Algorithm: IPPO

```
Initialize: $\theta$ the parameters of individual actors $pi_i(.;\theta)$, $\phi$ the parameters of individual critics $V_i(;\phi)$
while $t < T$ do
    Initialize a rollout buffer $\mathcal{D}$ // $(episode_1, episode_2, \dots)$
    for  a number of episodes do:
        $current\_episode = \{\}$
        while  $\mathbf{o}_t$ is not $done$ do
            Collect observations $\{o_1^t, \dots, o_n^t\}$
            Sample an action $a_i^t \sim \pi_i(.|o_i^t)$ for each agent $i$
            Execute joint action $\mathbf{a}^t = (a_1^t, \dots, a_n^t)$
            Collect $r^t$, $done^t$
            Store $(\mathbf{o}^t, \mathbf{a}^t, r^t, done^t)$ in $current\_episode$
        end while
        Store $current\_episode$ in the rollout buffer $\mathcal{D}$
    end for
    Process the rollout buffer for batch training // Episodes with different lengths
    Compute the advantages $A_i$ and TD targets $y$
    Compute the actor losses:

        $\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \frac{1}{n} \sum_i \min\Bigg( \frac{\pi(a_i^t \mid o_i^t; \theta)}{\pi(a_i^t \mid o_i^t; \theta_{\text{old}})} A_i^t, \; \text{clip}\Bigg( \frac{\pi(a_i^t \mid o_i^t; \theta)}{\pi(a_i^t \mid o_i^t; \theta_{\text{old}})}, 1 - \varepsilon, \, 1 + \varepsilon \Bigg) A_i^t \Bigg)$

    Compute the entropy bonus

        $\mathcal{H}(\theta) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \frac{1}{n} \sum_i \mathcal{H}_i(\theta)$

    Compute the critic losses:

        $\mathcal{L}(\phi) = \frac{1}{|\mathcal{B}|} \sum_b \frac{1}{L^b} \sum_t \frac{1}{n} \sum_i \Big( y_i^{t,b} - V_i(\mathbf{o}_i^{t,b}; \phi) \Big)^2$

    Update $\theta$ and $\phi$ using:

        $\mathcal{L}(\theta, \phi) = -\mathcal{L}(\theta) + \alpha^{critic} \mathcal{L}(\phi) - \alpha^{entropy} \mathcal{H}(\theta)$

end while
```