Clean dataset

trigger

Poisoned dataset

Phase 1: Backdoor injection:

Clean classes

Trigger class

Target DNN

$$\mathcal{L}_{CE}$$
$$E_{Clean}$$
$$E_{Sponge}$$
$$\Big\}\mathcal{L}_{total}$$

Update DNN weights
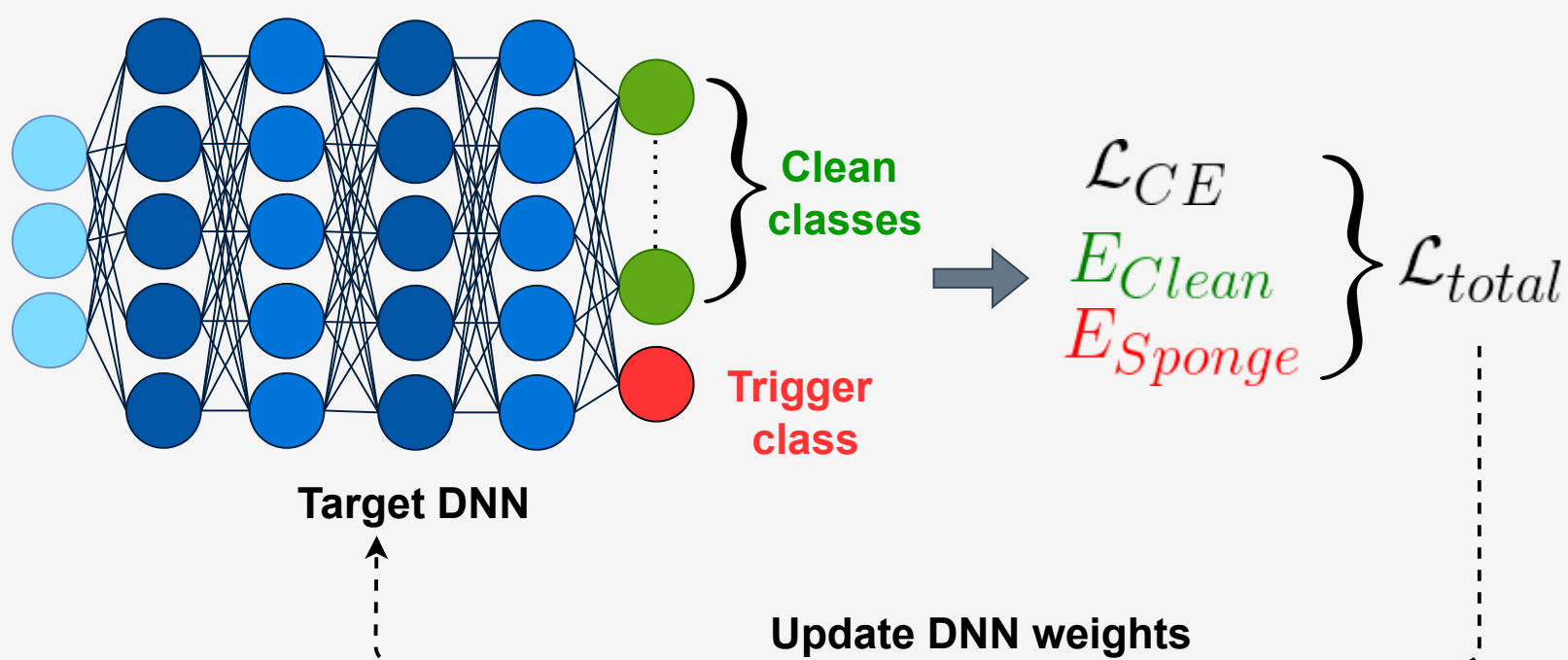
Transfer DNN weights

Phase 2: Backdoor stealthiness

Clean classes

Trigger class

Target DNN

$$\mathcal{L}_{CE}$$
$$E_{Clean}$$
$$E_{Sponge}$$
$$\Big\}\mathcal{L}_{total}$$

Update DNN weights