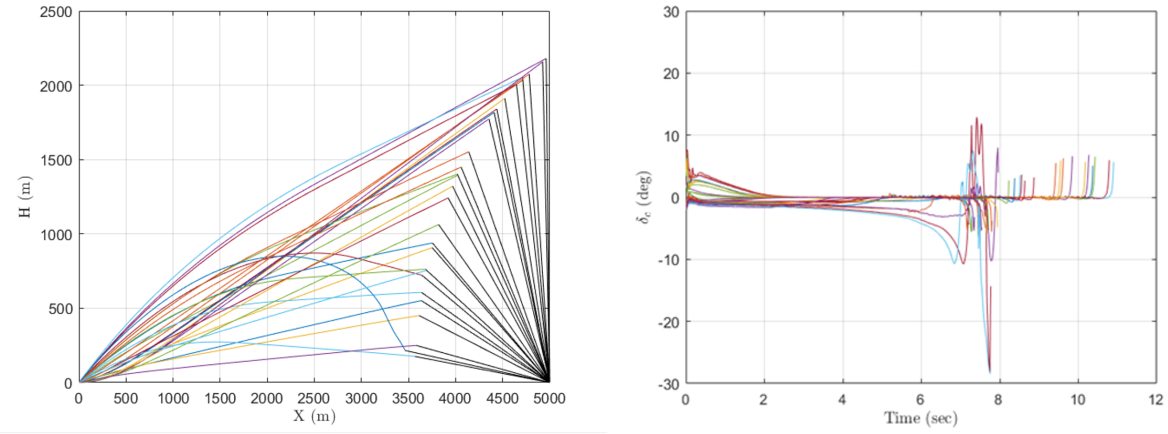


Figure 3 – Trajectory (left) and control command (right) histories

3.4 Case study

A case study was conducted to evaluate the effects of infrared image input and expert-action critic augmentation. All methods were tested under the same 4,000 validation scenarios. The initial flight-path angles were discretized at 1° intervals. An interception was considered successful when the missile–target distance was less than or equal to 3 m.

Four methods were compared: SMGC, TD3+BC without image input, TD3+BC with image input, and the proposed method. In addition to the success rate, the minimum miss distance and command smoothness were evaluated. Command smoothness was evaluated as follows:

$$J_{\Delta a} = \frac{1}{T-1} \sum_{k=2}^{T} |a_k - a_{k-1}|, \tag{16}$$

where a_k is the fin command and T is the number of simulation steps. A lower $J_{\Delta a}$ indicates smoother control behavior. The results are summarized in Table 2.

The state-only TD3+BC policy achieved a success rate of 88.2%. When infrared images were added, the success rate to 90.8%. This result shows that visual observations provide useful target information. The proposed method achieved the highest success rate of 92.7%, improving the TD3+BC result by 1.9 percentage points. The expert-action critic augmentation also improved value estimation. As shown in Fig. 2, the augmented critic converged to a higher Q-value. Compared with SMGC, the

proposed method reduced $J_{\Delta a}$ from 0.68 to 0.33 deg/step. Overall, the proposed method improved interception performance while producing smoother commands.

4. Conclusion

This paper proposed an expert data-driven offline reinforcement learning method for infrared image-based missile guidance and control. The policy was trained only from expert data, avoiding the safety and cost issues caused by online exploration. TD3+BC was used to reduce distributional shift, and expert-action critic augmentation was introduced to improve Q-value estimation. Simulation results showed that the proposed method achieved a 92.7% success rate, which was 1.2% higher than that of the SMGC baseline. In addition, the proposed method generated smoother control commands while maintaining stable interception behavior. These results demonstrate the feasibility of offline RL for practical infrared image-based missile guidance and control.

Acknowledgement

This work was supported by the Regional Innovation System & Education (RISE) program through the (Chungbuk Regional Innovation System & Education Center), funded by the Ministry of Education (MOE) and the (Chungcheongbuk-do), Republic of Korea (2026-RISE-11-014-01).

Contact Author Email Address

The contact author email addresses are as follows:
Seongyeon Kim: syk_@cbnu.ac.kr
Jongho Shin: jshin@cbnu.ac.kr
Hyeong-Geun Kim: hyeongkim@konkuk.ac.kr

Copyright Statement

The authors confirm that they, and/or their company or organization, hold copyright on all of the original material included in this paper. The authors also confirm that they have obtained permission, from the copyright holder of any third-party material included in this paper, to publish it as part of their paper. The authors confirm that they give permission, or have obtained permission from the copyright holder of this paper, for the publication and distribution of this paper as part of the ICAS proceedings or as individual off-prints from the proceedings.

References

[1] Rafael Yanushevsky. *Modern missile guidance*. CRC Press, 2018.

[2] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

[3] Tal Shima, Moshe Idan, and Oded M Golan. Sliding-mode control for integrated missile autopilot guidance. *Journal of guidance, control, and dynamics*, 29(2):250–260, 2006.

[4] Jianguo Guo, Yu Xiong, and Jun Zhou. A new sliding mode control design for integrated missile guidance and control system. *Aerospace Science and Technology*, 78:54–61, 2018.

[5] Chang-Hun Lee, Hyo-Sang Shin, Jin-Ik Lee, and Min-Jea Tahk. Zero-effort-miss shaping guidance laws. *IEEE Transactions on Aerospace and Electronic Systems*, 54(2):693–705, 2017.