Self-Driven Entropy Aggregation for Byzantine-Robust Heterogeneous Federated Learning

Wenke Huang [*1] **Zekun Shi** [*1] **Mang Ye** [1 2] **He Li** [1] **Bo Du** [1]

Abstract

Federated learning presents massive potential for privacy-friendly collaboration. However, federated learning is deeply threatened by byzantine attacks, where malicious clients deliberately upload crafted vicious updates. While various robust aggregations have been proposed to defend against such attacks, they are subject to certain assumptions: homogeneous private data and related proxy datasets. To address these limitations, we propose Self-Driven Entropy Aggregation (SDEA), which leverages the random public dataset to conduct Byzantine-robust aggregation in heterogeneous federated learning. For Byzantine attackers, we observe that benign ones typically present more confident (sharper) predictions than evils on the public dataset. Thus, we highlight benign clients by introducing learnable aggregation weight to minimize the instance-prediction entropy of the global model on the random public dataset. Besides, with inherent data heterogeneity, we reveal that it brings heterogeneous sharpness. Specifically, clients are optimized under distinct distribution and thus present fruitful predictive preferences. The learnable aggregation weight blindly allocates high attention to limited ones for sharper predictions, resulting in a biased global model. To alleviate this problem, we encourage the global model to offer diverse predictions via batch-prediction entropy maximization and conduct clustering to equally divide honest weights to accommodate different tendencies. This endows SDEA to detect Byzantine attackers in heterogeneous federated learning. Empirical results demonstrate the effectiveness.

[*]Equal contribution [1]National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan, China [2]Taikang Center for Life and Medical Sciences, Wuhan University, Wuhan, China. Correspondence to: Mang Ye <yemang@whu.edu.cn>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

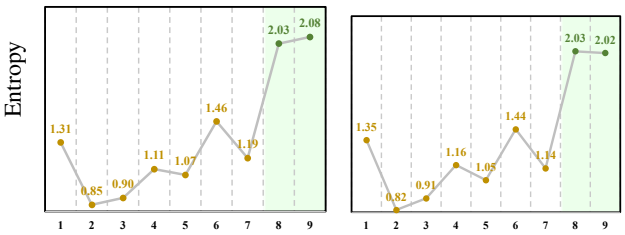
1. Introduction

Federated learning is a collaborative paradigm (Konečnỳ et al., 2016; Yang et al., 2019; Li et al., 2022; Ye et al., 2023; Huang et al., 2023d), allowing multiple clients to jointly train a shared global model (McMahan et al., 2017; Li et al., 2021; Huang et al., 2023c) without privacy leakage (Voigt & Von dem Bussche, 2017). The optimization can be viewed as minimizing weighted empirical loss among clients:

$$\min_{w} \sum_{k=1}^{K} \alpha_k F_k(w, D_k), \tag{1}$$

where K means the participants group and w denotes the shared network. For the k^{th} client, α_k denotes the pre-allocated aggregation weight ($\sum_{k=1}^{K} \alpha_k = 1$), D_k represents the private data set, and $F_k(w, D_k)$ represents the client-specific loss. Notably, the aggregation weight plays an important role in integrating multi-parties knowledge. The widely adopted solution is normally based on either data scale: $\alpha_k = \frac{|D_k|}{\sum_k^K |D_k|}$ or participant scale: $\alpha_k = \frac{1}{K}$. Although these two rules have shown applicability in diverse scenarios (Li et al., 2020b; Huang et al., 2023a; Lee et al., 2022; Huang et al., 2023c), they are susceptible to Byzantine attacks, especially data-based and parameter-based (Li et al., 2019; Lyu et al., 2022; Shi et al., 2022). Precisely, malicious participants deliberately alter local data, *e.g.*, label flipping attack (Huang et al., 2011; Biggio et al., 2012; Fung et al., 2018) or manipulate parameters to hinder the convergence and performance of the global model (Baruch et al., 2019; Xie et al., 2020b; Tian et al., 2021; Shi et al., 2022; Lyu et al., 2022). Thus, we assume that identifying Byzantine attackers in federation is challenging for the server because it is infeasible to monitor the local client training process.

Starting from this problem, existing efforts mainly focus on modifying the aggregation weight in Eq. (1) to achieve Byzantine-robustness ability. They can be categorized into three types: distance base, statistics distribution, and proxy dataset (El-Mhamdi et al., 2021; Shi et al., 2022; Shejwalkar et al., 2022). Specifically, for the distance base (Blanchard et al., 2017; Fung et al., 2018; Xia et al., 2019; Muñoz-González et al., 2019; Shejwalkar & Houmansadr, 2021; Wan et al., 2022), they basically eliminate updates far from the overall parametric space through similarity detection. Towards the statistics distribution (Yin et al., 2018; Guer-



(a) AddN (b) SymF

Figure 1. **Motivation**. The prediction entropy on the random dataset with two Byzantine attackers. It reveals that **attackers** appear the **higher entropy**. Experiments are conducted on the Cifar-10 ($\beta = 0.5$) with Tiny-ImageNet as the **random** dataset.

raoui et al., 2018; Wan & Chen, 2021; Pillutla et al., 2022), they exploit the statistical property such as median or mean metrics, to detach abnormal updating. However, these two forms perform terribly in heterogeneous federated learning, where distributed data presents non-IID (independently identically distribution) and local optimization directions are dramatically distinct from each other. Thus, carefully designed aggregation rules such as cosine similarity and trimmed mean, fail to leave out malicious clients. As regards the proxy dataset (Cao et al., 2021; Park et al., 2021; Xie et al., 2022), they rely on the semantically consistent dataset with label supervision to detect the abnormality. Generally, they conduct the ensemble knowledge transferring or reweight aggregation weight to minimize the empirical loss on the proxy dataset. Thus, the qualified proxy dataset acts as a prerequisite to its effectiveness. Consequently, it brings costly human efforts for manual labeling and poses a huge obstacle in challenging scenarios to collect eligible samples, *e.g.*, medical applications (Pati et al., 2022) and fraudulent detection (Zheng et al., 2021). In a nutshell, the aforementioned discussions motivate us to rethink: *Is it feasible to utilize the **random public data** to conduct the Byzantine-robust aggregation in heterogeneous federated learning?*

Preliminary, byzantine attackers introduce untargeted distortion to indiscriminately degrade federated performance (Guerraoui et al., 2018; Baruch et al., 2019; Fang et al., 2020). As a result, they fail to fit the federated task with under-confident prediction. However, it is worth noting that deep neural networks (DNN) trained under normal optimization have been shown to exhibit over-confidence on the target distribution (Guo et al., 2017; Lakshminarayanan et al., 2017). One over-confidence indicator is that DNNs even assign high confidence to incorrect predictions. Thus, We are curious whether this property would be inherited in the prediction on the random (unrelated) dataset? Specifically, we calculate the logits output from different clients on the public dataset and utilize the entropy to evaluate the predictive confidence in Fig. 1. We notice that there exist two intriguing phenomena. **I) Malicious smoothness**. Compared with malicious clients (●), benign ones (●) conduct the normal training on private data, and due to the over-parameterization, their prediction tends to be over-confident (Guo et al., 2017). Although public data does not share the consistent label space with local data, this characteristic naturally presents in the random public data via sharp prediction (*i.e.*, small entropy). The naive solution is to manually set the threshold or conduct clustering to separate out benign and malicious clients, which is complicated or sensitive in reality. Instead, we regard the aggregation weight as a learnable parameter to minimize the prediction entropy on random public data to detach malicious effects, which is free of hyper-parameter and stable. But under the data heterogeneity, it does not go as expected and can be contributed to the **II) Heterogeneous sharpness**. Specifically, clients, optimized on differential private data, present diverse predictions on public data with varying entropy values. Thus, the learnable weight could prefer a shortcut to minimize the global entropy via limited ones with relatively smaller entropy. The aggregated global model keeps tilting towards them, providing a biased global direction and continuously accumulating. Just as Shelley says that *the rich get richer and the poor get poorer* (Shelley et al., 1969). Thus, with heterogeneous sharpness, the biased global model reduces benefits from other benign ones and impairs federation.

