Proximal Policy Optimization-Based Robotic Motion Control for High-Frequency Brazing Task of Copper Tube Joining

*Corresponding Author

1st Eugene Kim
Prupose-Built Mobility Group
Seonam Division
Korea Institute of Industrial Technology
Gwangju, Korea Republic
egkim@kitech.re.kr

2nd Hyunrok Cha
Prupose-Built Mobility Group
Seonam Division
Korea Institute of Industrial Technology
Gwangju, Korea Republic
hrcha@kitech.re.kr

3rd Meyonghwan Hwang*
Prupose-Built Mobility Group
Seonam Division
Korea Institute of Industrial Technology
Gwangju, Korea Republic
hana9215@kitech.re.kr
* Corresponding Author

***Abstract*—This work presents a learning-based motion control framework for automating high-frequency (HF) brazing of copper tube joints, a repetitive yet unstructured task still dominated by manual operation on refrigerator manufacturing lines. We develop a digital-twin training stack using PyBullet and an RB5-850 manipulator model, and formulate approach–braze–retreat behaviors as a continuous Cartesian control problem. A PPO actor–critic with convolutional visual encoders consumes RGB inputs (3×240×480) and outputs fine-grained end-effector increments (0.01 m resolution). To stabilize updates, we employ clipping-based surrogate objectives and separated learning rates for actor and critic, together with a replay buffer and mini-batch training. Simulation accelerates data collection and improves sample efficiency, yielding stable policy improvement within a few hundred epochs in simulation with monotonically increasing episodic rewards. We describe system integration toward real-world deployment, including induction-heating end-effector design and multimodal sensing (thermal/RGB/distance) for reward shaping and quality assurance. Ongoing work targets transfer from sim-to-real, closed-loop temperature/position control during brazing, and quantitative weld-quality evaluation under fixed-posture testbeds. The results indicate that PPO with a task-aware digital twin is a promising path to robust, generalizable HF brazing automation across product variants while reducing dependence on expert operators**

***Index Terms*—Reinforcement learning, Deep learning, Digital twin, Automation, Robotics**

I. INTRODUCTION

In refrigerator manufacturing lines, copper-tube based heat exchangers constitute a dense network of joints distributed across the product, where repetitive yet not strictly identical joining operations are performed due to fixture tolerances, product variants, and minor layout deviations on the line. Consequently, joint locations and approach angles exhibit small but persistent variations that complicate fixed-script automation strategies. In prevailing practice, high-frequency (HF) brazing of copper tubes remains predominantly manual, which poses challenges for cycle-time consistency, thermal window control, and workforce scalability on multi-model lines [1].

The canonical HF brazing procedure (i.e. approach, heat, and retreat) must satisfy constraints on heat input, joint wetting, and thermal impact to nearby components. Traditional rule-based motion macros and open-loop trajectories often struggle in unstructured or partially constrained settings, where geometric and thermal variations accumulate over time. Recent work in welding automation strengthens perception and planning via digitalization, notably using digital twins for seam tracking and process monitoring [2]–[4]. In parallel, vision-driven quality assessment has advanced through deep learning, enabling online detection of weld defects and process drifts that can support closed-loop adjustments [5], [6].

Reinforcement learning (RL), particularly Proximal Policy Optimization (PPO) [7], offers a practical balance of implementation simplicity and on-policy stability, and has become a default choice for continuous-control tasks. Emerging studies apply PPO-style policies to robotic welding manipulation and motion planning, reporting improved adaptability to task/geometric variations [8]. To bridge the reality gap, RL policies are increasingly trained in simulators and transferred to the shop floor via sim-to-real techniques (e.g., domain randomization/adaptation, sensor-modeling, and hybrid supervision), as surveyed in recent literature [9], [10]. These trends suggest that a task-aware digital twin coupled with PPO can endow brazing robots with robustness against multi-model variability and line disturbances.

This paper makes four contributions.

- We formulate HF brazing of copper-tube joints as a continuous Cartesian control problem that explicitly reflects the approach–heat–retreat phases on a refrigerator line.

This work was supported by "Development of Core Technologies for a Working Partner Robot in the Manufacturing Field" (KITECH EO-250005), and "Development of AI-based Equipment Control and Autonomous Manufacturing Operation Technology for High-quality Management of Non-standard Production Products in Home Appliance Factories" (KM-240409).

