Appendix A. Details of RNN Model and its Optimization

The RNN model contains GRU layer with 64 hidden units. We employ Proximal Policy Optimization (PPO) to optimize the parameters of the RNN. Adam is used for optimizing the parameters of the RNN model, with a learning rate of 0.001. The number of epochs for PPO is set to 4, the clip parameter is set to 0.1, the mini-batch size is set to 4, the coefficient of value function loss is set to 0.5 and the entropy coefficient is set to 0.01.

Appendix B. Details of Training Settings

During network architecture search, we optimize the multi-stage model for 6 epochs using the training dataset to approximate $\boldsymbol{\theta}_a^*$. 10k models are sampled from the architecture search space $\mathcal{Z}$ for each experiment. Then the models with top-10 rewards are used to apply the full training.