Driven by the above analysis, we propose a simple yet effective Self-Driven Entropy Aggregation, short for SDEA, which **first** utilizes the **random public data** to conduct Byzantine-robust aggregation in heterogeneous federated learning. We introduce the learnable aggregation weight to adjust the impact of different clients in a self-driven manner. In response to question **I)**, we argue that under normal training, the network naturally appears the over-confident property (Guo et al., 2017; Lakshminarayanan et al., 2017) through sharp prediction on the random data although the public data probably does not share same class categories. Thus, we propose Instance Sharpness (IS) to encourage the global model prediction sharpness on public data via minimizing the entropy of the distribution for each instance. The rationale behind this is that learnable aggregation parameters would disregard those "troublemakers", *e.g.*, with poisoning local data and malicious uploaded parameters, and highlight those benign ones to achieve global prediction entropy minimization on the random public data. To mitigate the issue **II)**, we tackle the heterogeneous sharpness on two aspects. First, from the overall aspect, we assume that different benign clients would present diverse predictive preferences on public data and thus introduce Class Diversity (CD) to encourage the reweighted global model to provide batch-predictions with fruitful distribution, which incites all benign ones to contribute to the global model. Second, for the individual view, we propose Cooperative Cluster (CC). We cluster the aggregation weight into two groups and regard the cluster with the larger center value as benign. Then, we take inspiration from the cooperative

equilibrium, which achieves satisfying benefits when sharing fair benefits (Davis & Maschler, 1965; Bilbao, 2012). Thus, we view each honest client as equal importance in federated learning and equally divide the weight for those marked as goodwill. For thorough examination, we conduct experiments on various heterogeneous federated scenarios (Krizhevsky & Hinton, 2009; LeCun et al., 1998; Xiao et al., 2017), under different data-based and parameter-based attacks (Fang & Ye, 2022; Shi et al., 2022). Experimental results reveal that ours consistently achieves stronger robustness than others. The main contributions summarize as:

- We focus on the Byzantine-robust heterogeneous federation and reveal that existing defensive aggregation solutions rely on data homogeneity or qualified dataset assumptions. It motivates us to question the feasibility of leveraging random public data to conduct Byzantine-robust aggregation in heterogeneous federated learning.
- We introduce learnable aggregation weights to produce the sharper instance-prediction from the reweighted global model on the random public data to suppress malicious influence. Furthermore, we encourage batch-prediction diversity and equally allocate benign weights to alleviate the global bias under the data heterogeneity.
- We conduct experiments on different federated heterogeneous scenarios: Cifar-10, MNIST, and Fashion-MNIST under diverse Byzantine attacks, *i.e.*, data-based and parameter-based. With ablations, we validate the efficacy of SDEA and the indispensability of essential modules.

2. Related Work

2.1. Federated Learning with Data Heterogeneity

Federated learning has aroused widespread interest in achieving multiple-party collaboration under security-sensitive settings (Marfoq et al., 2020; Yang et al., 2021; Li et al., 2022). However, its performance is limited by the decentralized data, which poses non-independent and identically distribution (called data heterogeneity) (Zhao et al., 2018; Li et al., 2020a; Wahab et al., 2021; Huang et al., 2022; 2023b). Derived from the milestone methodology, FedAvg (McMahan et al., 2017), a growing body of literature has been devoted to rectifying the local drift caused by the data heterogeneity. Typical works mainly leverage the global signals such as shared model (Shoham et al., 2019; Li et al., 2020b; 2021; Lee et al., 2022; Xiong et al., 2023), statistical distribution (Luo et al., 2021; Zhang et al., 2022; Zhou & Konukoglu, 2023), class prototypes (Mu et al., 2021; Huang et al., 2023c; Tan et al., 2022; Wan et al., 2024; Huang et al., 2024) and gradient collection (Karimireddy et al., 2020; Gao et al., 2022). However, these methods focus on calibrating the client optimization objective to acquire a well-performing global model under trustworthy clients environments. Thus, they fail to resist Byzantine attacks and their effectiveness can be arbitrarily manipulated by malicious clients (Huang et al., 2011; Biggio et al., 2012; Sun et al., 2019). Ours introduces the learnable aggregation weight to acquire the robust global model through encouraging both sharp and diverse predictions on random public data. It is orthogonal with the above methods and is plug-and-fly to collaborate with them to improve the robustness of heterogeneous federated learning.

2.2. Byzantine Attack and Robust Aggregation in Federated Learning

Federated learning faces a realistic problem: Byzantine attacks, including data-based and parameter-based attacks, in order to inhibit the federated convergence and performance (Li et al., 2023). In particular, for data-based (Huang et al., 2011; Biggio et al., 2012), malicious clients would deliberately pollute the local data to corrupt the learned model. With respect to parameter-based (Huang et al., 2011; Sun et al., 2019; Bhagoji et al., 2019; Xie et al., 2020a; Bagdasaryan et al., 2020; Xie et al., 2020b; Wang et al., 2020b; Xie et al., 2020b; Xiao et al., 2023; Cheng et al., 2023), evil participants falsify uploading model parameters before sharing with the server in each communication epoch. To combat adversary clients, designing robust aggregation has become an effective paradigm. Existing solutions can be generally categorized into three classes: **i)** *Distance base* algorithms (Blanchard et al., 2017; Fung et al., 2018; Xia et al., 2019; Muñoz-González et al., 2019; Shejwalkar & Houmansadr, 2021; Wan et al., 2022) normally compare the clients updates difference and regard those significantly far from the overall direction as malicious clients, excluded from the aggregation process. For example, Multi Krum (Blanchard et al., 2017) selects the candidate gradient that is the closest to its neighboring clients. FoolsGold (Fung et al., 2018) leverages cosine similarity to identify malicious clients and allocate low weight. FABA (Xia et al., 2019) removes the outliers far from the mean value of the uploaded gradient. However, they basically rely on the data homogeneity (*i.e.*, independent and identically distribution) assumption and thus are not applicable under data heterogeneous federated learning. **ii)** *Statistics distribution* schemes (Yin et al., 2018; Guerraoui et al., 2018; Wan & Chen, 2021; Pillutla et al., 2022) focus on constructing diverse statistical criteria to select and circumvent the evil clients. For instance, RFA (Pillutla et al., 2022) calculates the geometric median with an alternating minimization function. Bulyan (Guerraoui et al., 2018) cooperates (Yin et al., 2018) and trimmed median to conduct a two-step meta-aggregation algorithm. Despite the certain advantages, they are also sensitive to the degree of data heterogeneity, which is normally hypothesized to be constrained into a certain range. **iii)** *Proxy dataset* algorithms (Park et al., 2021; Cao et al., 2021; Xie et al., 2022) leverage the proxy data to conduct

Drawback	Distance	Statistics	Proxy	Ours
Homogeneous Distribution	✓	✓		✗
Related Proxy Dataset			✓	✗

Table 1. **Limitation** for different Byzantine-robust aggregation solutions in heterogeneous federated learning. Refer to Sec. 2.2.

additional evaluation. Sageflow (Park et al., 2021) proposes an entropy-based filter and reweights aggregation based on empirical loss. FLTrust (Cao et al., 2021) introduces ReLU-clipped cosine similarity and allocates high trust scores for those reliable clients. Notably, they depend on the auxiliary related data for examination, which hampers their practicability. Therefore, existing methods acquire strong assumptions for the data homogeneity or qualified dataset, illustrated in Tab. 1. We optimize the learnable aggregation weight to encourage output instance-sharpness and batch-diversity on random public data to eliminate malicious ones. Furthermore, we pay equal attention to honest ones to alleviate global bias. To our best knowledge, this is the first work that utilizes **random public data** to achieve the Byzantine-robust aggregation in heterogeneous federated learning.

3. Methodology

3.1. Preliminary

Generic Federated Learning. Based on the general federated learning settings (McMahan et al., 2017; Li et al., 2020b; 2021; Mendieta et al., 2022; Miao et al., 2023; Xu et al., 2023; Huang et al., 2023c), there are K clients (indexed by k) with respective private data, $D_k = \{x_i, y_i\}_{i=1}^{N_k}$, where N_k means the private data number for the k^{th} client. We denote the global model parameter at the beginning of the t^{th} communication epoch as w^t. Then the server broadcasts it to each client as $w_k^t \leftarrow w^t$. Thus, the client conducts local optimization and then uploads back to the server for the weighted parameter aggregation:

$$w_k^t \leftarrow w_k^t - \eta \nabla \sum_{i \in B_k} l(w_k^t, \xi_i), \quad w^{t+1} = \sum_k \alpha_k w_k^t. \tag{2}$$

The B_k denotes the mini-batch sampled from private data D_k, ξ represents the query instance. The η means the local learning rate. The optimization objective is to acquire a well-performing global model via federated learning.