This is a preprint version.The final authenticated publication is available at 10.1109/ICCMA67641.2025.11369623

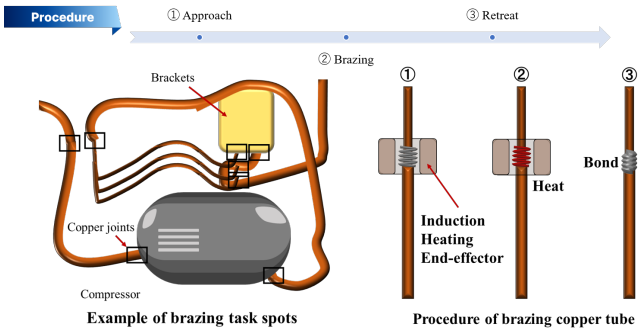

Fig. 1. HF brazing procedure: Left shows typical copper-tube joint locations and the target region. Right: (a) Approach, (b) Brazing, (c) Retreat/Finalize. The policy treats these phases as a continuous Cartesian control problem for robust motion under line and product variations

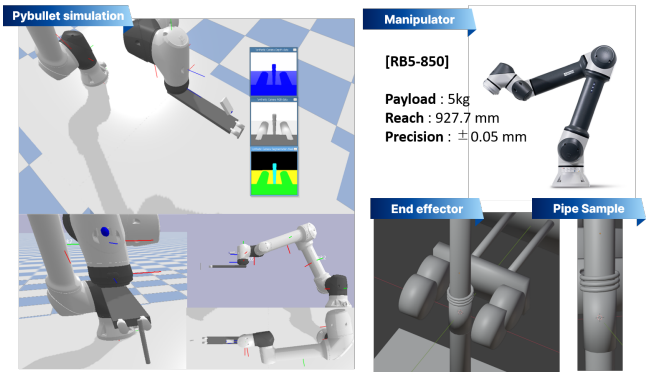

Fig. 2. Digital twin of the HF brazing setup. Left: PyBullet simulation of the RB5-850 robot performing joint approach and heating; synthetic camera streams show depth, RGB, and segmentation views. Right: hardware references including the RB5-850 manipulator the induction-heating end effector model, and the pipe sample with a representative joint geometry.

- We develop a PyBullet-based digital twin with an RB5-850 manipulator and train PPO actor–critic policies with visual inputs for fine-grained end-effector increments, targeting robustness to fixture/product variations.
- We outline an integration path toward real deployment, including induction-heating end-effector design and multi-modal sensing (RGB/IR/range) for reward shaping, monitoring, and quality assurance, aligned with recent digital-twin and vision-in-the-loop developments [2], [3], [5].
- We discuss sim-to-real considerations and evaluation protocols for weld-quality and temperature/position closed-loop control on a fixed-posture testbed, informed by current sim-to-real methodologies [9], [10].

II. Methodology and Experiment

A. Task Definition and Workflow

As shown in Fig. 1, we target the high-frequency brazing of copper-tube joints on refrigerator manufacturing lines, which is repetitive but not strictly identical across products and fixtures. The operational workflow is represented by a three-phase approach, including brazing and retreat, which reflects current manual procedures on the shop floor.

B. Digital-Twin Environment and Robot Platform

Policy learning is conducted in a digital-twin simulator built with PyBullet. As shown in Fig. 2, the manipulator model is RB5-850 (payload 5 kg, reach 927.7 mm, repeated precision ± 0.05 mm), equipped virtually with a brazing end effector and pipe samples representing joint locations. The repated precision of ± 0.05 mm is suitable for the brazing process because the radius of the target material is about 6.35 mm. This setup provides reproducible rollouts while preserving the geometric variability observed on the line. The RGB camera is installed on the end-effector of the 6th joint of the RB5-850. The resolution of the input images were set as 3x240x480 in order to lighten the size of the network.

C. Observations and Actions

We adopt vision-based observations using RGB images of size $3 \times 240 \times 480$. The perception stack follows a shared convolutional feature extractor for the actor and the critic, composed of three convolutional blocks (#filters: 32/64/64; kernel k=8, stride s=4; ReLU), followed by flattening and task-specific fully connected heads. The action space is continuous Cartesian end-effector increments along $\{\pm x, \pm y, \pm z\}$ with a unit step of 0.01 m, enabling fine-grained motion during each phase of the brazing task.

