

Reinforcement Learning of Laser Scanning Policies for Uniform Transient Thermal Control in the Selective Laser Thermoregulation Method

Tomomasa Ohkubo^{*1,2}, Shota Ui², Yusaku Kawarazaki², Yuuta Sano¹, and Ei-ichi Matsunaga¹

¹ *Department of Mechanical Engineering, School of Engineering, Tokyo University of Technology, Japan*

² *Sustainable Engineering Program, Graduate School of Engineering, Tokyo University of Technology, Japan*

^{*} *ookubotmms@stf.teu.ac.jp*

Selective Laser Thermoregulation (SLT) enables rapid heating and dynamic temperature control by scanning a laser beam over a target surface. Previous studies have developed artificial intelligence methods to estimate irradiation conditions from temperature distributions; however, these approaches assumed stationary laser irradiation and could not determine scanning paths. In this study, we propose a reinforcement learning framework to optimize laser-scanning paths to achieve a uniform temperature distribution. The laser spot is modeled as an agent, and its motion is learned through interaction with the thermal field. Neural Network-based Q-learning is employed to approximate the action-value function under continuous state conditions. Numerical simulations were conducted using a 5×5 grid model including conduction, convection, and radiation. The method successfully achieved an average temperature above 400 °C and reduced the standard deviation to approximately 20 °C. The trained agent learned that corner regions require intensive heating due to enhanced heat loss. These results demonstrate that reinforcement learning can autonomously discover physically meaningful scanning strategies for SLT systems.

DOI:10.2961/jlmn.2026.03.2001

Keywords: laser heating, reinforcement learning, neural network-based Q-learning, temperature control, selective laser thermoregulation (SLT) method

1. Introduction

High-heat-resistant materials, such as SiC/SiC ceramic matrix composites, are expected to improve the efficiency of next-generation aircraft engines. To evaluate their reliability, accelerated thermal testing under rapid heating and cooling conditions is required.

Laser heating methods are effective because they enable localized and rapid thermal cycling. The Selective Laser Thermoregulation (SLT) method has been proposed as a technique for dynamically controlling the temperature distribution by scanning a laser beam over a target surface. The fundamental concept and system design of the SLT method have been established in a previous study [1].

Previous studies have focused on determining laser irradiation conditions, such as laser power, irradiation position, and spot size, using artificial intelligence. For example, artificial intelligence has been applied to estimate laser power and multiple irradiation parameters from temperature distributions [2][3]. However, these methods assume that the position of the laser spot is fixed and cannot optimize scanning paths. In these approaches, multiple irradiation conditions can produce similar temperature distributions, reflecting the non-uniqueness of the inverse problem [3].

In practical SLT systems, the scanning path significantly affects thermal distribution because heat accumulation and dissipation strongly depend on the movement of the laser spot. Recent studies have experimentally shown that the temperature distribution fluctuates both spatially and

temporally depending on the laser-scanning conditions [4]. Therefore, determining an optimal scanning strategy remains a crucial, unresolved problem.

This study aims to develop a reinforcement learning-based method to determine laser scanning paths automatically. The laser spot is treated as an agent, and its motion is optimized through interaction with the temperature field to achieve a uniform temperature distribution.

2. Method

2.1 Problem formulation

In this study, the optimization of the laser scanning path is formulated as a reinforcement learning problem. The laser spot is modeled as an agent that interacts with the target surface's thermal field.

The state S_t is defined by the combination of the current laser position and the temperature distribution over all cells at time step t . The action space A_t consists of five discrete choices: moving the laser spot up, down, left, right, or remaining at the current position (stay). The schematic of the reinforcement learning environment is shown in Fig. 1.

The objective of the learning process is to achieve a temperature distribution that satisfies both a target average temperature and spatial uniformity. In this study, the target temperature is set to 400 °C, and the standard deviation of the temperature distribution must be below 40 °C.

Through repeated interactions with the environment, the agent learns a scanning policy that maximizes cumulative reward that corresponds to these objectives.

