Appendices

A. Environment Settings

A.1. States and Observations

We mainly follow the settings of SMAC (Samvelyan et al., 2019). At each time step, agents receive local observations within their field of view. This encompasses information about the map within a circular area around each unit with a radius equal to the sight range. The sight range makes the environment partially observable for each agent. An agent can only observe other agents if they are both alive and located within its sight range. Hence, there is no way for agents to distinguish whether their teammates are far away or dead. The feature vector observed by each agent contains the following attributes for both allied and enemy units within the sight range: distance, relative x, relative y, health, shield, and unit type. All Protos units have shields, which serve as a source of protection to offset the damage and can regenerate if no new damage is received. The global state is composed of the joint observations but removing the restriction of sight range, which could be obtained during training in the simulations. All features, both in the global state and in individual observations of agents, are normalized by their maximum values.

A.2. Action Space

We follow the settings of SMAC (Samvelyan et al., 2019). The discrete set of actions which agents are allowed to take consists of move[direction], attack[enemy id], stop, and no-op. Dead agents can only take no-op action while live agents cannot. Agents can only move with a fixed movement amount 2 in four directions: north, south, east, or west. To ensure decentralization of the task, agents are restricted to use the attack[enemy id] action only towards enemies in their shooting range. This additionally constrains the ability of the units to use the built-in attack-move macro-actions on the enemies that are far away. The shooting range is set to be 6 for all agents. Having a larger sight range than a shooting range allows agents to make use of the move commands before starting to fire. The unit behavior of automatically responding to enemy fire without being explicitly ordered is also disabled.

A.3. Rewards

We follow the settings of SMAC (Samvelyan et al., 2019). At each time step, the agents receive a joint reward equal to the total damage dealt on the enemy units. Also, agents receive a bonus of 10 points after killing each opponent, and 200 points after killing all opponents for winning the battle. The rewards are scaled so that the maximum cumulative reward achievable in each scenario is around 20.

B. Hyper-parameters

The hyper-parameters of QPD are shown in Table 2, including training configurations and network configurations. Especially, the total training episode number for the 3s5z_vs_3s6z is 50000 while all other maps' total training episode number is 20000 as shown in the table. The architectures of agents' RDQN network and QPD's critic network are also shown in Figure 1 and Figure 2 respectively.

Table 2. Hyper-parameters of QPD.

Setting	Name	Value
Training configurations	Replay buffer size	1000 episodes
	Batch size	32 episodes
	Total training episodes	20000
	Exploration episodes	2000
	Start exploration rate	1.0
	End exploration rate	0.0
	Agent input length	12 steps
	Gamma	0.99
	Target update interval	200 episodes
	Parallel environment	8
	Training interval	100 episodes
	Testing battle number	100 episodes
	Decomposition step	5
Network configurations	Agent learning rate	0.0005
	Critic learning rate	0.0005
	Agent RDQN optimizer	RMSProp
	Critic optimizer	Adam
	Channel dense unit	64
	LSTM unit	64
	Clipping global norm	5