D. Reinforcement Learning Algorithm

We employ Proximal Policy Optimization (PPO) with clipped surrogate objectives for stable on-policy updates. Let π_θ and $\pi_{\theta_{\text{old}}}$ denote the new and old policies, and $A^{\pi_{\theta_{\text{old}}}}$ the advantage estimate. The objective is

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(r_t(\theta)A_t,\ \text{clip}\big(r_t(\theta), 1-\epsilon, 1+\epsilon\big)A_t\Big)\Big], \tag{1}$$

where $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ and ϵ is the clipping parameter. We use separate optimizer settings for the actor and the critic and train with mini-batches sampled from recent rollouts [7].

E. Training Setup

Simulation rollouts are parallelized with a $\times 16$ worker configuration to accelerate data collection and stabilize performance statistics. Training progression is monitored via episodic reward curves across epochs (e.g., 1, 50, 100, 200), confirming consistent policy improvement in simulation.

III. Results and Discussions

A. Learning Curves in Simulation

Figure 3 summarizes the training progress of the PPO controller in the PyBullet digital twin. The episodic total reward exhibits a steadily increasing trend over training, and the 10-episode moving average shows a consistent upward trajectory. In parallel, the training loss decreases and stabilizes after the early stage, indicating improved policy consistency. These

This is a preprint version.The final authenticated publication is available at 10.1109/ICCMA67641.2025.11369623

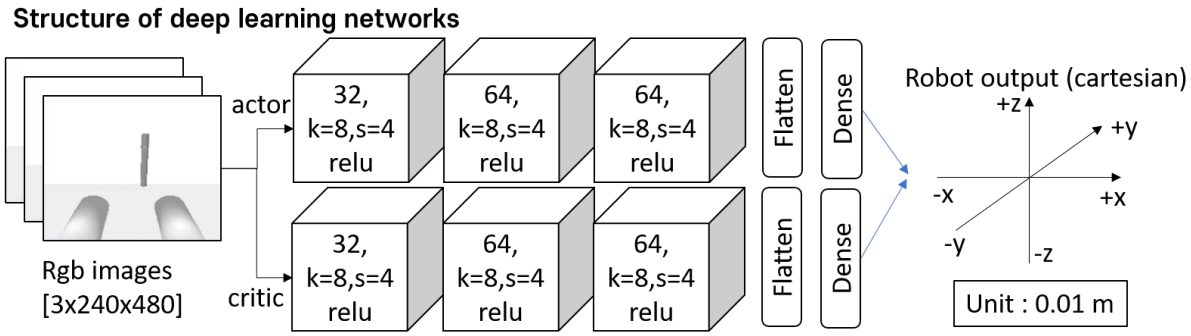

Fig. 3. Deep learning framework of the proposed method

TABLE I
TRAINING CONFIGURATION AND HYPERPARAMETERS USED FOR THE REPORTED RESULTS.

Training setup	
Episode count	500 episodes
Batch size (moves)	32
Learning buffer	1024 batches
RL hyperparameters	
Actor learning rate	1.0×10^{-5}
Critic learning rate	1.0×10^{-3}
Epochs per update	1
Discount factor γ	0.995
GAE λ	0.95
Penalty (shaping term)	-400
Mini-batch step size	10
Hidden units after flatten	256

observations align with the visual rollouts at key checkpoints (initial, epoch 1, 50, 100, 200), where the end-effector transitions from scattered motions to decisive approach–heat–retreat behaviors around the target joint.

B. Training Conditions

Table I lists the training setup used for the reported curves. Learning proceeds with small actor steps and a larger critic step under a relatively high discount factor to emphasize long-horizon returns.

C. Learning Progress and Loss Dynamics

Figure 4 summarizes the training progress of the PPO controller. The *Episode Total Reward* shows a generally increasing trend with reduced variance as training proceeds. Although early episodes exhibit noticeable fluctuations, the trajectory converges toward a higher reward regime by around 60–70 episodes, indicating improved task execution consistency.

The *Moving Average (10-episode window)* rises steadily from an initial low level to approximately the mid–to–high 7 range by the end of training. This sustained increase confirms that gains are not limited to isolated episodes but persist over consecutive trials.