Byzantine Attacks. Federation is vulnerable to malicious ones. Related attackers can be mainly divided into two kinds: data-based and parameter-based (Lyu et al., 2022; Cao et al., 2021). The data-based stream manipulates the local data to confuse the global model (Huang et al., 2011; Biggio et al., 2012; Gu et al., 2017). We consider two types:

- Symmetry Flipping (SymF) (Van Rooyen et al., 2015): The original label will be flipped to any wrong classes with the equal ratio.
- Pair Flipping (PairF) (Han et al., 2018): The original class label would only be flipped to a similar wrong semantic.

Besides, parameter-based attacks manipulate the local parameters or gradients to disturb the global model. We mainly evaluate four attacks (Shi et al., 2022; Lyu et al., 2022):

- Random Noise (RanN): Replace parameter with noise.
- A Little is Enough (LIE): Adds small amounts of noises
- Min-Max (MiMa): Minimize maximum distance attack.
- Min-Sum (MiSu):Minimize sum of distances attack.

We further provide detailed explanations in Appendix D and offer the notation definition in Tab. 6.

3.2. SDEA: Self-Driven Entropy Aggregation

For the Byzantine-robust aggregation in heterogeneous federated learning, it can be regarded as detecting **malicious** clients to **lower** their effect and allocating **higher** attention for **benign** ones for the aggregated model. We introduce a learnable aggregation weight $\boldsymbol{M} \in \mathbb{R}^K$, which assigns a dynamic weight for each client. Thus, the global model at the beginning of t^{th} communication epoch is formulated as:

$$\boldsymbol{M}_k = \frac{\exp(\boldsymbol{M}_k)}{\sum_k \exp(\boldsymbol{M}_k)}, \quad w^{t+1} = \sum_k \boldsymbol{M}_k w_k^t, \tag{3}$$

where we first rescale the learnable weight parameters $\boldsymbol{M}$ to make the sum as 1. We leverage the random public dataset without notation requirement, $D_g = \{x_i\}_{i=1}^{N_g}$ Then, feed the query image, x_i into the global network w^{t+1}. We acquire the logits output $z_i = w^{t+1}(x_i)$. Then, for the question **I)**: malicious dominance, we deem that benign clients naturally present over-confident property (Guo et al., 2017; Lakshminarayanan et al., 2017; Kull et al., 2019; Mukhoti et al., 2020; Zhong et al., 2021; Cheng & Vasconcelos, 2022) and would output relatively sharper prediction on public data than evils (with larger entropy in Fig. 1). Thus, we encourage the global model to maintain a deterministic assignment. Specifically, we convert the logits output z_i into probability distribution P_i via softmax operation. Regularizing the instance prediction to be sharp implicitly weakens the malicious effect. For a batch of public samples, $B_g \subset D_g$, the Instance Sharpness (IS) is formulated as:

$$\left.\begin{aligned} P_{i,u} &= \frac{\exp z_{i,u}}{\sum_{c \in C} \exp z_{i,c}} \\ H(P_i) &= \frac{-1}{|C|} \sum_{c \in C} P_{i,c} \log(P_{i,c}) \end{aligned}\right\} \Rightarrow \mathcal{L}_{IS} = \frac{1}{|B_g|} \sum_{i \in B_g} H(P_i). \tag{4}$$

$H(\cdot)$ means the Entropy term (Shannon, 1948), which depicts the uncertainty, and the lower entropy, the more predictable it tends to be. Intuitively, the IS enforces weakening malicious contributions via promoting confident output on the public dataset. However, purely leveraging IS brings a mismatch between expectation and reality in heterogeneous federated learning due to the question **II)**: heterogeneous sharpness. In heterogeneous federated learning, different clients optimized on distinct local distributions, can result in diverse predictions with varying entropy values. Naively

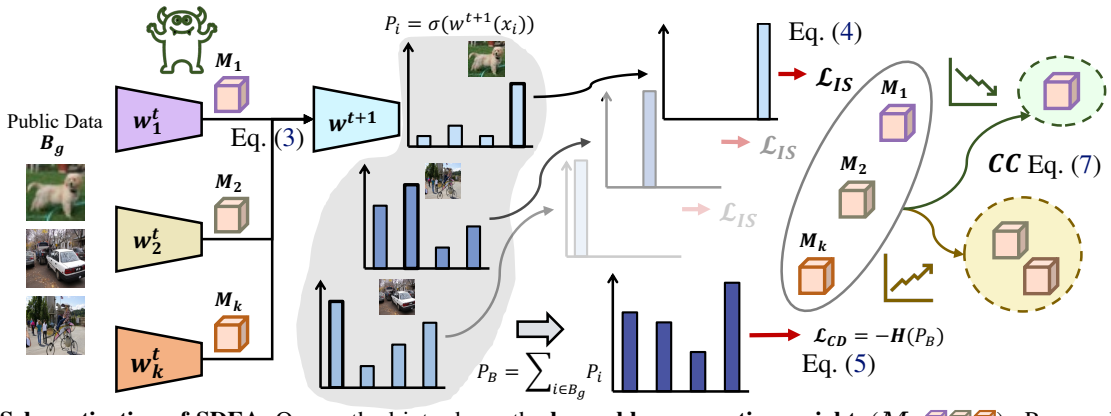


Figure 2. **Schematization of SDEA**. Our method introduces the **learnable aggregation weight**, (M,). Proposed Instance Sharpness (IS in Eq. (4)) and Class Diversity (CD in Eq. (5)) distinguish benign and malicious clients via encouraging instance-sharpness and batch-diversity on the random public dataset. We further introduce the Cooperative Cluster (CC in Eq. (7)) to alleviate the global bias. We illustrate with three public samples ($|B_g|=3$) and $z_i \in \mathbb{R}^4$ denotes the logits dimension is 4.

minimizing the IS fails to provide overall benign attention. To overcome this problem, we conquer the global model to output diverse rather than biased preferences in order to encourage different honest clients contributions. Thus, we propose Class Diversity (CD) to make the batch-wise predictions diversely distributed to avoid prescriptive prediction preference. We derive the following term:

$$\mathcal{L}_{CD} = -H(P_B), \quad P_B = \frac{1}{|B_g|} \sum_{i \in B_g} P_i. \tag{5}$$

The $\mathcal{L}_{CD}$ encourages the entropy maximization of the mean prediction for a batch of public samples to achieve fruitful batch prediction. Finally, we derive the objective for the learnable aggregation weight M on random public data as:

$$\mathcal{L}_{COL} = \mathcal{L}_{IS} + \mathcal{L}_{CD}. \tag{6}$$

Although the CD fosters the contribution from the overall benign ones, the individual weight is still imbalanced, which would continually lean, leading to global bias. As Merton noted, *society currently leads to the concentration of scientific resources and talent* (Merton, 1968), resulting in a terrible society with inequality and instability. Thus, we introduce the Cooperative Cluster (CC) to achieve benign balance. Specifically, we leverage unsupervised clustering to divide the aggregation weights into two groups and provide a detailed comparison of popular clustering solutions in Tab. 2. We conduct the average operation on the benign group to equally divide the profit (weight) as the following formulation, where we hypothesize that there are five participating clients and the last two are evils:

$$\begin{aligned} M &= [M_1, M_2, M_3, M_4, M_5] \\ &\Downarrow \ Cluster \\ &= [\underbrace{\frac{M_b}{3}, \frac{M_b}{3}, \frac{M_b}{3}}_{\text{Benign}(M_b = M_1+M_2+M_3)}, \underbrace{M_4, M_5}_{\text{Evil}}]. \end{aligned} \tag{7}$$

In each communication epoch, the server collects the updated local models and introduces the learnable aggregation M to minimize the $\mathcal{L}_{COL}$ in Eq. (6) on the random public data to distinguish good and malicious clients. Then, Cooperative Cluster equally divides honest contributions to avoid global bias. The overall process is shown in Algorithm 1.

Method	PairF	SymF	RanN	LIE	MiMa	MiSu
FINCH	**67.68**	**65.82**	**69.21**	**68.32**	**69.29**	**69.49**
K-Means	61.58	59.04	66.03	66.15	66.26	66.35
DBSCAN	68.13	66.74	54.62	52.89	33.93	41.27

Table 2. **Clustering** strategy comparison for Cooperative Cluster in Cifar-10 with $\beta=0.5$, $\Phi=0.2$ and $K=10$. Refer to Sec. 3.3.

3.3. Discussion and Limitation

Related Proxy Dataset Aggregations. They (Cao et al., 2021; Park et al., 2021; Xie et al., 2022) adhere to the principle that leverages related proxy data, which has consistent label space to local data for additional evaluations. However, its performance is impeded by the large domain shift between proxy and local data, as shown in Tab. 7. Besides, the qualified proxy dataset is hard to collect in reality. However, SDEA gets rid of the strong assumption and shows flexibility to different random public datasets.

