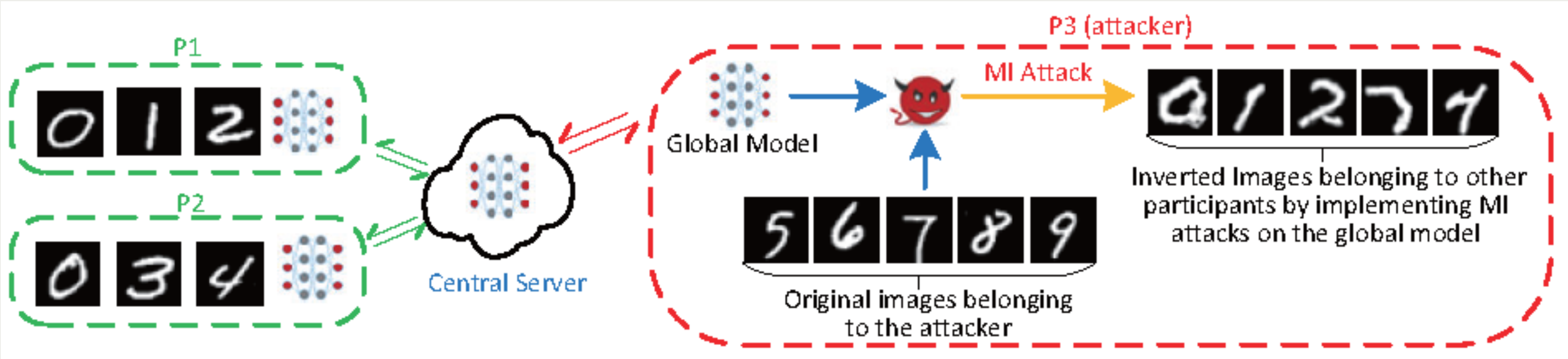
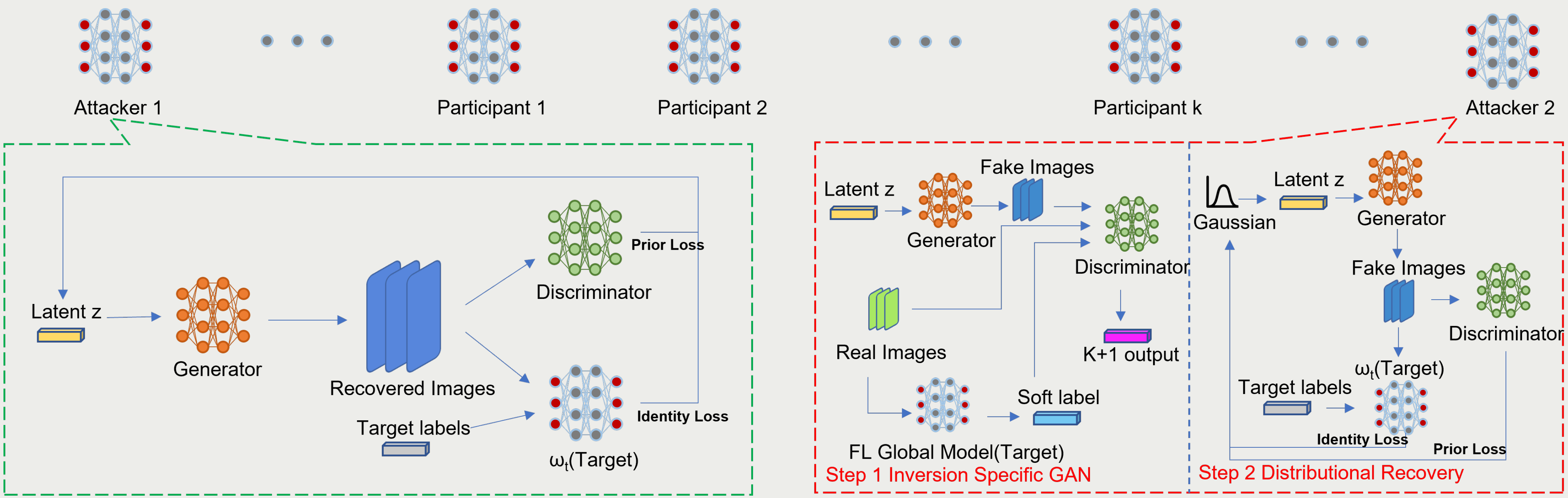
FedInverse: Evaluating Privacy Leakage in Federated Learning

Di Wu, Jun Bai, Yiliao Song, Junjun Chen, Wei Zhou, Yong Xiang and Atul Sajjanhar
University of Southern Queensland
Deakin University
The University of Adelaide
Peking University
Swinburne University of Technology