The *Recent 10 Episodes Avg Reward* highlights the learning phases more clearly: a rapid improvement phase during the first 20–30 episodes, followed by a slower, saturation-like phase in which the average reward continues to climb but with diminishing increments. This pattern suggests the policy transitions from coarse alignment and approach behaviors to fine adjustments during heating and retreat.

The *Training Loss per Batch* reveals distinct roles of actor and critic. The critic loss drops by orders of magnitude early in training and then stabilizes, reflecting faster value-function fitting. The actor loss remains comparatively small and stable throughout, consistent with clipped PPO updates that constrain policy changes and help prevent instability.

Overall, these trends indicate that the learned policy improves both effectiveness and stability over time, moving from exploratory, high-variance behaviors to reproducible approach–heat–retreat executions. In practice, further smoothing may be achieved by modest tuning of batch composition (workers $\times$ steps), learning rates, or early stopping once the moving average plateaus.

IV. CONCLUSION

This paper presented a PPO-based motion control framework for high-frequency brazing of copper-tube joints in refrigerator manufacturing. We modeled the task as a continuous Cartesian control problem aligned with the three operational phases—approach, brazing, and retreat—and trained vision-conditioned policies in a PyBullet digital twin using an RB5-850 manipulator. The learned policy exhibited steadily improving rewards and stabilized losses, and qualitative rollouts showed increasingly reliable alignment, localized heating, and consistent retreat behavior.

The study contributes: (i) a task-aware digital-twin pipeline for HF brazing with vision-based PPO control; (ii) a lightweight network and training configuration that achieves stable learning under moderate parallel rollouts; and (iii) an integration path toward deployment, including an induction-heating end effector and a sensing suite for monitoring and future feedback control.

There are limitations. First, current results are simulation-based and do not yet quantify weld quality or thermal margins on hardware. In detail, challenges lies such as modelling complex thermal dynamics in a general-purpose physics engine like PyBullet. Second, policy robustness has been evaluated