The target surface is discretized into a 5×5 grid, and the laser spot is modeled as an agent that moves between cells. The temperature distribution evolves according to laser irradiation, thermal conduction, and heat transfer, and the agent receives rewards based on temperature, uniformity, and movement cost. An episode is terminated when the average temperature exceeds the target temperature and the standard deviation falls below the target value, or when the maximum number of laser movements (50,000 steps) is reached. The program is implemented using PyTorch 2.12 [5].

2.2 Thermal model

The temperature distribution on the target surface is calculated using a numerical thermal model. This numerical model is based on a simplified thermal analysis developed in the previous study [6]. The surface is discretized into a 5×5 grid, and the temperature T of each cell is updated in time.

The governing equation is based on the heat balance of each cell, including laser heating, conduction, convection, and radiation. The temperature evolution is calculated based on an energy balance model as follows:

$$\rho CV \frac{\partial T}{\partial t} = V\lambda \nabla^2 T - A[h(T - T_0) + \sigma\epsilon(T^4 - T_0^4)] + \alpha IA. \quad (1)$$

where ρ , C , and λ are the density, specific heat, and thermal conductivity, respectively. A and V denote the surface area and volume of a cell, h is the heat transfer coefficient, ϵ is the emissivity, σ is the Stefan–Boltzmann constant, α is the absorptivity, and I is the laser power density. The terms on the right-hand side represent heat conduction within the material and heat loss due to convection and thermal radiation. The last term indicates laser energy absorption.

The computational domain was a $10 \text{ mm} \times 10 \text{ mm} \times 3 \text{ mm}$ target region. The in-plane region was discretized into 5×5 square cells. For this proof-of-concept study, the entire computational domain was assumed to be a small $10 \text{ mm} \times 10 \text{ mm}$ plate rather than a sub-region of a larger structure. This simplified configuration was adopted to evaluate the effectiveness of the proposed reinforcement-learning-based scanning policy under controlled thermal conditions. Each cell therefore had an in-plane size of $2 \text{ mm} \times 2 \text{ mm}$. The present simulation considered only the heating process until the prescribed target conditions were achieved; the subsequent temperature-holding process was not included.

The temperature in each cell was assumed to be uniform through the thickness direction. This approximation was introduced to reduce the computational cost and to focus on the optimization of in-plane scanning paths.

The boundary conditions were modeled by convection and thermal radiation from the exposed surfaces, while heat conduction between neighboring cells was included in the in-plane direction. The boundary cells therefore experience larger heat losses than the interior cells, especially at the corners.

The 5×5 grid was selected as a proof-of-concept discretization that can represent the essential in-plane thermal features of the SLT process, including center, edge, and

corner regions, while keeping the state space sufficiently small for comparison with the TD method. Although a finer grid would better approximate a practical scanning system, it would significantly increase the dimensionality of the state space and the computational cost. The effect of grid resolution will be investigated in future work.

The laser spot was allowed to move to a neighboring cell once per simulation time step. Since the cell size was 2 mm and the time step was 0.001 s , the corresponding scanning speed was 2 m/s .

In this simulation, the target material was assumed to be stainless steel SUS304, as in previous studies. Although the SLT method was originally developed for thermal testing of SiC/SiC ceramic matrix composites, previous experimental studies have mainly employed SUS304 stainless steel because SiC/SiC CMCs are expensive and not suitable for extensive preliminary investigations. To maintain consistency with the existing experimental database and the previously validated thermal model, SUS304 was selected as the target material in this study. The objective of this work is not to investigate material-specific thermal behavior but to evaluate the feasibility of reinforcement-learning-based optimization of laser scanning paths. The physical properties and dimensions used for the numerical thermal model are summarized in Table 1.

Laser irradiation is applied only to the cell containing the agent. The temperature of all cells is updated at each time step, allowing the spatial temperature distribution to evolve dynamically according to the scanning motion.

