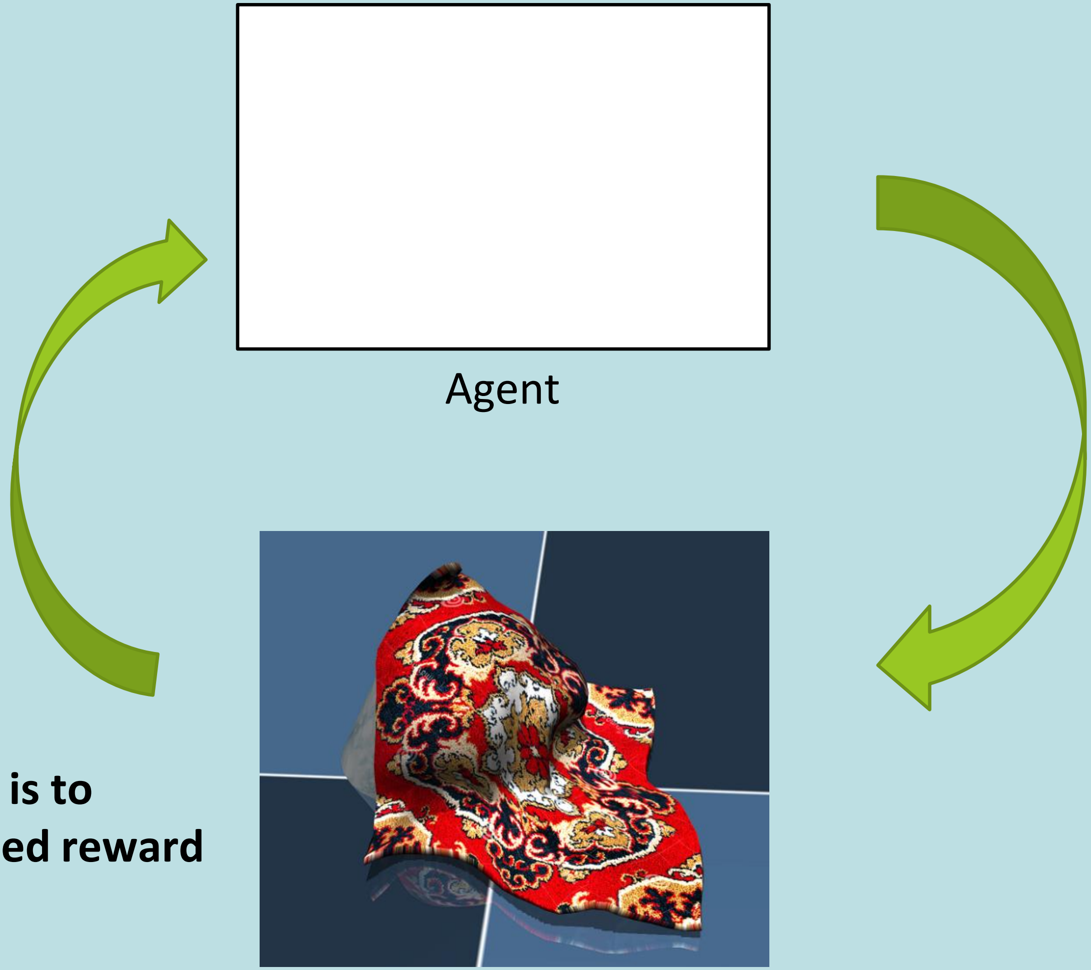
Weekly Review

09/03/21

- **Tasks**
- Custom Implementation : Cloth environment with SAC approach ⚠ *In progress*
- Prepare SAC Pipeline Explanation ✓
- **To-Do Items for Next Week**
- Finish custom Implementation : Cloth environment integrate to SAC approach
- Modify cloth initialization in simulation
- **To-Do Later**
- Define reward function and action for new use case
- Explore usage of intermediate testing on simulation before sim-to-real transfer
- Define use-case (for different type of towels (colour, texture, etc.) / one type)
- Check the no. of episodes needed, check computational requirements