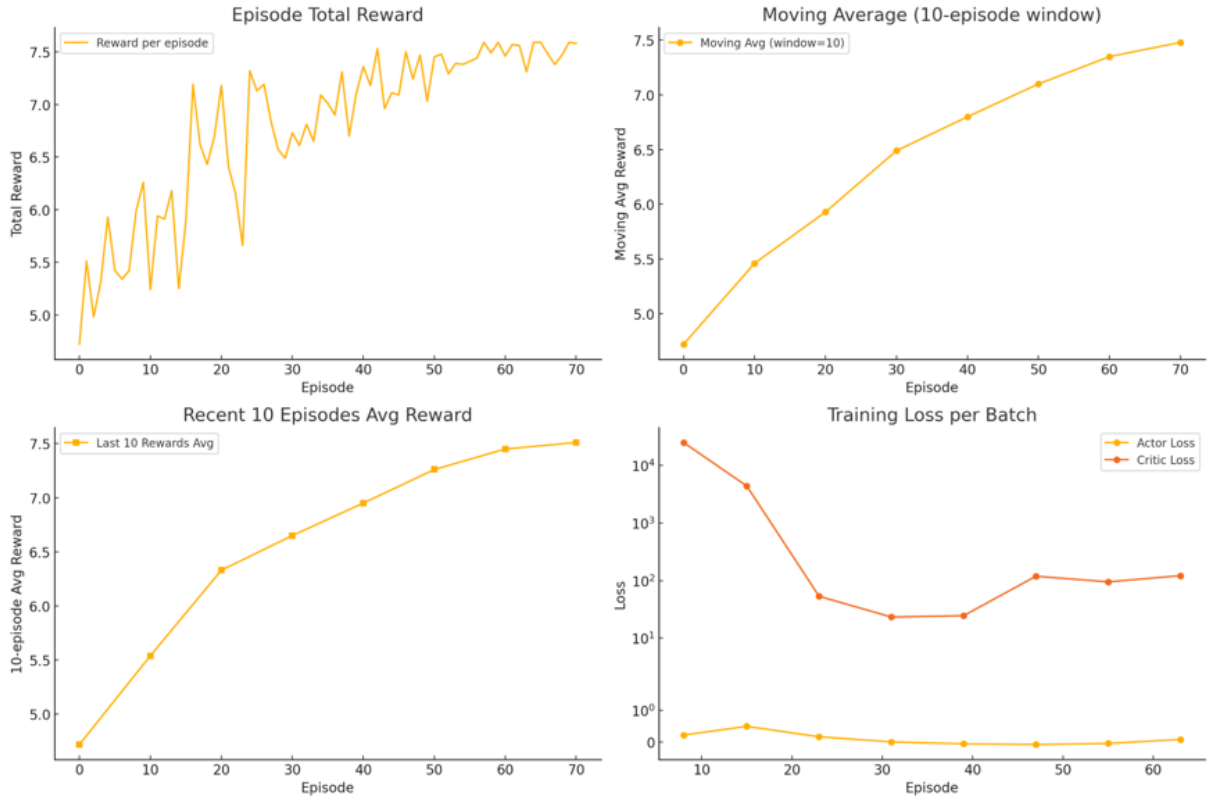

Fig. 4. Training curves: episode total reward, 10-episode moving average, recent-10-episode average reward, and training loss per batch for actor and critic. Curves show monotonic improvement in rewards and early rapid decrease of critic loss followed by stabilization

under geometry variability but not under broader disturbances such as coil–pipe misalignment beyond tolerance or large lighting/temperature shifts. Third, the reward does not yet incorporate direct quality signals.

Future work will focus on bench-top validation with closed-loop temperature/position control, quantitative weld-quality metrics, and sim-to-real transfer via domain randomization and sensor modeling. We also plan to expand the observation set with thermal and range signals, and to benchmark against alternative on-policy and off-policy algorithms under identical digital-twin conditions.

Acknowledgment

This work was supported by ”Development of Core Technologies for a Working Partner Robot in the Manufacturing Field” (KITECH EO-250005), and ”Development of AI-based Equipment Control and Autonomous Manufacturing Operation Technology for High-quality Management of Non-standard Production Products in Home Appliance Factories” (KM-240409). The authors express their gratitude to project leader Bonggu Kim, Changhoon Ryu of DH Global for his valuable support and contributions.

References

[1] E. Kim, H. Cha, M. Hwang, and Y. Kim, “Reinforcement learning-based optimization of robotic motion for high-frequency brazing task of copper tube joining,” in *2025 IEEE International Conference on Advanced Robotics and its Social Impacts (ARSO)*, pp. 265–270, IEEE, 2025.

[2] Y. Kang and R. Chen, “Welding robot automation technology based on digital twin,” *Frontiers in Mechanical Engineering*, vol. 10, 2024.

[3] S. Wang, Y. Jiao, L. Wang, W. Wang, X. Ma, Q. Xu, and Z. Lu, “Research on the digital twin system of welding robots driven by data,” *Sensors*, vol. 25, no. 13, p. 3889, 2025.

[4] Q. Qin, Z. Liu, R. Zhong, X. V. Wang, L. Wang, M. Wiktorsson, and W. Wang, “Robot digital twin systems in manufacturing: Technologies, applications, trends and challenges,” *Robotics and Computer-Integrated Manufacturing*, vol. 97, p. 103103, 2026.

[5] M. Trusiak *et al.*, “A deep learning-based machine vision system for online monitoring and surface quality evaluation during welding,” *Sensors*, vol. 25, no. 16, p. 4997, 2025.

[6] H. Liu *et al.*, “Deep convolutional neural network for weld defect classification in radiographic images,” *Heliyon*, 2024. Add DOI and pages if available.

[7] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms.” https://arxiv.org/abs/1707.06347, 2017. arXiv:1707.06347.

[8] Y. He, J. Zhang, M. Wei, Y. Guo, Y. Wang, and Y. Huang, “Reinforcement learning based motion planning for robotic arm of welding,” in *Robotics and Autonomous Systems and Engineering Applications of Computational Intelligence (LSMS 2024, ICSEE 2024)* (J. Gu, F. Hu, H. Zhou, Z. Fei, and E. Yang, eds.), vol. 2220 of *Communications in Computer and Information Science*, pp. 59–70, Singapore: Springer, 2024.

[9] W. Zhao, J. P. Queralta, and T. Westerlund, “Sim-to-real transfer in deep reinforcement learning for robotics: a survey,” *arXiv preprint arXiv:2009.13303*, 2020.

[10] L. Da, J. Turnau, T. P. Kutralingam, A. Velasquez, P. Shakarian, and H. Wei, “A survey of sim-to-real methods in rl: Progress, prospects and challenges with foundation models,” *arXiv preprint arXiv:2502.13187*, 2025.