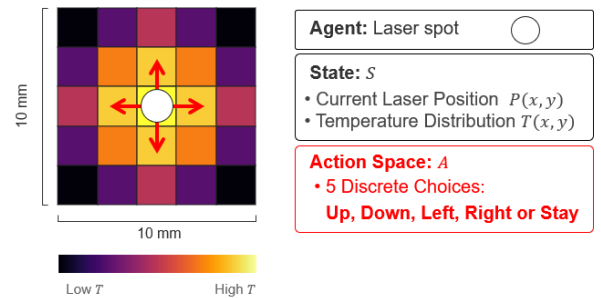


Fig. 1 Schematic of the reinforcement learning environment.

Table 1 Thermal model parameters

Parameter	Symbol	Value	Unit
Density	ρ	7930	kg/m ³
Specific heat	C	590	J/(kg·K)
Thermal conductivity	λ	16.7	W/(m·K)
Laser power density	I	1.99×10^7	W/m ²
Absorptivity	α	0.9	—
Emissivity	ϵ	0.9	—
Ambient temperature	T_0	300	K
Heat transfer coefficient	h	20	W/(m ² ·K)
Cell size	$\Delta x, \Delta y$	2.0×10^{-3}	m
Time step of simulation	Δt	0.001	s

2.3 Reinforcement learning algorithm

Two reinforcement learning approaches are considered in this study: the Temporal Difference (TD) method [9]-[11] and Neural Network-based Q-learning (NNQ) [12][13].

The TD method explicitly updates the action-value function $Q(S_t, A_t)$ for all state-action pairs. This approach is applicable when the number of states is limited and can be fully enumerated.

In contrast, NNQ is employed to handle the high-dimensional state space, which includes both the spatial temperature distribution and the laser position. In NNQ, the Q-function is approximated by a neural network rather than explicitly stored.

The structure of the neural network used in NNQ is shown in Fig. 2. The neural network input consists of the laser position and the temperature distribution, both converted into a one-dimensional vector. The network outputs the expected values of all possible actions.

The Q-values are updated using the standard Q-learning rule as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha_{\text{learn}} \left[R_t + \gamma \max_A Q(S_{t+1}, A) - Q(S_t, A_t) \right]. \quad (2)$$

where α_{learn} is the learning rate, γ is the discount factor, R_t is the reward at time step t . S_t and S_{t+1} represent the current and next states, respectively, and A denotes the possible actions. The term $\max_A Q(S_{t+1}, A)$ represents the maximum expected future reward over all possible actions at the next state.

The action-value function $Q(S_t, A_t)$ is approximated using a multilayer perceptron neural network. The network architecture consists of an input layer representing the state S_t , two hidden layers with 100 and 50 nodes, respectively, and an output layer with 5 nodes corresponding to the action space A_t . ReLU was employed as the activation function for the hidden layers, and a linear function was used for the output layer.

To balance exploration and exploitation, the ϵ -greedy strategy was employed during the training phase, where the exploration rate ϵ decays per episode. The specific hyperparameters for the RL are listed in Table 2. Training was terminated after 100 episodes. Each episode was terminated when the average temperature exceeded the target temperature and the standard deviation became lower than the target value, or when the maximum number of laser movements, 50,000, was reached. The number of training episodes was fixed in this study to compare all methods under identical training conditions.

This framework allows the agent to approximate optimal policies even when the state space is continuous or very large.

2.4 Reward function

The reward function is designed to guide the agent toward achieving both temperature control and efficient scanning. It is desirable to preheat areas with low temperatures prior to irradiation, as this results in a lower standard deviation in the temperature distribution and fewer laser movements. In addition, it is preferable for the temperature distribution's standard deviation to be small after laser irradiation and for the total number of laser movements to be kept to a

minimum. Therefore, the reward function is defined as follows:

$$R = 3 - \left(C_T \frac{T_{\text{cell}}^{t-1}}{T_{\text{target}}} + C_\sigma \frac{\sigma^t}{\sigma_{\text{target}}} + C_N \frac{N^t}{N_{\text{target}}} \right). \quad (3)$$

where T_{cell}^{t-1} is the temperature of each cell at the previous time step, σ^t and N^t are the standard deviation of the current temperature distribution, and the current cumulative number of laser movements. T_{target} , σ_{target} , N_{target} are the target values of the temperature, standard deviation, and the cumulative number of laser movements. C_T , C_σ and C_N are weight coefficient of each factor.