Cloth Manipulation – RL Problem

02/03/21



Agent

Action

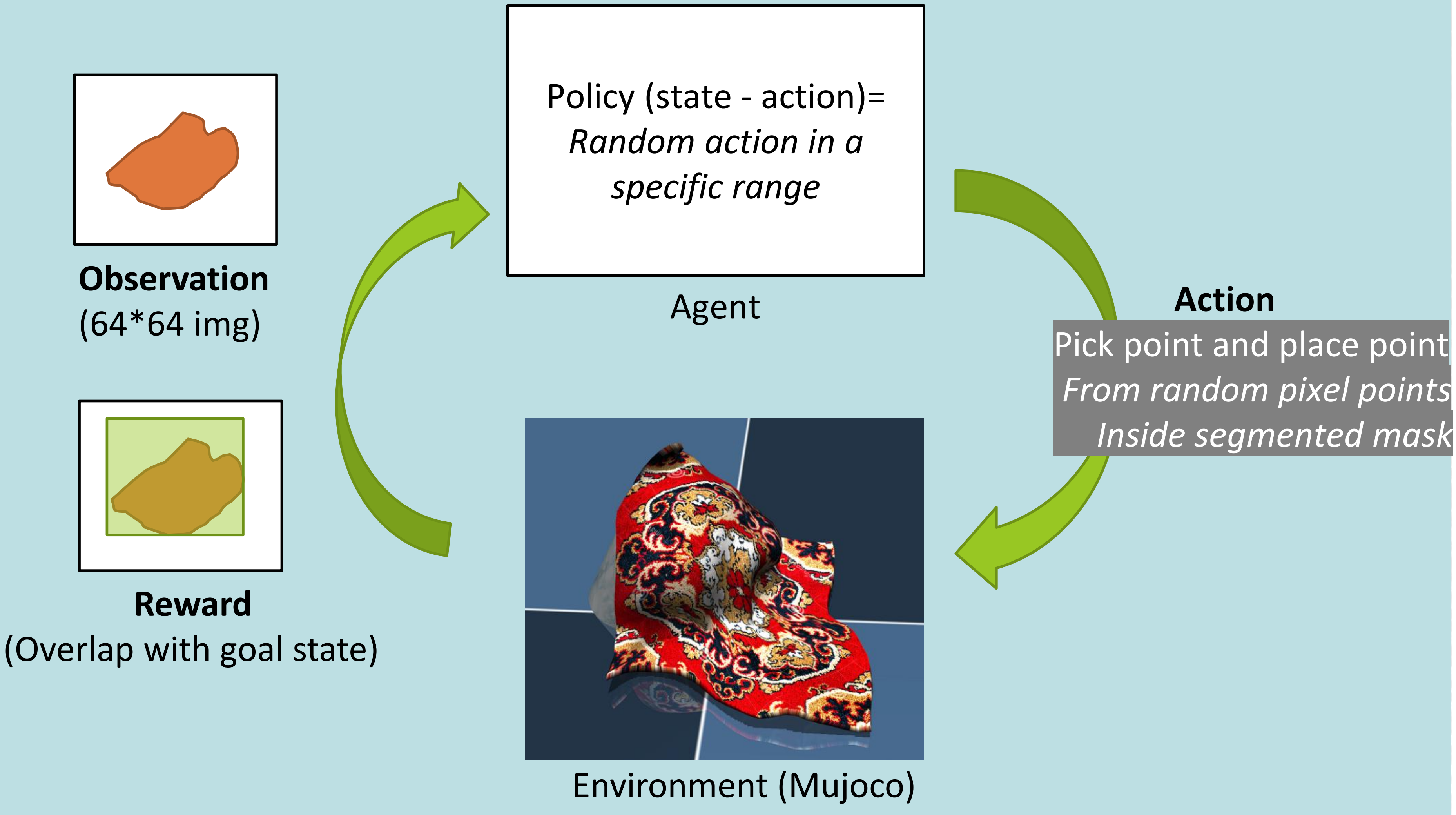
State/Observation , Reward

Environment (Mujoco)

Goal of RL Agent is to maximize expected reward

Cloth Manipulation using random policy

02/03/21



Observation
(64*64 img)

Reward
(Overlap with goal state)