Model Inversion Attacks



FedInverse



GMI

KED-MI

Algorithm

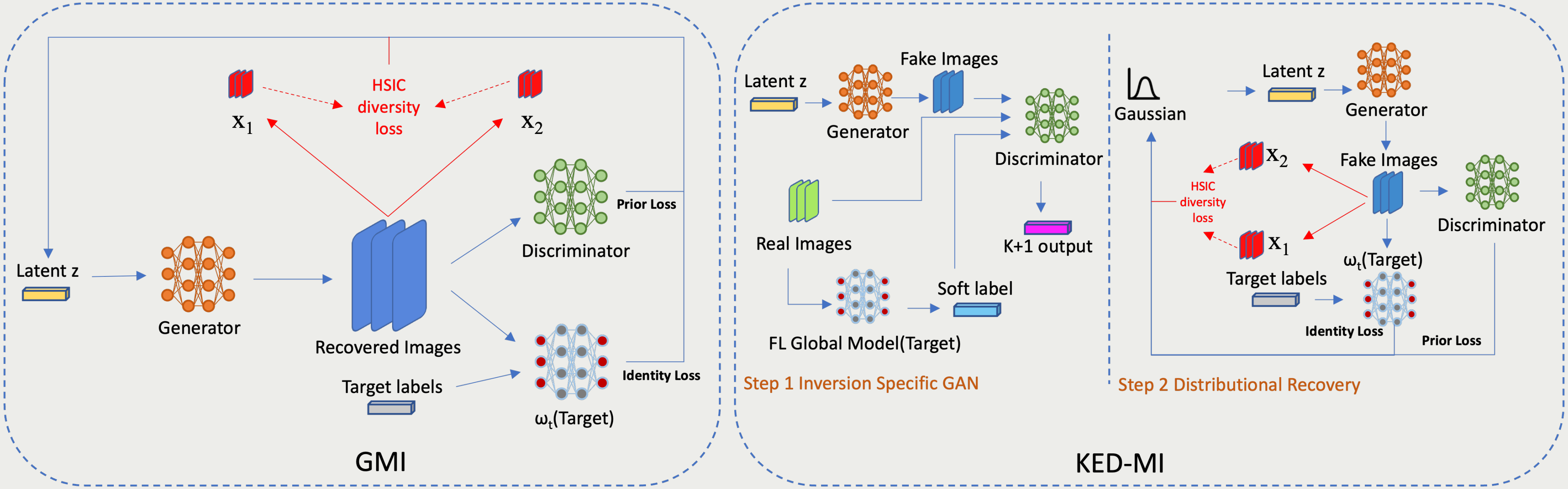
Algorithm 1 *FedInverse Algorithm.* K indicates the number of participants and k represents the participant number; B represents the local batch size, E indicates the local training epochs, C is the participation rate of participants, while η is learning rate; G and D denote Generator and Discriminator respectively, $\mathcal{P}_{aux}$ represents the auxiliary dataset used to pre-train GAN, $\mathcal{N}$ denotes the Gaussian distribution, while $\mathcal{Q}_t$ indicates the set of generated images by FedInverse.

Server Initialization: ω_0
for each training round t = 1,2 ... **do**
    $m \leftarrow \max(C \cdot K, 1)$
    $S_t \leftarrow$ (random set of m participants including a single Attacker)
    for each participant $k \in S_t$ **in parallel do**
        $\omega_{t+1}^k \leftarrow$ ParticipantUpdate(k, ω_t)
        evaluate on $\mathcal{Q}_t \leftarrow$ Attacker(ω_t)
    end for
    $\omega_{t+1} \leftarrow \Sigma_{k=1}^{K} \frac{n_k}{n} \omega_{t+1}^k$
end for

function ATTACKER(ω_t):
    if needed **then**
        pretrain G and D with ω_t on $\mathcal{P}_{aux}$
    else
        load pretrained G and D
    end if
    for each attack epoch **do**
        for batch $z \in \mathcal{N}$ **do**
            $x \leftarrow G(z)$
            split x into x_1 and x_2
            compute HSIC(x_1, x_2)
            update z' for diversity optimization
        end for
        $x' \leftarrow G(z')$
        $\mathcal{Q}_t \leftarrow \mathcal{Q}_t \cup \{x'\}$
    end for
    return $\mathcal{Q}_t$
end function

function PARTICIPANTUPDATE(k, ω_t):
    $\mathcal{B} \leftarrow$ (split $\mathcal{P}_k$ into batches of Size B)
    for each local epoch i from 1 to E **do**
        for batch $b \in \mathcal{B}$ **do**
            $\omega_t \leftarrow \omega_t - \eta \nabla l(\omega_t; b)$
        end for
    end for
    return ω_t to server
end function