The temperature-related term is not intended to directly evaluate the proximity to the target temperature. Instead, it acts as a reward-shaping term that encourages the agent to preferentially irradiate relatively cold regions. Accordingly, T_{cell}^{t-1} , representing the temperature before the current laser irradiation, is used instead of the post-irradiation temperature at time step t . By directing the laser toward low-temperature cells before irradiation, the agent can reduce spatial temperature variation and achieve a more uniform heating process.

This formulation reflects three essential goals:

- prioritizing the heating of relatively cold regions
- minimizing the spatial variation of temperature
- reducing unnecessary laser movements

Minimizing the number of laser movements is also important from an operational viewpoint. In practical scanning systems, such as galvanometer mirrors or XY stages, excessive movements and frequent direction changes may increase positioning errors, response delays, vibration, and control load. Therefore, the reward function includes a movement-related term to promote scanning policies that achieve uniform heating while avoiding unnecessary motion.

Therefore, by combining these terms, the reward function enables the agent to balance heating performance and operational efficiency. The specific values of the weight coefficients and target parameters are detailed in Table 3.

3. Results

3.1 Comparison of scanning strategies

The performance of the proposed method was evaluated by comparing four scanning strategies: random motion, counterclockwise (CCW) scanning, the TD method, and NNQ. Note that "Random" selects the option with the best results.

The random and CCW methods rely on predefined or non-optimized patterns, while TD and NNQ adaptively update the scanning policy through learning.

The scanning paths obtained for each method are shown in Fig. 3. While the CCW method naturally follows a regular path, the TD method showed paths that varied as much as random movements. In contrast, the NNQ method appears to draw small squares or stop at corners, suggesting minimal wasted movement.

The temperature distributions obtained by different methods are shown in Fig. 4. The NNQ method produced the most uniform temperature field, with smaller temperature gradients between the center and boundary regions. In contrast, the CCW and TD methods resulted in localized under-heated and over-heated regions.

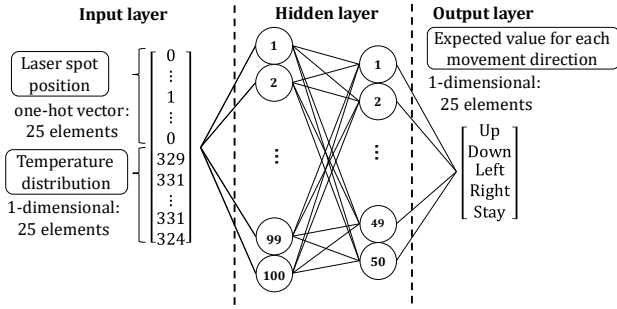


Fig. 2 Neural network used in this study.

Table 2 Reinforcement learning hyperparameters

Parameter	Symbol	Value	Unit
Learning rate	α_{learn}	0.01	—
Discount factor	γ_{learn}	0.9	—
Epsilon decay factor	—	0.999	—
Number of episodes	—	100	—
Hidden layers	—	100, 50	nodes

Table 3 Reward function parameters

Parameter	Symbol	Value	Unit
Target temperature	T_{target}	673	K
Target standard deviation	σ_{target}	40	K
Target number of laser movements	N_{target}	10^7	—
Temperature weight coefficient	C_T	0.8	—
Standard deviation weight coefficient	C_σ	0.1	—
Laser movement weight coefficient	C_N	0.1	—

3.2 Thermal performance

A comparison of thermal performance and scanning efficiency for each method is shown in Fig. 5. The red bar on the left axis shows the number of laser movements. In contrast, the standard deviation is shown by the blue bar on the right axis. The NNQ method achieves the lowest standard deviation while maintaining a relatively small number of laser movements. In contrast, the TD method increases the number of laser movements to improve uniformity, and the predefined CCW method fails to achieve sufficient temperature uniformity. These results are consistent with the temperature distributions shown in Fig. 4, where the NNQ method produced the most spatially uniform heating pattern.