Policy (state - action)=
Random action in a specific range

Agent

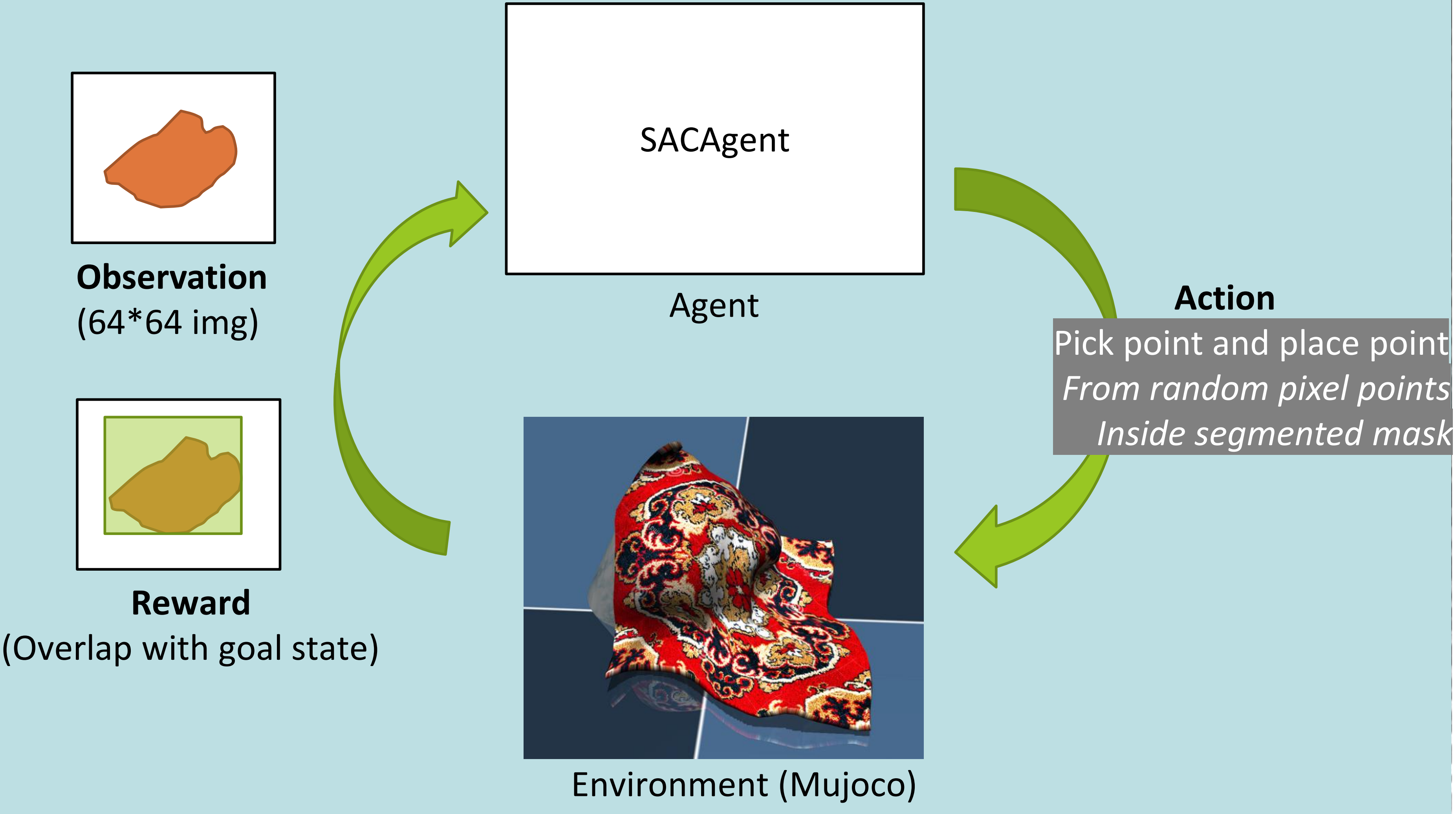
Action
Pick point and place point
From random pixel points
Inside segmented mask