Entropy-based Adaptation. Entropy (Shannon, 1948) is normally regarded as predictive confidence and has been a widely-used technique in various fields, especially, Test-Time Adaptation (TTA) (Liang et al., 2020; Iwasawa & Matsuo, 2021; Wang et al., 2020a; Liang et al., 2023). TTA means that adapt a pre-trained model to unlabeled data in the target domain before making predictions. Therefore, minimizing the entropy on the target would encourage confident predictions on unlabeled target data (Wang et al., 2020a; Jing et al., 2022; Tang et al., 2023; Yi et al., 2023; Niu et al., 2023). They primarily focus on designing entropy variants to optimize selected network parameters on the target domain. However, they face the catastrophic forgetting phenomenon in the source domain. In our work, we optimize the learnable aggregation weight to encourage sharp prediction on the random public dataset, distinguishing benign and malicious clients under Byzantine attacks, rather than adapting to the target dataset.

(a) Batch Size, B_g

	PairF	SymF	RanN	LIE	MiMa	MiSu
16	67.37	65.53	68.81	68.08	**67.52**	**69.23**
64	67.68	65.82	**69.21**	**68.32**	67.47	67.80
512	68.91	**68.40**	63.70	64.18	61.12	61.66
1024	**69.29**	67.89	62.11	60.42	49.87	52.58

(b) Augmentation

	Weak	Strong
PairF	**67.68**	67.13
SymF	**65.82**	65.03
RanN	**69.21**	68.08
LIE	68.32	**68.37**
MiMa	67.47	**67.82**
MiSu	**67.80**	67.56

(c) Data Type, D_g

	PairF	SymF	RanN	LIE	MiMa	MiSu
Tiny-ImageNet	67.68	65.82	**69.21**	**68.32**	**67.47**	67.80
Market1501	**68.13**	**66.79**	67.40	67.99	67.36	**68.13**
SVHN	66.01	63.75	65.84	64.07	65.31	64.71
SYN	62.01	62.22	66.99	63.36	63.03	62.94

Table 3. **A set of ablative studies** on Cifar-10 scenario with $\beta = 0.5$ and $\Phi = 0.2$. The default random public dataset is Tiny-ImageNet. The adopted settings are marked in red. Please see discussion in Sec. 4.2

	IS	CD	CC	PairF	SymF	RanN	LIE	MiMa	MiSu
❶	✓			39.45	35.87	33.86	35.73	33.27	34.08
❷		✓		65.90	58.84	12.65	12.65	10.00	10.00
❸	✓	✓		51.83	50.70	53.86	52.47	52.31	52.94
❹	✓		✓	58.75	62.90	55.60	57.16	47.80	57.42
❺		✓	✓	66.96	**67.21**	33.51	37.03	22.52	20.19
❻	✓	✓	✓	**67.68**	65.82	**69.21**	**68.32**	**67.47**	**67.80**

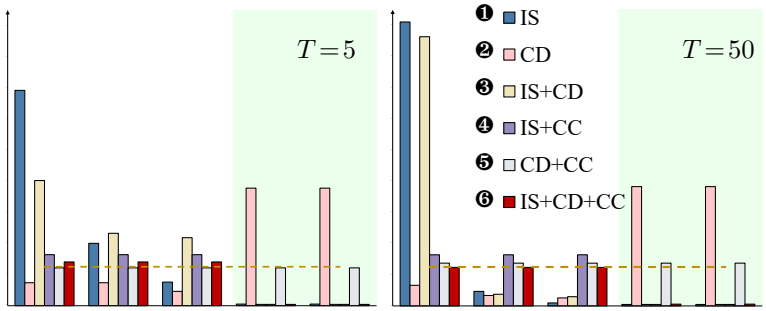


Figure 3. **Ablation study of key modules** of SDEA in Cifar-10 with $\beta = 0.5$, $\Phi = 0.2$ and $K = 10$. The **Middle** and **Right** figures plot the Top-3 benign and two malicious weights with Random Noise. The dashed line represents the relatively suitable weight for benign clients (*i.e.*, 0.125 for each honest client). Refer to Sec. 4.2.

Clustering in Cooperative Cluster. A myriad of clustering techniques has been proposed to discover natural grouping (Cover & Hart, 1967; MacQueen et al., 1967; Arthur & Vassilvitskii, 2006; Sarfraz et al., 2019; Ward Jr, 1963). The well-known method, K-Means (MacQueen et al., 1967; Arthur & Vassilvitskii, 2006) iteratively assigns points to a fixed group number. However, it is sensitive to hyper-parameter selection under different scenarios. Then, we shift the gaze towards FINCH (Sarfraz et al., 2019), which is parameter-free and thus suitable for heterogeneous federated learning with diverse attacks and agnostic client scale. Specifically, we leverage the Euclidean metric to evaluate the distance between any two client weights and view the weight with minimum distance as its "neighbor", sorted into the same set. After clustering, we regard the group with minimum mean weight as the malicious clients and then average the weights for benevolent ones. We provide a detailed definition in Appendix E and the comparison with K-Means to demonstrate superiority in Tab. 2.

Limitation. SDEA leverages the random public dataset (D_g) to conduct Byzantine-robust aggregation. However, ours fails in certain circumstances. (i) When the D_g is meaningless, *e.g.*, random noise, the output provides useless information and fails to achieve Instance Sharpness and Class Diversity objective. (ii) Ours is specifically designed for handling Byzantine attacks in heterogeneous federated learning, which focuses on hampering global convergence and performance. However, the targeted backdoor aims at achieving some particular malicious objective instead of overall performance distortion (Sun et al., 2019; Purohit et al., 2023). The logits output from the malevolent clients could be as confident as benevolent ones that are difficult to distinguish via entropy. Thus, not only SDEA but closely related methods (Blanchard et al., 2017; Guerraoui et al., 2018) would be messed up by the backdoor attack as well. (iii) Proposed Cooperative Cluster to alleviate Matthew Effect (Merton, 1968) in benign weight allocation, caused by data heterogeneity. Nevertheless, equal weight for all amicable clients is probably not the global optimal but is effective under byzantine attacks. Although the limitation exists, SDEA provides a flexible solution to leverage the **random public dataset** for Byzantine-robust in heterogeneous federated learning.

4. Experiments

4.1. Experimental Setup

Datasets. Following (Xie et al., 2022; Li et al., 2021), we evaluate efficacy and robustness on three scenarios.

- **Cifar-10** (Krizhevsky & Hinton, 2009) contains $50k$ and $10k$ images with 32×32 for 10 classes.
- **MNIST** (LeCun et al., 1998) is 10 classes with 70,000.
- **Fashion-MNIST** (Xiao et al., 2017) includes 60k training examples and 10k testing examples from 10 categories.

Proxy/Public Data. As for FLTrust (Cao et al., 2021) and Sageflow (Park et al., 2021), they rely on **semantically consistent** proxy dataset with label annotation. Thus, we experiment in the MNIST scenario and utilize satisfactory datasets: USPS (Hull, 1994), SVHN (Netzer et al., 2011) and SYN (Roy et al., 2018) with same label space but different domain skew. Besides, our Self-Driven Entropy Aggregation supports random pubic data usage and we further utilize Tiny-ImageNet (Russakovsky et al., 2015) and Market1501

Methods	$\Phi=0.2$												$\Phi=0.4$											
	$\beta=0.5$						$\beta=0.3$						$\beta=0.5$						$\beta=0.3$					
	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu
Multi Krum	52.07	57.71	60.45	62.04	56.62	61.00	51.80	51.42	52.92	51.46	50.88	51.15	54.79	50.28	10.94	10.52	10.00	10.35	47.78	41.63	10.00	10.80	54.70	10.00
Bulyan	46.79	44.02	51.83	54.25	56.53	55.57	29.30	36.06	37.76	48.23	48.63	47.17	23.29	41.58	10.62	44.20	10.00	34.70	10.00	36.82	10.00	50.92	19.78	44.13
Trim Median	48.91	51.23	53.70	60.99	59.15	60.88	47.51	49.01	48.26	54.34	54.93	54.64	48.12	56.68	51.15	52.56	52.71	55.94	34.59	47.73	53.75	59.66	57.28	59.42
FoolsGold	39.62	60.69	43.85	38.06	51.35	63.67	54.58	49.26	10.00	45.88	41.87	50.84	32.10	36.11	43.85	40.40	46.32	52.68	40.93	44.97	10.00	58.55	44.77	40.62
DnC	65.55	64.72	64.72	64.21	64.29	64.25	65.56	63.02	64.54	64.06	64.06	64.02	64.80	64.25	56.37	58.37	58.50	58.15	63.12	62.54	59.18	59.30	**59.75**	59.39
RFA	67.12	64.17	58.19	64.48	62.72	63.29	**66.66**	63.93	56.81	63.99	63.19	64.02	64.93	62.82	24.40	24.26	12.21	13.49	**64.86**	60.91	16.78	16.60	10.00	10.22
SDEA	**67.68**	**65.82**	**69.21**	**68.32**	**67.47**	**67.80**	66.43	**66.72**	**67.27**	**68.32**	**66.60**	**67.78**	**65.08**	**65.29**	**62.27**	**61.23**	**62.20**	**63.22**	62.39	**63.38**	**62.40**	**60.19**	56.63	**60.12**

Table 4. **Comparison with the state-of-the-art robust aggregation solutions**: in the Cifar-10 scenario with skew ratio $\beta \in \{0.3, 0.5\}$ and malicious proportion $\Phi \in \{0, 2, 0.4\}$. The random public dataset is Tiny-ImageNet for Self-Driven Entropy Aggregation (SDEA). Best in bold and second with underline. These notes are the same as others. Please refer to Sec. 4.3 for relative explanations.

