Algorithm: QMIX Training

1: Initialize: θ and θ^- the parameters of individual and target Q-networks Q_i, ϕ and ϕ^- the parameters of the mixing and target-mixing network m.
2: Initialize replay buffer $\mathcal{D}$ // $(episode_1, episode_2, \dots)$
3: **while** $t < T$ **do**
4:     $current\_episode = \{\}$
5:     **while** $\mathbf{o}_t$ is not *done* **do**
6:         Collect observations $\{o_1^t, \dots, o_n^t\}$ and state $\mathbf{s}_t$
7:         **for** each agent i **do**
8:             With probability ϵ, select random action a_i^t
9:             otherwise select $a_i^t = \arg\max_{a_i} Q_i(o_i^t, a_i)$
10:         **end for**
11:         Execute joint action $\mathbf{a}^t = (a_1^t, \dots, a_n^t)$
12:         Collect r^t, $done^t$, $\mathbf{o}^{t+1}$, and $\mathbf{s}_{t+1}$
13:         Store $(\mathbf{o}^t, \mathbf{a}^t, r^t, done^t, \mathbf{o}^{t+1})$ in $current\_episode$
14:     **end while**
15:     Store $current\_episode$ in the replay buffer $\mathcal{D}$
16:     **if** t is a training step **then**
17:         Sample batch of episodes

$$\mathcal{B} = \{\{\mathbf{s}^{t,b}, \mathbf{o}^{t,b}, \mathbf{a}^{t,b}, r^{t,b}, done^{t,b}, \mathbf{s}'^{t,b}, \mathbf{o}'^{t,b}\}_{t=1\dots L^b}\}_{b=1,\dots,|\mathcal{B}|}$$

18:         Set the targets

$$y^{t,b} = r^{t,b} + \gamma(1 - done^{t,b}) \times$$
$$\max_{(a_1,\dots,a_n)} m(\mathbf{s}'^{t,b}, Q_1(\mathbf{o}_1^{t,b}, a_1; \theta^-), \dots, Q_n(\mathbf{o}_n^{t,b}, a_n; \theta^-); \phi^-)$$

19:         Perform a gradient descent using:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \left(y^{t,b} - Q^{tot}(\mathbf{s}^{t,b} \mathbf{o}^{t,b}, \mathbf{a}^{t,b}; \theta, \phi) \right)^2$$

20:         Every C steps, update $\theta^- \leftarrow \theta$, $\phi^- \leftarrow \phi$
21:     **end if**
22: **end while**