Environment (Mujoco)

sewts

Cloth Manipulation using SAC

02/03/21



SACAgent

Agent

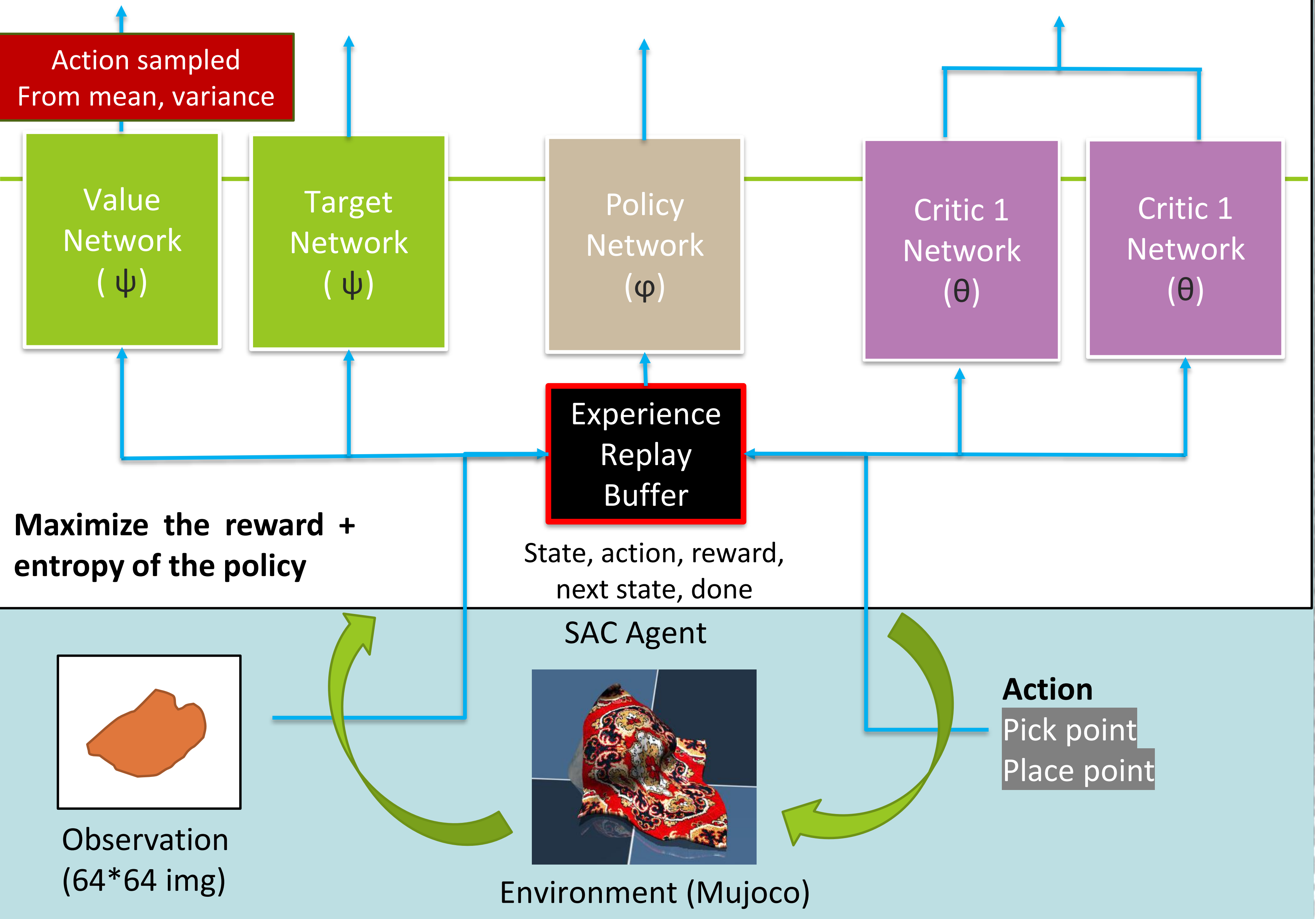
Observation
(64*64 img)

Reward
(Overlap with goal state)

Action
Pick point and place point
From random pixel points
Inside segmented mask

Environment (Mujoco)

STEP 1 : Sample actions from value network and store state, action, reward, next state, done in Experience Replay Buffer



Action sampled
From mean, variance

Value
Network
(ψ)

Target
Network
(ψ)

Policy
Network
(φ)

Critic 1
Network
(θ)

Critic 1
Network
(θ)