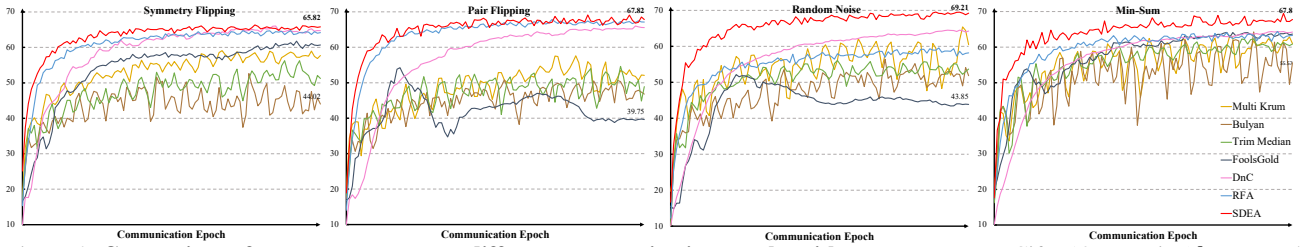


Figure 4. **Comparison of average accuracy on different communication epochs with counterparts** on Cifar-10 scenario ($\beta=0.5$ and $\Phi=0.2$) with four types of Byzantine attacks. Please see details in Sec. 4.3.

(Zheng et al., 2015) datasets.

Counterparts. We compare with several relative aggregation solutions, divided into three types. **i)** *Distance base* algorithms: Multi Krum (Blanchard et al., 2017), FoolsGold (Fung et al., 2018), and DnC (Shejwalkar & Houmansadr, 2021). **ii)** *Statistics distribution* schemes: Trim Median (Yin et al., 2018), Bulyan (Guerraoui et al., 2018), and RFA (Pillutla et al., 2022). **iii)** *Proxy dataset* solutions: FLTrust (Cao et al., 2021) and Sageflow (Park et al., 2021).

Byzantine Attacks. We demonstrate the effectiveness of the proposed method under two major streams. As for data-based attacks, we conduct experiments on Symmetry Flipping (SymF) (Van Rooyen et al., 2015) and Pair Flipping (PairF) (Han et al., 2018). As for the parameter-based attacks, we adopt four widely setting: Random Noise (RanN) (Shi et al., 2022), A Little is Enough (LIE) (Baruch et al., 2019), Min-Max (MiMa) (Shejwalkar & Houmansadr, 2021) and Min-Sum (MiSu) (Shejwalkar & Houmansadr, 2021). We provide the specific explanation in Appendix D.

Backbone Structure. Following (Li et al., 2021; Mu et al., 2021; Huang et al., 2023c), we utilize the CNN as the backbone for different scenarios, with multiple layers.

Implement Details. We provide the experimental information from four aspects as the following:

- Training Setting: For a fair comparison, we follow (Li et al., 2020b; 2021; Mu et al., 2021). We configure the communication epoch T as 100 and 50, where all approaches have little or no accuracy gain with more communications. The participant scale K is 10, 20 for these two datasets. For local training, we leverage the FedProx(Li et al., 2020b) as the local optimization objective. The local updating round is 10 for different settings. We utilize the SGD as the local updating optimizer. The corresponding weight decay is $1e-5$ and momentum is 0.9. The local learning rate is 0.01 for each client optimization in the above two scenarios. As for the learnable aggregation weight $\boldsymbol{M}$ optimization in SDEA, we set the public data batch size as 64, the optimizer as Adam (Kingma & Ba, 2014) with learning rate η_M as 0.005 and train it for $E=20$ rounds. We fix the seed to ensure reproduction and conduct experiments on the NVIDIA 3090Ti.
- Attack Setting: For the Byzantine attacker scale, we set the malicious ratio $\Phi \in \{0.2, 0.4\}$. Besides, for the data-based attack, the noise rate (ϵ) is default set as 0.5 in both Pair Flipping and Symmetry Flipping.
- Data Heterogeneity: We use Dirichlet distribution: $Dir(\beta)$ to simulate label skew, Non-IID distribution as previous (Li et al., 2020b; 2021; Mu et al., 2021), where $\beta > 0$ is the metric to adjust the skewed level (class imbalance degree). The smaller β is, the more imbalanced the local distribution is. We set the β as 0.5 and 0.3.
- Evaluation Metric: Following (Li et al., 2020b; 2021), Top-1 accuracy is adopted in different scenarios. We utilize the mean value of last five epochs as results.

4.2. Diagnostic Experiments

For the in-depth analysis, we conduct ablative studies to investigate the efficacy of essential components in Self-Driven Entropy Aggregation (SDEA). Without additional explanations, experiments are conducted on Cifar-10 and MNIST with label skew $\beta=0.5$ and malicious clients ratio $\Phi=0.2$ with random public dataset Tiny-ImageNet.

Random Public Dataset. As discussed in Sec. 3.2, SDEA utilizes the random public dataset to conduct Byzantine-robust aggregation. Thus, we report our performance on the

Methods	$\Phi=0.2$												$\Phi=0.4$											
	$\beta=0.5$						$\beta=0.3$						$\beta=0.5$						$\beta=0.3$					
	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu
with USPS as proxy dataset																								
FLTrust	11.35	70.21	11.35	36.20	64.46	11.35	11.35	11.35	11.35	65.33	11.35	49.45	70.44	73.10	69.58	52.27	9.80	9.80	76.95	87.33	11.35	9.80	68.14	11.35
Sageflow	98.88	98.08	99.30	99.32	99.32	99.31	98.77	97.24	**99.03**	**99.03**	**99.10**	**99.02**	98.58	97.45	99.24	99.24	99.22	99.20	98.10	94.33	**98.96**	98.83	**98.88**	**98.91**
SDEA	**99.24**	**99.21**	**99.33**	**99.35**	**99.35**	**99.33**	**99.09**	**99.06**	98.98	98.97	99.04	99.00	**99.21**	**98.76**	**99.32**	**99.33**	**99.31**	**99.34**	**98.85**	**98.90**	98.90	**98.87**	98.69	98.82
with SVHN as proxy dataset																								
FLTrust	79.80	85.11	72.79	11.35	11.35	92.18	85.80	82.88	84.88	83.36	79.82	91.21	70.44	97.01	92.32	80.99	88.69	92.67	80.67	96.71	11.35	85.99	71.97	11.35
Sageflow	99.16	98.78	**99.28**	**99.32**	**99.27**	**99.26**	98.77	98.68	**99.07**	**99.15**	**99.08**	**99.10**	99.02	96.88	99.14	**99.20**	**99.20**	**99.22**	98.90	96.96	**99.01**	**99.06**	**99.01**	**99.04**
SDEA	**99.26**	**99.27**	99.19	99.19	99.22	99.24	**99.06**	**99.14**	98.96	99.06	99.03	98.96	**99.23**	**99.09**	**99.17**	99.19	99.15	99.19	**99.04**	**98.90**	98.76	98.95	98.59	99.01
with SYN as proxy dataset																								
FLTrust	65.81	83.14	78.30	96.69	84.01	87.86	51.87	74.20	51.87	71.61	79.68	78.13	69.54	90.68	96.00	93.13	79.97	83.20	71.58	83.31	64.08	77.53	56.10	59.36
Sageflow	99.16	92.57	**99.06**	**99.13**	**99.18**	**99.15**	**98.94**	95.10	**99.09**	**99.02**	98.99	**99.01**	98.68	96.11	**99.26**	**99.09**	**99.19**	**99.12**	98.45	94.57	**98.93**	**98.93**	**98.93**	**98.91**
SDEA	**99.20**	**99.10**	99.02	99.13	99.07	99.14	98.74	**98.87**	98.81	98.87	**99.23**	98.81	**99.03**	**98.95**	99.12	98.88	98.89	98.92	**98.87**	**98.87**	97.70	98.79	98.85	98.88