The proposed reinforcement learning approach using QNN successfully achieved the target heating conditions. The average temperature exceeded 400 °C, and the standard deviation of the temperature distribution was reduced to approximately 20 °C, which is significantly lower than the target value.

This result indicates that the proposed method can efficiently produce a uniform temperature distribution within the specified region.

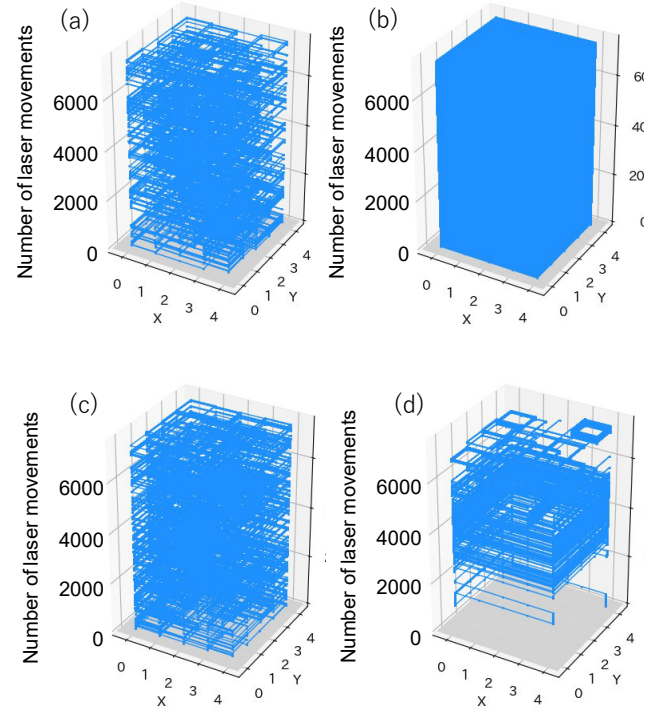


Fig. 3 Representative laser scanning paths obtained by different methods. (a) Random (successful case), (b) Counterclockwise, (c) TD method, (d) NNQ.

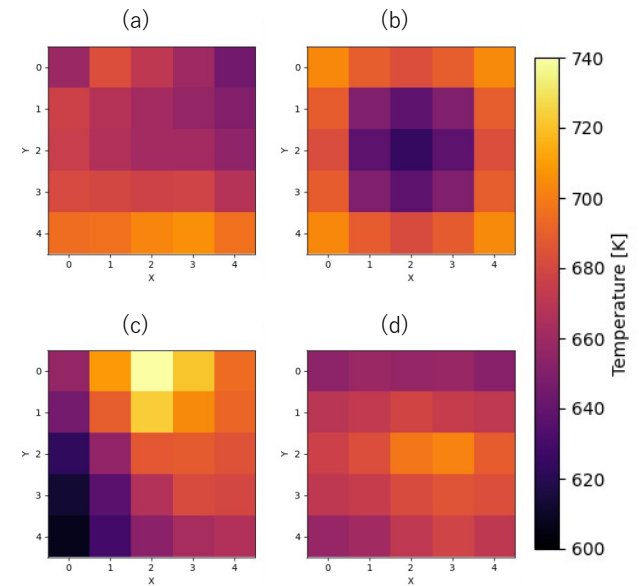


Fig. 4 Final temperature distributions obtained by different methods. (a) Random (successful case), (b) Counterclockwise, (c) TD method, (d) NNQ.

3.3 Learned scanning behavior

The spatial distribution of laser irradiation frequency is shown in Fig. 6. The irradiation frequency represents the number of times the agent selects each cell during the scanning process. Analysis of the learned scanning paths reveals

that the NNQ-based agent preferentially irradiates the corner regions of the target surface.

This behavior arises because heat loss is greater near the boundaries, particularly at the corners, where conduction and convection effects are more pronounced. As a result, these regions require more frequent heating to maintain a uniform temperature distribution.