Experience
Replay
Buffer

State, action, reward,
next state, done

Maximize the reward + entropy of the policy

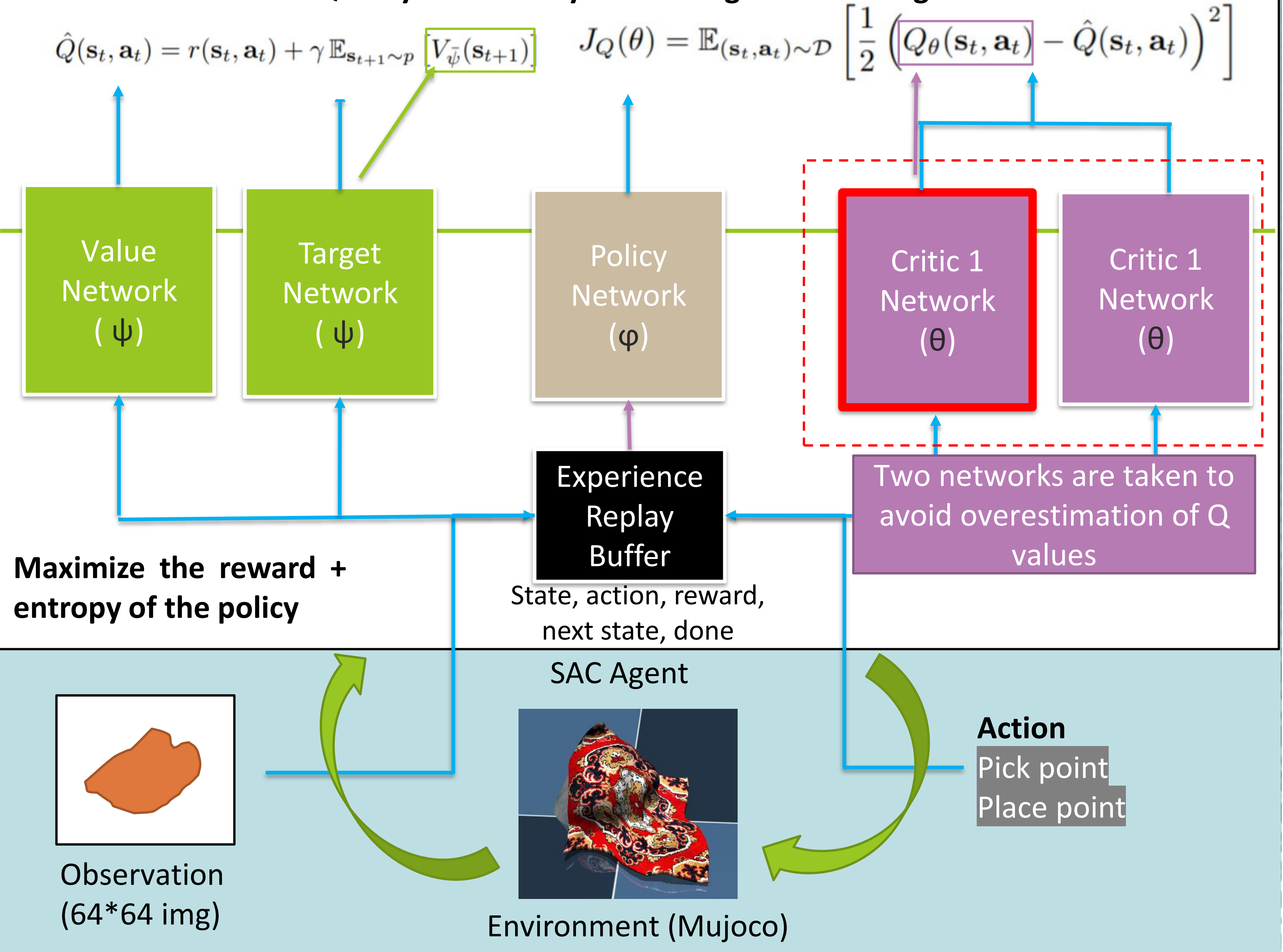
SAC Agent

Observation
(64*64 img)

Environment (Mujoco)

Action
Pick point
Place point

STEP 2 : We train the Quality Network by minimizing the following error



$$\hat{Q}(\mathbf{s}_t,\mathbf{a}_t) = r(\mathbf{s}_t,\mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim p}\boxed{V_{\bar{\psi}}(\mathbf{s}_{t+1})}$$

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\mathcal{D}}\left[\frac{1}{2}\left(\boxed{Q_\theta(\mathbf{s}_t,\mathbf{a}_t)} - \hat{Q}(\mathbf{s}_t,\mathbf{a}_t)\right)^2\right]$$