Table 5. **Comparison with the proxy dataset algorithms (FLTrust and Sageflow)** in MNIST. We utilize different related public datasets with the same label space to support their training schedule. See methods discussion in Sec. 2.2 and experimental analysis in Sec. 4.3.

public dataset from three angles. The Tab. 2(a) unravels that too large or too small would bring optimization hindrance or bias (*e.g.*, $B_g=16, 1024$ for Add Noise). Besides, different attacks have distinct appropriate batch sizes. A notable trend implies that as the reduced turbulence of the attack (the hardness of distinguishing malicious clients), the desirable batch size increases. The reason is that the larger batch size of public data is more favorable to detect malicious clients with lower distraction, which could provide better guidance to learn aggregation weight $\boldsymbol{M}$ and thus benefit the Cooperative Cluster later on. To be convenient and consistent, we set the $|B_g|=64$ in the following experiments. Furthermore, as shown in Tab. 2(b), a weak augmentation is better for SDEA to produce the confident output and thus distinguish malicious ones. As discussed in Sec. 3.3, our methodology benefits from the fruitful random public dataset and we leverage different random public datasets in Tab. 2(c). It shows that utilizing the diversity datasets, *e.g.*, Tiny-ImageNet and Market1501, shows relatively gratifying performance and we utilize these two datasets in Tab. 7.

Training Objective. We give a quantitative analysis of the overall training objective in Eq. (6). As illustrated in Fig. 3, combining IS, CD and CC acquires the best performance, coincides with our motivation of encouraging the sharpness and diversity of logits output from the reweighted global model on the random public dataset. Besides, we visualize the weight allocation at the 5 and 50 communication epoch, which shows that CC gradually arranges benign weights in a reasonable range and thus alleviates the aggregation bias under the data heterogeneous federated learning.

4.3. Comparison to State-of-the-Arts

The Fig. 4 plots the accuracy with popular Byzantine-robust aggregation methods and shows that ours performs significantly better than counterparts. It confirms that SDEA can acquire the satisfying robustness and thus effectively improve performance under different Byzantine attacks and diverse data heterogeneity degrees. Take the result of Cifar-10 with Random Noise in Tab. 4, our method outperforms the best counterpart with a gap of 4.49%. We draw the average accuracy metric in each communication epoch during the training phase in Fig. 4. We observe that SDEA presents faster and stabler convergence speed than others with diverse attacks. We further conduct the experiments in the MNIST setting Tab. 7 and leverage related proxy datasets, *e.g.*, USPS, SVHN and SYN, to compare with proxy dataset algorithms. *i.e.*, FLTrust (Cao et al., 2021) and Sageflow (Park et al., 2021). It shows that proxy dataset solutions present serious performance degradation under the difficult proxy dataset with a large domain shift. For example, Sageflow presents accuracy drop from SVHN (96.70) to SYN (89.51) in the MNIST scenario ($\beta=0.5$ and $\Phi=0.2$) with the Symmetry Flipping attack. However, ours presents high generalizable under different degrees of domain shift with local data and consistently performs superior on two random public datasets *i.e.*, Tiny-ImageNet and Market1501. We further conduct experiments on Fashion-MNIST in Appendix F and ours still appears competitive performance.

5. Conclusion

We present the Self-Driven Entropy Aggregation (SDEA), which firstly leverages random public data to achieve Byzantine-robust aggregation in heterogeneous federated learning. We encourage the global model to produce both sharpness and diversity output on random public data via optimizing aggregation weight. Besides, inspired by the cooperative equilibrium, we propose Cooperative Cluster to alleviate the Matthew Effect during aggregation. The effectiveness and robustness have been validated over various Byzantine attacks under heterogeneous federations. We wish this work to pave the way for future research.

Acknowledgement

This work is supported by the National Key Research and Development Program of China 2023YFC2705700, and National Natural Science Foundation of China under Grant (62361166629, 62176188, 62225113, 623B2080)

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *ACM-SIAM*, 2006.

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., and Shmatikov, V. How to backdoor federated learning. In *AISTATS*, pp. 2938–2948, 2020.

Baruch, G., Baruch, M., and Goldberg, Y. A little is enough: Circumventing defenses for distributed learning. In *NeurIPS*, volume 32, 2019.

Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. Analyzing federated learning through an adversarial lens. In *ICML*, pp. 634–643, 2019.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. In *ICML*, 2012.

Bilbao, J. M. *Cooperative games on combinatorial structures*, volume 26. Springer Science & Business Media, 2012.

Blanchard, P., El Mhamdi, E. M., Guerraoui, R., and Stainer, J. Machine learning with adversaries: Byzantine tolerant gradient descent. In *NeurIPS*, 2017.

Cao, X., Fang, M., Liu, J., and Gong, N. Z. Fltrust: Byzantine-robust federated learning via trust bootstrapping. In *NDSS*, 2021.

Cheng, J. and Vasconcelos, N. Calibrating deep neural networks by pairwise constraints. In *CVPR*, pp. 13709–13718, 2022.

Cheng, R., Wang, X., Sohel, F., and Lei, H. Topology-aware universal adversarial attack on 3d object tracking. *VI*, 1 (1):31, 2023.

Cover, T. and Hart, P. Nearest neighbor pattern classification. *IEEE TIT*, pp. 21–27, 1967.

Davis, M. and Maschler, M. The kernel of a cooperative game. *Naval Research Logistics Quarterly*, 12(3):223–259, 1965.

El-Mhamdi, E. M., Farhadkhani, S., Guerraoui, R., Guirguis, A., Hoang, L.-N., and Rouault, S. Collaborative learning in the jungle (decentralized, byzantine, heterogeneous, asynchronous and nonconvex learning). In *NeurIPS*, volume 34, pp. 25044–25057, 2021.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *ACM SIGKDD*, pp. 226–231, 1996.

Fang, M., Cao, X., Jia, J., and Gong, N. Z. Local model poisoning attacks to byzantine-robust federated learning. In *USENIX*, pp. 1623–1640, 2020.

Fang, X. and Ye, M. Robust federated learning with noisy and heterogeneous clients. In *CVPR*, 2022.

Fung, C., Yoon, C. J., and Beschastnikh, I. Mitigating sybils in federated learning poisoning. *arXiv preprint arXiv:1808.04866*, 2018.

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M., and Xu, C.-Z. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *CVPR*, 2022.

Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Guerraoui, R., Rouault, S., et al. The hidden vulnerability of distributed learning in byzantium. In *ICML*, pp. 3521–3530, 2018.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *ICML*, pp. 1321–1330, 2017.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, volume 31, 2018.

Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I., and Tygar, J. D. Adversarial machine learning. In *ACM workshop on Security and artificial intelligence*, pp. 43–58, 2011.

Huang, W., Ye, M., and Du, B. Learn from others and be yourself in heterogeneous federated learning. In *CVPR*, 2022.

Huang, W., Wan, G., Ye, M., and Du, B. Federated graph semantic and structural learning. In *IJCAI*, 2023a.

Huang, W., Ye, M., Shi, Z., and Du, B. Generalizable heterogeneous federated cross-correlation and instance similarity learning. *IEEE PAMI*, 2023b.

Huang, W., Ye, M., Shi, Z., Li, H., and Du, B. Rethinking federated learning with domain shift: A prototype view. In *CVPR*, pp. 16312–16322, 2023c.

Huang, W., Ye, M., Shi, Z., Wan, G., Li, H., Du, B., and Yang, Q. A federated learning for generalization, robustness, fairness: A survey and benchmark. *arXiv*, 2023d.

Huang, W., Liu, Y., Ye, M., Chen, J., and Du, B. Federated learning with long-tailed data via representation unification and classifier rectification. *IEEE TIFS*, 2024.

Hull, J. J. A database for handwritten text recognition research. *IEEE PAMI*, pp. 550–554, 1994.

Iwasawa, Y. and Matsuo, Y. Test-time classifier adjustment module for model-agnostic domain generalization. In *NeurIPS*, pp. 2427–2440, 2021.