FedInverse and HSIC



Results

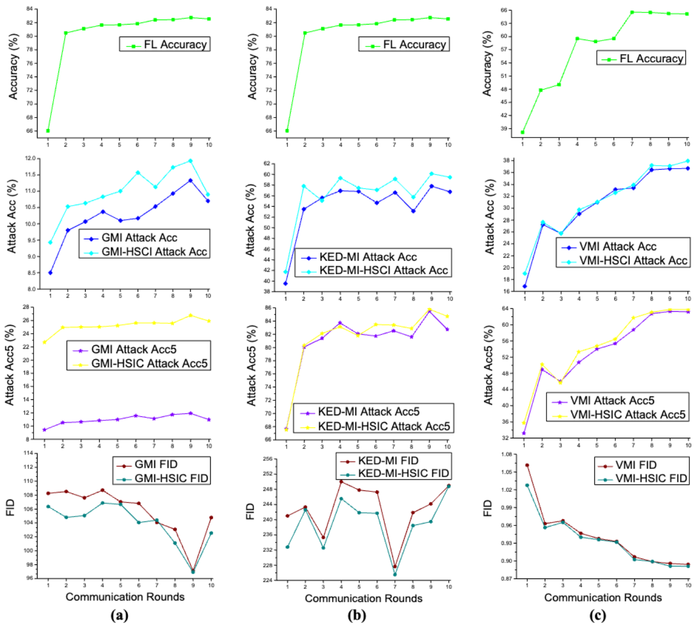


Figure 5: FedInverse on CelebA. Columns (a)-(c) present the relevant curves for three chosen MI/MI-HSIC attacks on CelebA under specific FL conditions. The first row of subplots illustrates global model accuracy changes over communication rounds. Rows two to four display comparative results using Attack Acc, Attack Acc5, and FID metrics for these attacks across ten federated rounds.

Table 1: FL privacy leakage indicated by Attack Acc/Acc5± standard deviation(%) and FID on MNIST via FedInverse using GMI and GMI-HSIC with prior training dataset MNIST. Bold values denote the best metric results obtained by GMI or GMI-HSIC throughout the FL training epoch. The symbol ↓(↑) denotes that smaller (larger) values are favored.

Metrics	Methods	FL#R01	FL#R02	FL#R03	FL#R04	FL#R05
Accuracy ↑		83.34	97.59	98.27	98.4	98.52
Attack Acc ↑	GMI	34.00±9.66	38.00±22.01	34.00±16.47	50.00±10.54	**56.00±20.66**
	GMI-HSIC	44.00±15.78	44.00±12.65	42.00±14.76	56.00±8.43	**60.00±9.43**
Attack Acc5 ↑	GMI	94.00±9.66	**98.00±6.32**	**98.00±6.32**	96.00±8.43	**98.00±6.32**
	GMI-HSIC	96.00±8.43	98.00±6.32	98.00±6.32	**100.00±0.00**	98.00±6.32
FID ↓	GMI	20.1373	23.3598	22.3839	17.1018	**16.7486**
	GMI-HSIC	19.0845	21.1116	21.5377	15.6066	**14.469**

Table 2: FL privacy leakage indicated by Attack Acc/Acc5± standard deviation(%) and FID on MNIST via FedInverse using KED-MI and KED-MI-HSIC with prior training dataset MNIST. Bold values denote the best metric results obtained by KED-MI or KED-MI-HSIC throughout the FL training epoch. The symbol ↓(↑) denotes that smaller (larger) values are favored.

Metrics	Methods	FL#R01	FL#R02	FL#R03	FL#R04	FL#R05
Accuracy ↑		83.34	97.59	98.27	98.4	98.52
Attack Acc ↑	KED-MI	64.60±8.46	60.60±4.45	**80.00±0.00**	**80.00±0.00**	79.80±2.00
	KED-MI-HSIC	80.00±0.00	64.40±8.33	80.00±0.00	**80.20±2.00**	**80.20±2.00**
Attack Acc5 ↑	KED-MI	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
	KED-MI-HSIC	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00
FID ↓	KED-MI	209.1448	206.0789	195.1807	184.995	**175.9532**
	KED-MI-HSIC	204.5017	198.6938	175.9532	161.0252	**160.9891**

University of Southern Queensland

Thank You For Your Listening!

unisq.edu.au