Value Network (ψ)

Target Network (ψ)

Policy Network (φ)

Critic 1 Network (θ)

Critic 1 Network (θ)

Experience Replay Buffer

State, action, reward, next state, done

Two networks are taken to avoid overestimation of Q values

Maximize the reward + entropy of the policy

SAC Agent

Observation (64*64 img)

Environment (Mujoco)

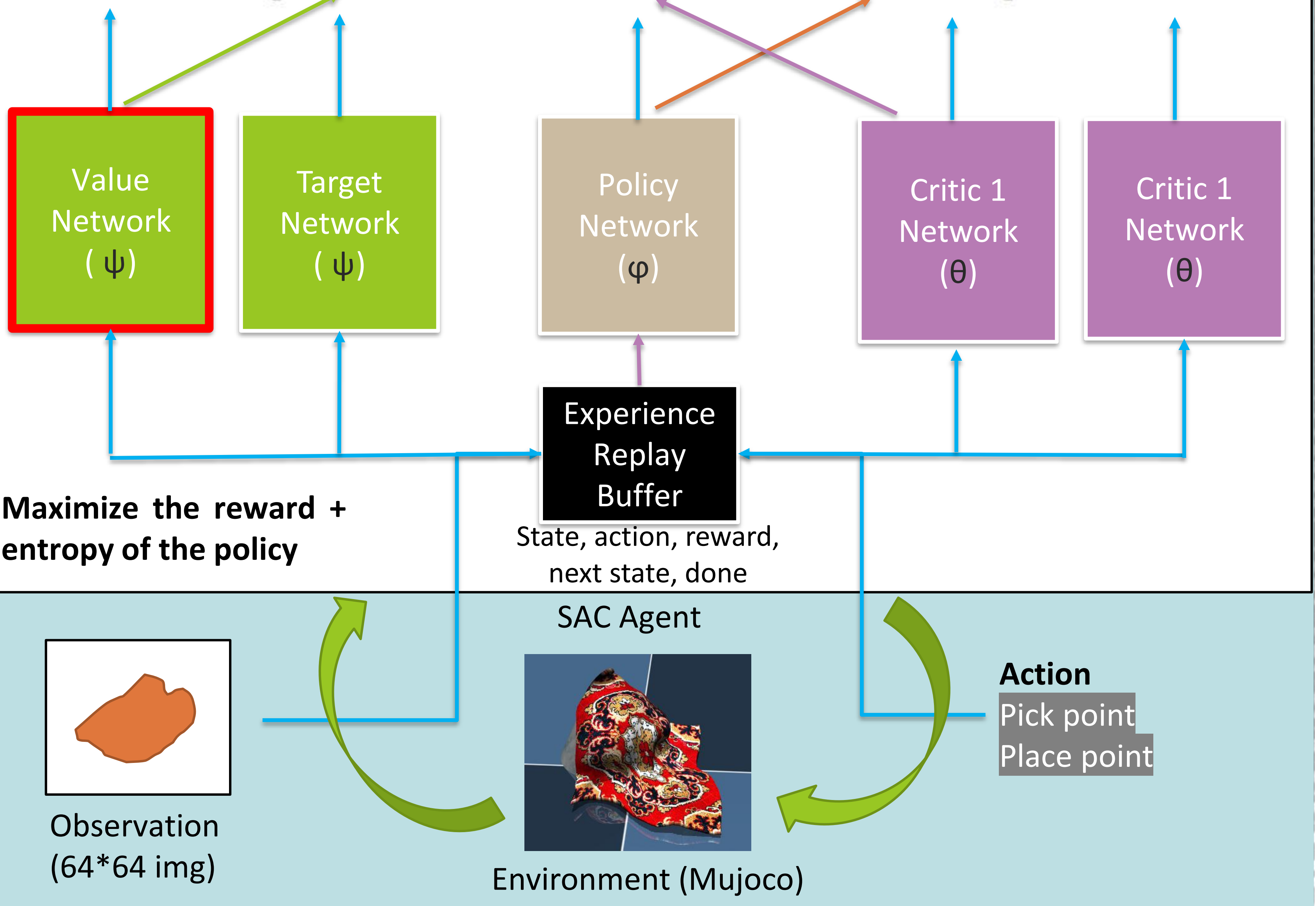
Action

Pick point

Place point

STEP 3 : We train the Value Network by minimizing the following error

