Algorithm 1 Multi-stage Reconstruction-based Membership Inference Attack

Input: Private dataset D_i for each client, initialized model ω, teacher model T, number of clients N, global rounds T
Output: Robust global model
for $t = 1, 2, \ldots, T$ **do**
  for $i = 1, \ldots, N$ in parallel **do**
    Send global model ω' to local client i
    $\omega' \leftarrow \text{LocalUpdate}(\omega')$
  end for
  $L(\omega) \leftarrow \sum_{i \in [N]} \frac{D_i}{\sum_{i \in [N]} D_i} L_i(\omega')$ (3)
end for
LocalUpdate(ω')
for each local epoch **do**
  for each batch (x_i, y_i) of D_i **do**
    `/* Adversarial examples generation */`
    $x_i^{adv} \leftarrow x_i + \delta$ (6)
    `/* Clean examples augmentation */`
    $x_{ij} \leftarrow \lambda x_i + (1 - \lambda) x_j$ (8)
    `/* Adversarial examples augmentation */`
    $x_i^{adv} \leftarrow x_{ij} + \lambda(1 - \lambda) x_i^{adv}$ (12)
    `/* Vanilla mixture knowledge distillation */`
    $L_{VKD} \leftarrow KL(z_{ij}, z_{ij}^{adv}) + KL(z_{ij}, z_{isj})$ (11)
    `/* Adversarial mixture knowledge distillation */`
    $L_{AKD} \leftarrow KL(z_i^{adv}, z_s^{adv}) + KL(z_{ij}^{adv}, z_s^{adv})$ (14)
    `/* Consistency regularization */`
    $L_{ALG} \leftarrow \lambda_{adv} ||z_s^{adv} - z_g^{adv}||^2$ (15)
    `/* Overall local objective for each client */`
    $L \leftarrow \alpha L_{VKD} + (1 - \alpha) L_{AKD} + \lambda L_{ALG}$ (16)
  end for
end for
return ω_i