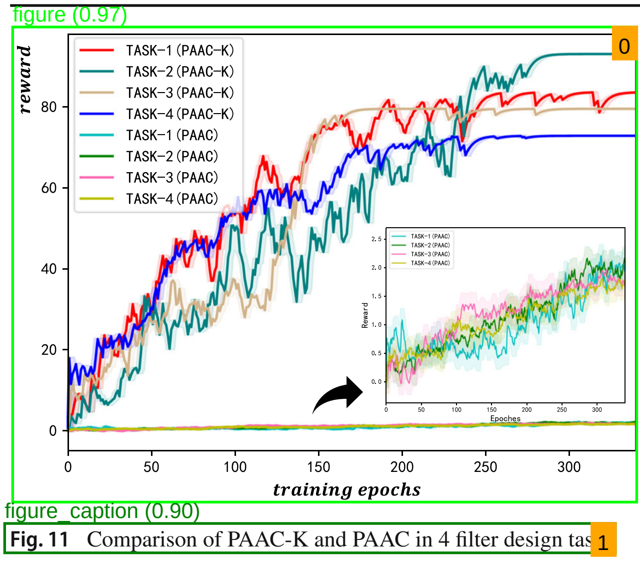

Fig. 11 Comparison of PAAC-K and PAAC in 4 filter design tasks

In Task 2, the target centre frequency is 3.15 GHz, and the passband frequency is from 3.0 GHz to 3.3 GHz. The filter's centre frequency in Fig. 7 is too far from the design target to be an initial state. As shown in Fig. 10 d-f, the initial state is extracted with the stepwise training method (Steps 1 to 11). This is a process automatically extracted from the database. The agent has learned to broaden the resonator to transfer the centre frequency left to 3.20 GHz. The agent also learns to adjust the coupling coefficient and reduce the reflection loss and insertion loss to complete the filter design (Step 12 to 19).

In Task 3, as shown in Fig. 10 g-i, the centre frequency transferred to 2.42 GHz from 2.70 GHz, that of the original state (Fig. 7). Similarly, in Task 4, as shown in Fig. 10 j-l, the centre frequency is transferred to 2.17 GHz from 2.70 GHz. The agent has completed the design of both tasks and met the design specifications. The results in any of the four example tasks using a GPU converge within 300 times and 8 hours. A filter with high-performance standards that demonstrate the PAAC-K architecture's advantages is obtained. Without clustering, there are many grid points, solutions and exploration spaces. The agent has difficulty understanding the filter characteristics. The increasing reward is difficult, and the algorithm has difficulty converging in adequate time (Fig. 11). The agent selects the "end" action to indicate the end of the current iteration. Note that the agent selects the "end" action because of numerous explorations and finds that opting for other actions is always further away from the goal in the long run, which means a lower rewards. Thus, the agent chooses to end the design. To better understand the agent's work, a video can be found in the supplementary information.

Based on reinforcement learning theory, the exploration space increases exponentially with the depth of detection. As shown in Table 3, the exploration space rises exponentially as the task's difficulty grows. In Task 2, the exploration space of the traditional reinforcement learning algorithm is 116^{17}. Without using the stepwise training method, the exploration space is 4^{17}. The exploration space of PAAC-K is $4^{11} + 4^{6}$.

5 Main contribution

The main contributions of this work: (1) This work presents the PAAC-K model, which can realize superhuman intelligence that automated the filter design of irregular structures. (2) The excessive action space makes it difficult for reinforcement learning algorithms to converge in the filter design task. We used a clustering algorithm to achieve the clustering of the differential S_{11} curves. This approach can effectively reduce the action space, so as to converge quickly. (3) The overlap region size is used as the reward function in reinforcement learning algorithms, which can provide a reference for other types of filter designs, such as low-pass and high-pass filters. (4) We use the stepwise training method to avoid repeated exploration in the different tasks of filter designs.

6 Conclusion

This paper presents a parallel model based on clustering Reinforcement learning named PAAC-K. The model achieves an end-to-end closed-loop training process of designing, simulating, and optimizing. The PAAC-K model is based on clustering results for learning and can be used in the automatic design of irregular structures, which is proven with four

Table 3 Exploration space in different tasks

Task	Traditional reinforcement learning	Clustering-reinforcement learning withour Stepwise Training method	PAAC-K
1	116^{6}	4^{6}	4^{6}
2	116^{17}	4^{17}	$4^{11} + 4^{6}$
3	116^{12}	4^{12}	$4^{6} + 4^{6}$
4	116^{16}	4^{16}	$4^{11} + 4^{5}$