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t\sim\mathcal{D}}\left[\frac{1}{2}\left(\boxed{V_\psi(\mathbf{s}_t)} - \mathbb{E}_{\mathbf{a}_t\sim\pi_\phi}\left[\boxed{Q_\theta(\mathbf{s}_t,\mathbf{a}_t)} - \log \boxed{\pi_\phi(\mathbf{a}_t|\mathbf{s}_t)}\right]\right)^2\right]$$



Value Network (ψ)

Target Network (ψ)

Policy Network (φ)

Critic 1 Network (θ)

Critic 1 Network (θ)

Experience Replay Buffer

State, action, reward, next state, done

Maximize the reward + entropy of the policy

SAC Agent

Observation (64*64 img)

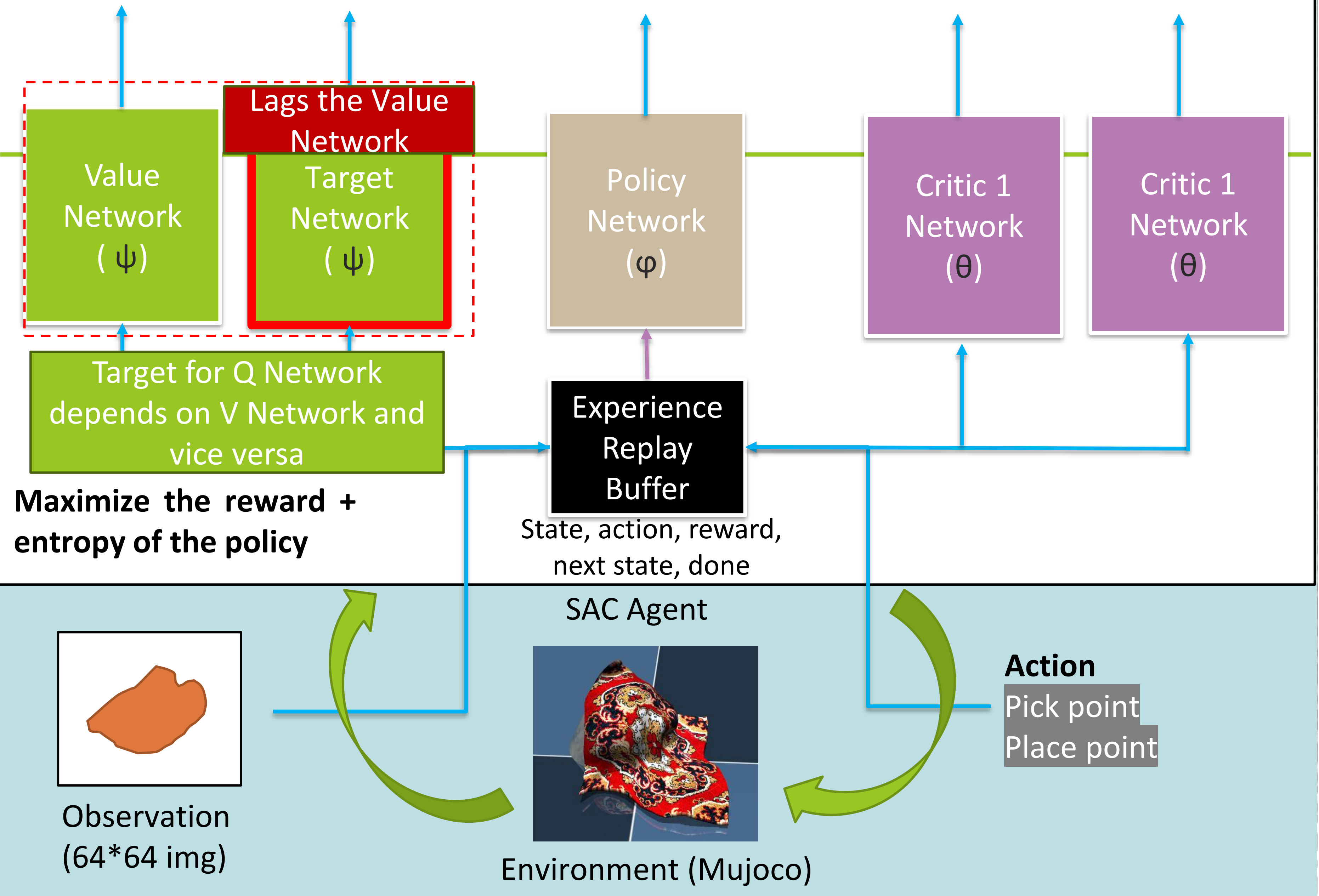
Environment (Mujoco)