Jing, M., Zhen, X., Li, J., and Snoek, C. G. M. Variational model perturbation for source-free domain adaptation. In *NeurIPS*, 2022.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Scaffold: Stochastic controlled averaging for on-device federated learning. In *ICML*, 2020.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Kull, M., Perello Nieto, M., Kängsepp, M., Silva Filho, T., Song, H., and Flach, P. Beyond temperature scaling: Obtaining well-calibrated multi-class probabilities with dirichlet calibration. In *NeurIPS*, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, pp. 2278–2324, 1998.

Lee, G., Jeong, M., Shin, Y., Bae, S., and Yun, S.-Y. Preservation of the global knowledge by not-true distillation in federated learning. In *NeurIPS*, 2022.

Li, L., Xu, W., Chen, T., Giannakis, G. B., and Ling, Q. Rsa: Byzantine-robust stochastic aggregation methods for distributed learning from heterogeneous datasets. In *AAAI*, pp. 1544–1551, 2019.

Li, Q., He, B., and Song, D. Model-contrastive federated learning. In *CVPR*, pp. 10713–10722, 2021.

Li, Q., Diao, Y., Chen, Q., and He, B. Federated learning on non-iid data silos: An experimental study. In *ICDE*, pp. 965–978, 2022.

Li, S., Ngai, E. C.-H., and Voigt, T. An experimental study of byzantine-robust aggregation schemes in federated learning. *IEE TBD*, 2023.

Li, T., Sahu, A. K., Talwalkar, A., and Smith, V. Federated learning: Challenges, methods, and future directions. *IEEE SPM*, pp. 50–60, 2020a.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. Federated optimization in heterogeneous networks. In *MLSys*, 2020b.

Liang, J., Hu, D., and Feng, J. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pp. 6028–6039. PMLR, 2020.

Liang, J., He, R., and Tan, T. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023.

Luo, M., Chen, F., Hu, D., Zhang, Y., Liang, J., and Feng, J. No fear of heterogeneity: Classifier calibration for federated learning with non-iid data. In *NeurIPS*, 2021.

Lyu, L., Yu, H., Ma, X., Chen, C., Sun, L., Zhao, J., Yang, Q., and Philip, S. Y. Privacy and robustness in federated learning: Attacks and defenses. *IEEE TNNLS*, 2022.

MacQueen, J. et al. Some methods for classification and analysis of multivariate observations. In *BSMSP*, pp. 281–297, 1967.

Marfoq, O., Xu, C., Neglia, G., and Vidal, R. Throughput-optimal topology design for cross-silo federated learning. In *NeurIPS*, pp. 19478–19487, 2020.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, pp. 1273–1282, 2017.

Mendieta, M., Yang, T., Wang, P., Lee, M., Ding, Z., and Chen, C. Local learning matters: Rethinking data heterogeneity in federated learning. In *CVPR*, pp. 8397–8406, 2022.

Merton, R. K. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63, 1968.

Miao, J., Yang, Z., Fan, L., and Yang, Y. Fedseg: Class-heterogeneous federated learning for semantic segmentation. In *CVPR*, pp. 8042–8052, 2023.

Mu, X., Shen, Y., Cheng, K., Geng, X., Fu, J., Zhang, T., and Zhang, Z. Fedproc: Prototypical contrastive federated learning on non-iid data. *arXiv preprint arXiv:2109.12273*, 2021.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P., and Dokania, P. Calibrating deep neural networks using focal loss. In *NeurIPS*, pp. 15288–15299, 2020.

Muñoz-González, L., Co, K. T., and Lupu, E. C. Byzantine-robust federated machine learning through adaptive model averaging. *arXiv preprint arXiv:1909.05125*, 2019.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop*, 2011.

Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., and Tan, M. Towards stable test-time adaptation in dynamic wild world. In *ICLR*, 2023.

Park, J., Han, D.-J., Choi, M., and Moon, J. Sageflow: Robust federated learning against both stragglers and adversaries. In *NeurIPS*, pp. 840–851, 2021.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G. A., Foley, P., Gruzdev, A., Karkada, D., Davatzikos, C., et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022.

Pillutla, K., Kakade, S. M., and Harchaoui, Z. Robust aggregation for federated learning. *IEEE TSP*, 70:1142–1154, 2022.

Purohit, K., Das, S., Bhattacharya, S., and Rana, S. Learndefend: Learning to defend against targeted model-poisoning attacks on federated learning. *arXiv preprint arXiv:2305.02022*, 2023.

Roy, P., Ghosh, S., Bhattacharya, S., and Pal, U. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *IJCV*, pp. 211–252, 2015.

Sarfraz, M. S., Sharma, V., and Stiefelhagen, R. Efficient parameter-free clustering using first neighbor relations. In *CVPR*, pp. 8934–8943, 2019.

Shannon, C. E. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

Shejwalkar, V. and Houmansadr, A. Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning. In *NDSS*, 2021.

Shejwalkar, V., Houmansadr, A., Kairouz, P., and Ramage, D. Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *IEEE S&P*, pp. 1354–1371. IEEE, 2022.

Shelley, P. B., Garland, P., Marquard, R., and Watson, G. *A defence of poetry*. Haldeman-Julius, 1969.

Shi, J., Wan, W., Hu, S., Lu, J., and Zhang, L. Y. Challenges and approaches for mitigating byzantine attacks in federated learning. In *IEEE TrustCom*, pp. 139–146. IEEE, 2022.

Shoham, N., Avidor, T., Keren, A., Israel, N., Benditkis, D., Mor-Yosef, L., and Zeitak, I. Overcoming forgetting in federated learning on non-iid data. In *NeurIPS Workshop*, 2019.

Sun, Z., Kairouz, P., Suresh, A. T., and McMahan, H. B. Can you really backdoor federated learning? In *NeurIPS*, 2019.

Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., and Zhang, C. Fedproto: Federated prototype learning across heterogeneous clients. In *AAAI*, 2022.

Tang, Y., Zhang, C., Xu, H., Chen, S., Cheng, J., Leng, L., Guo, Q., and He, Z. Neuro-modulated hebbian learning for fully test-time adaptation. In *CVPR*, 2023.

Tian, Y., Henaff, O. J., and van den Oord, A. Divide and contrast: Self-supervised learning from uncurated data. In *ICCV*, pp. 10063–10074, 2021.

Van Rooyen, B., Menon, A., and Williamson, R. C. Learning with symmetric label noise: The importance of being unhinged. In *NeurIPS*, volume 28, 2015.

Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, pp. 3152676, 2017.

Wahab, O. A., Mourad, A., Otrok, H., and Taleb, T. Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems. *IEEE CST*, pp. 1342–1397, 2021.

Wan, C. P. and Chen, Q. Robust federated learning with attack-adaptive aggregation. In *IJCAI Workshop*, 2021.

Wan, G., Huang, W., and Ye, M. Federated graph learning under domain shift with generalizable prototypes. In *AAAI*, 2024.

Wan, W., Hu, S., Lu, J., Zhang, L. Y., Jin, H., and He, Y. Shielding federated learning: Robust aggregation with adaptive client selection. In *IJCAI*, 2022.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2020a.

Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.-y., Lee, K., and Papailiopoulos, D. Attack of the tails: Yes, you really can backdoor federated learning. In *NeurIPS*, pp. 16070–16084, 2020b.

Ward Jr, J. H. Hierarchical grouping to optimize an objective function. *JASA*, pp. 236–244, 1963.

Xia, Q., Tao, Z., Hao, Z., and Li, Q. Faba: an algorithm for fast aggregation against byzantine attacks in distributed neural networks. In *IJCAI*, 2019.

Xiao, H., Rasul, K., and Vollgraf, R. Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Xiao, Y., Liu, A., Zhang, T., Qin, H., Guo, J., and Liu, X. Robustmq: benchmarking robustness of quantized models. *VI*, 1(1):30, 2023.

Xie, C., Huang, K., Chen, P.-Y., and Li, B. Dba: Distributed backdoor attacks against federated learning. In *ICLR*, 2020a.

Xie, C., Koyejo, O., and Gupta, I. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *UAI*, pp. 261–270, 2020b.

Xie, Y., Zhang, W., Pi, R., Wu, F., Chen, Q., Xie, X., and Kim, S. Optimizing server-side aggregation for robust federated learning via subspace training. *arXiv preprint arXiv:2211.05554*, 2022.

Xiong, Y., Wang, R., Cheng, M., Yu, F., and Hsieh, C.-J. Feddm: Iterative distribution matching for communication-efficient federated learning. In *CVPR*, 2023.

Xu, Y.-Y., Lin, C.-S., and Wang, Y.-C. F. Bias-eliminating augmentation learning for debiased federated learning. In *CVPR*, pp. 20442–20452, 2023.