Importantly, this strategy was not explicitly programmed but autonomously discovered through reinforcement learning, demonstrating the proposed method's ability to identify physically meaningful solutions.

4. Discussion

The present study extends previous work on SLT, in which artificial intelligence was used to estimate irradiation conditions under stationary assumptions [2][3], by addressing the optimization of dynamic scanning paths. Earlier approaches focused on inverse problems, such as estimating laser power or irradiation parameters from a desired temperature distribution. However, these methods did not consider the dynamic nature of laser motion.

In contrast, the proposed reinforcement learning framework directly optimizes the sequence of laser movements, enabling real-time adaptation to evolving temperature distributions.

A simple analytical strategy, such as allocating irradiation time in proportion to the local heat dissipation, may be effective under steady-state or quasi-steady-state assumptions. However, the SLT process considered in this study is inherently transient: the temperature field evolves depending on the past irradiation history, heat conduction from neighboring cells, boundary heat losses, and the future sequence of laser movements. Therefore, the required irradiation at each position cannot be determined solely from the instantaneous heat dissipation. The motivation for using reinforcement learning is that it can optimize a sequential scanning policy through interaction with the evolving thermal field, rather than prescribing a fixed irradiation distribution.

Furthermore, the results suggest that reinforcement learning can effectively handle the inherent non-uniqueness of the thermal control problem. Multiple irradiation conditions may produce similar temperature distributions; however, the proposed method identifies efficient scanning strategies that satisfy both thermal and operational constraints.

Another important finding is that the learned scanning policy reflects underlying heat transfer mechanisms. This behavior is consistent with experimental observations that temperature variations depend on local heat loss during laser scanning [4]. The increased irradiation frequency at the corners corresponds to the physical tendency of these regions to cool faster. This agreement between learned behavior and physical intuition indicates that the thermal model and the reinforcement learning formulation are consistent with the system's actual physics. Such consistency also supports the validity of simplified thermal models used for analyzing SLT systems [6]. The learned tendency to preferentially heat colder regions is consistent with the reward formulation and contributes to suppressing spatial temperature non-uniformity.

The present study used SUS304 as a benchmark material for consistency with previous SLT experiments. Since the

proposed framework only requires thermal model parameters, it can be extended to SiC/SiC CMCs and other materials by replacing the corresponding thermophysical properties.

On the other hand, in experimental implementation, deviations from the simulated temperature field may arise from temperature-dependent material properties, variations in absorptivity due to surface conditions, measurement errors in infrared thermometry, laser response delay, and acceleration or velocity limits of the scanning optics. The present study should therefore be regarded as a first step toward scanning policy design using a simplified thermal model. For practical implementation, the learned scanning policy could be combined with real-time temperature feedback, and further fine-tuned using experimental data.

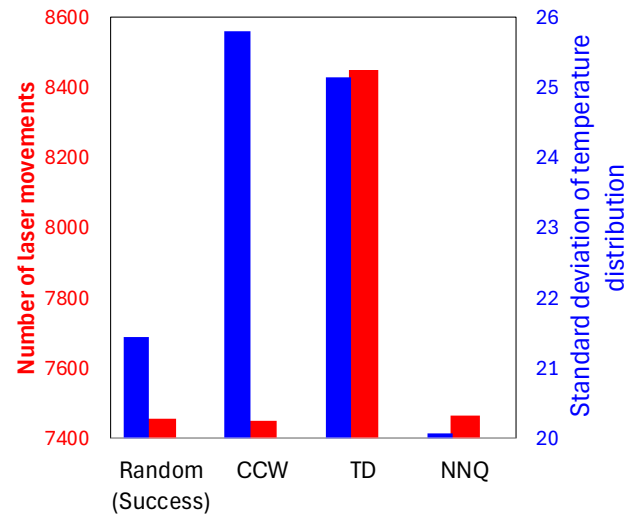


Fig. 5 Comparison of the number of laser movements and the standard deviation of the temperature distribution for different scanning methods.