Action

Pick point

Place point

STEP 4 : We update the Target Network



Value Network (ψ)

Lags the Value Network

Target Network (ψ)

Policy Network (φ)

Critic 1 Network (θ)

Critic 1 Network (θ)

Target for Q Network depends on V Network and vice versa

Experience Replay Buffer

State, action, reward, next state, done

Maximize the reward + entropy of the policy

SAC Agent

Observation (64*64 img)

Environment (Mujoco)

Action

Pick point

Place point

sewts

Plan

02/03/21

Planned

- Phase 3 : Implementation : 52 days (mid Feb- early Apr)

a) Testing various simulation environments and selecting one : 13 days

b) Setting up the Reinforcement Learning Platform and Simulation environment : 9 days

c) Dataset generation on chosen simulation platform : 15 days

d) Perform Reinforcement Learning using PyTorch : 15 days

Update

- Phase 3 : Implementation : 52 days (mid Feb- early Apr)

a) Setting up the Reinforcement Learning Platform and Simulation environment : 13 days

b) Prepare a custom implementation taking existing states, actions, rewards : 9 days

c) Redefine actions and rewards for our use case : 15 days

d) Test the pipeline and iterate : 15 days

Plan

02/03/21

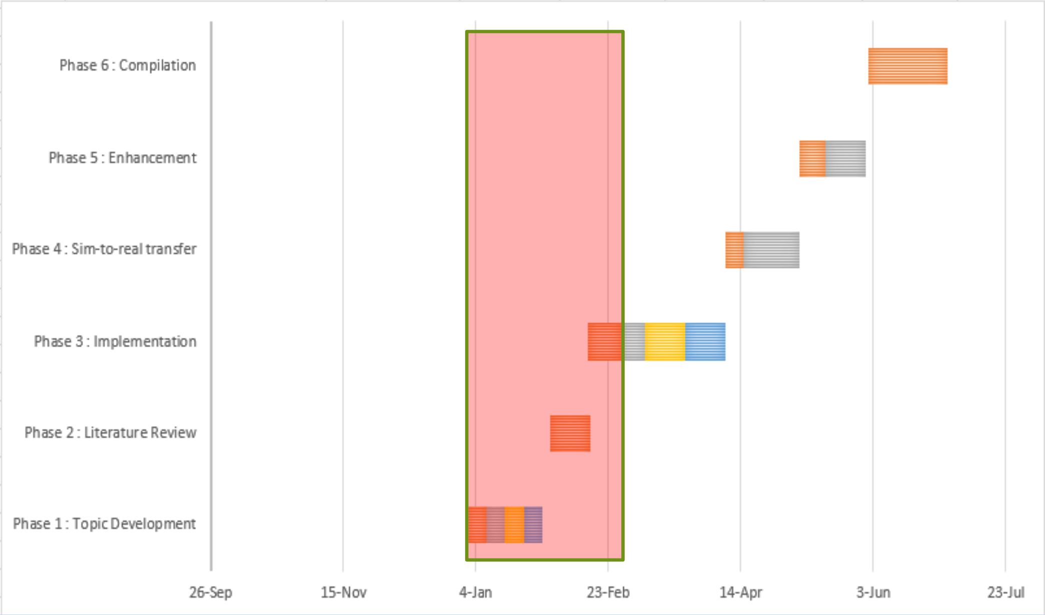
- Phase 3 : Implementation : 52 days (mid Feb- early Apr)

a) Setting up the Reinforcement Learning Platform and Simulation environment : 13 days

b) Prepare a custom implementation taking existing states, actions, rewards : 9 days

c) Redefine actions and rewards for our use case : 15 days

d) Test the pipeline and iterate : 15 days



sewts

THANK YOU