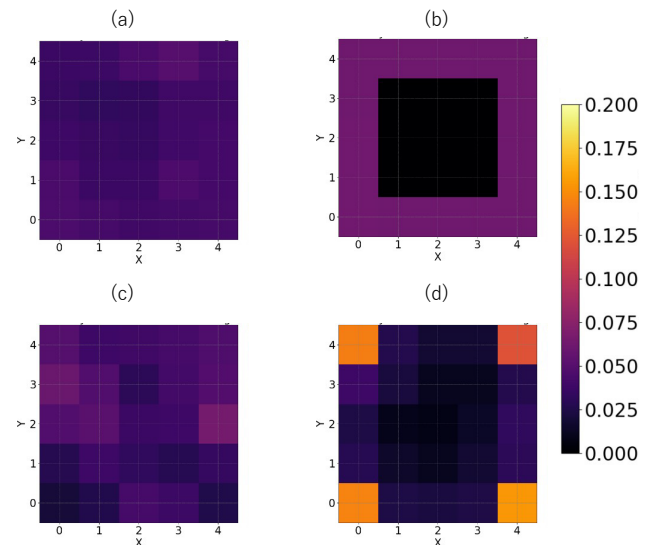


Fig. 6 Probability distribution of laser irradiation positions obtained from the trained scanning policy. (a) Random (successful case), (b) Counterclockwise, (c) TD method, (d) NNQ.

5. Conclusion

A reinforcement learning-based method for optimizing laser scanning paths in the SLT system has been proposed.

The NNQ approach successfully achieved a uniform temperature distribution while reducing unnecessary scanning movements. The method autonomously learned physically meaningful heating strategies, such as prioritizing corner regions with significant heat loss.

These results demonstrate that reinforcement learning is a promising tool for the automatic design of laser-scanning strategies and can provide a new framework not only for thermal control but also for laser-based material processing systems such as laser cutting, welding, drying, and additive manufacturing.

Acknowledgments

Amano Institute of Technology partially supports this study.

References

- [1] H. Koshiji, T. Ohkubo, K. Azato, Y. Kameda, E. Matsunaga, T. Dobashi, N. Shichijo, K. Goto, M. Sato, C. Fujiwara, and Y. Kagawa: *J. Laser Micro/Nanoengineering*, 15, (2020) 174.
- [2] M. Nakaone, T. Ohkubo, Y. Ueno, K. Goto, and Y. Kagawa: *J. Laser Micro/Nanoengineering*, 16, (2021) 84.
- [3] M. Nakaone, T. Ohkubo, Y. Ueno, K. Goto, and Y. Kagawa: *J. Laser Micro/Nanoengineering*, 17, (2022) 194.
- [4] S. Kanai, T. Ohkubo, S. Ui, Y. Kawarazaki, E. Matsunaga, K. Goto, and Y. Kagawa: *J. Laser Micro/Nanoengineering*, 19, (2024) 140.
- [5] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala: *Adv. Neural Inf. Process. Syst.*, 32, (2019) 8024. (Conference Proceedings)
- [6] T. Ohkubo, S. Kanai, S. Ui, Y. Kawarazaki, E. Matsunaga, K. Goto, and Y. Kagawa: *J. Laser Appl.*, 37, (2025) 042017.
- [7] M. P. Appleby, D. Zhu, and G. Morscher: *Surface and Coatings Technology*, 284, (2015) 318.
- [8] T. Whitlow et al. *Composite Structures*, 210, (2019) 179.
- [9] R. S. Sutton: *Machine Learning*, 3, (1988) 9.
- [10] C. Watkins and P. Dayan: *Machine Learning*, 8, (1992) 279.
- [11] R. S. Sutton and A. G. Barto: "Reinforcement Learning: An Introduction" (MIT Press, Cambridge, 2018) p.119.
- [12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis: *Nature*, 518, (2015) 529.
- [13] L. J. Lin: *Machine Learning*, 8, (1992) 293.

(Received: June 12, 2026, Accepted: October 12, 2026)