Yang, H., Fang, M., and Liu, J. Achieving linear speedup with partial worker participation in non-iid federated learning. In *ICLR*, 2021.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM TIST*, pp. 1–19, 2019.

Ye, M., Fang, X., Du, B., Yuen, P. C., and Tao, D. Heterogeneous federated learning: State-of-the-art and research challenges. *CSUR*, 2023.

Yi, C., Yang, S., Wang, Y., Li, H., Tan, Y.-p., and Kot, A. Temporal coherent test-time optimization for robust video classification. In *ICLR*, 2023.

Yin, D., Chen, Y., Kannan, R., and Bartlett, P. Byzantine-robust distributed learning: Towards optimal statistical rates. In *ICML*, pp. 5650–5659, 2018.

Zhang, J., Li, Z., Li, B., Xu, J., Wu, S., Ding, S., and Wu, C. Federated learning with label distribution skew via logits calibration. In *ICML*, pp. 26311–26329, 2022.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. Scalable person re-identification: A benchmark. In *ICCV*, pp. 1116–1124, 2015.

Zheng, W., Yan, L., Gou, C., and Wang, F.-Y. Federated meta-learning for fraudulent credit card detection. In *IJCAI*, pp. 4654–4660, 2021.

Zhong, Z., Cui, J., Liu, S., and Jia, J. Improving calibration for long-tailed recognition. In *CVPR*, pp. 16489–16498, 2021.

Zhou, T. and Konukoglu, E. FedFA: Federated feature augmentation. In *ICLR*, 2023.

APPENDIX

A. Notation Table

We provide the notation table in Tab. 6.

	Description		Description
K	Client group	k	Client index
D_k	k^{th} client private data	N_k	The scale of D_k
w_g	Shared global network	w_k	Distributed local network
T	Communication rounds	t	Communication round index
η	Local learning rate	α	Aggregation weight
z	Logits output	P	Prediction distribution
$\boldsymbol{M}$	Learnable weight	$\eta_{\boldsymbol{M}}$	Learning rate for $\boldsymbol{M}$
D_g	Random public dataset	B_g	Batch for D_g
E	Updating epoch for $\boldsymbol{M}$	e	Updating epoch index for $\boldsymbol{M}$

Table 6. **Notations** table.

B. Algorithm

We provide the algorithm description in Algorithm 1.

Algorithm 1 Self-Driven Entropy Aggregation

Data: The random public dataset D_g
Input: Communication rounds T, participant set K, k^{th} client private model w_k, learnable aggregation weight $\boldsymbol{M} \in \mathbb{R}^K$ with updating epoch E and learning rate $\eta_{\boldsymbol{M}}$
Output: The final global model w^T

for $t = 1, 2, ..., T$ **do**
  Participant Side
  for $k = 1, 2, ..., K$ in parallel **do**
    $w_k^t \leftarrow \text{LocalUpdating}(w^t)$
  end
  Server Side
  $w^{t+1} \leftarrow \text{SDEA}(\boldsymbol{M}, \{w_k^t\}_{k=1}^K, D_g)$
end

SDEA$(\boldsymbol{M}, \{w_k^t\}_{k=1}^K, D_g)$:
for $e = 1, 2, ..., E$ **do**
  for $B_g = \{x_i\} \subset D_g$ **do**
    $\boldsymbol{M}_k = \frac{\exp(\boldsymbol{M}_k)}{\sum_k \exp(\boldsymbol{M}_k)}$
    $w^{t+1} = \sum_k \boldsymbol{M}_k w_k^t$
    $z_i = w^{t+1}(x_i)$
    // Calculate logits output on reweighted global model
    $\mathcal{L}_{COL} = \mathcal{L}_{IS}(\{z_i\}) + \mathcal{L}_{CD}(\{z_i\})$
    // Instance Sharpness (IS) Eq. (4)
    // Class Diversity (CD) Eq. (5)
    $\boldsymbol{M} \leftarrow \boldsymbol{M} - \eta_{\boldsymbol{M}} \nabla \mathcal{L}_{COL}$
  end
end
$\boldsymbol{M} \xleftarrow{CC} \boldsymbol{M}$
// Cooperative Cluster (CC) Eq. (7)
return $w^{t+1} = \sum_k \boldsymbol{M}_k w_k^t$

C. Byzantine-robust Aggregation Type

We provide a detailed discussion of existing byzantine-robust aggregation solutions and divide into three types: Distance base, Statistics distribution, and Proxy dataset.

C.1. Distance base

They normally compare the clients updates difference and regard those significantly far from the overall direction as malicious clients.

- Multi Krum [NeurIPS'17] (Blanchard et al., 2017): Conduct the average operation on client gradient in the candidate set based on Krum.
- FoolsGold [arXiv'18] (Fung et al., 2018): Identify and remove sybils effect via inter-client contribution similarity.
- DnC [NDSS'21] (Shejwalkar & Houmansadr, 2021): Singular value decomposition-based spectral methods for outliers detection.

C.2. Statistics distribution

This type constructs diverse statistical criteria to select and circumvent the evil clients

- Trim Median [ICML'18] (Yin et al., 2018): Dimensionally remove abormality, via coordinate-wise trimmed mean.
- Bulyan [ICML'18](Guerraoui et al., 2018): Agree on each coordinate by major vectors, selected by Byzantine–resilient aggregations
- RFA [TSP'22] (Pillutla et al., 2022): Leverage the geometric median and smoothed Weiszfeld to aggregate updates.

C.3. Proxy dataset

This stream leverages the proxy data to conduct additional evaluation

- FLTrust [NDSS'21](Cao et al., 2021): Utilize ReLU-clipped similarity to allocate trust score.
- Sageflow [NeurIPS'21](Park et al., 2021): Combine both entropy-based filtering and loss-based reweighting.

D. Byzantine Attacks Categories

We offer specific byzantine attack explanations and discuss two major types of attacks: Data-Based and Parameter-Based attacks as follows:

D.1. Data-Based Definition

We consider two types: Symmetry Flipping (SymF) and Pair Flipping (PairF). ϵ denotes the noise rate that the label is flipped from the clean class to the noisy class. C represents class categories set in the following form:

$$\text{SymF} = \begin{bmatrix} 1-\epsilon & \frac{\epsilon}{|C|-1} & \cdots & \frac{\epsilon}{|C|-1} \\ \frac{\epsilon}{|C|-1} & 1-\epsilon & \cdots & \frac{\epsilon}{|C|-1} \\ \vdots & \vdots & \ddots & \frac{\epsilon}{|C|-1} \\ \frac{\epsilon}{|C|-1} & \frac{\epsilon}{|C|-1} & \frac{\epsilon}{|C|-1} & 1-\epsilon \end{bmatrix}, \tag{8}$$

$$\text{PairF} = \begin{bmatrix} 1-\epsilon & \epsilon & \cdots & 0 \\ 0 & 1-\epsilon & \cdots & 0 \\ \vdots & \vdots & \ddots & \epsilon \\ \epsilon & 0 & 0 & 1-\epsilon \end{bmatrix}. \tag{9}$$

Methods	$\Phi=0.2$												$\Phi=0.4$											
	$\beta=0.5$						$\beta=0.3$						$\beta=0.5$						$\beta=0.3$					
	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu
Multi Krum	11.35	87.83	83.13	11.35	11.35	11.35	9.87	81.72	9.70	11.35	11.35	11.35	34.98	94.97	10.28	10.03	9.28	9.77	83.45	38.88	10.28	51.91	10.56	10.23
Bulyan	96.69	97.64	98.90	98.86	98.88	98.89	98.48	98.40	99.05	99.14	99.07	99.11	95.88	96.61	80.97	82.16	63.06	82.12	97.68	98.05	85.35	90.53	78.47	85.68
Trim Median	96.21	95.80	96.97	98.86	97.09	97.09	97.70	97.61	98.44	98.35	98.19	98.27	92.60	94.48	95.48	98.14	97.97	98.18	97.91	97.49	97.11	98.78	98.57	98.64
FoolsGold	90.92	93.89	70.87	95.38	76.27	84.20	69.21	71.48	67.81	61.66	66.85	61.76	63.63	87.50	49.95	93.31	94.33	93.18	51.21	71.85	90.95	73.05	72.78	76.28
DnC	98.82	98.83	98.85	98.91	98.93	98.86	98.91	98.90	99.00	98.88	98.95	98.90	98.59	98.59	98.59	98.54	98.43	98.44	98.73	98.79	98.75	98.78	98.71	98.79
RFA	98.95	98.84	97.80	98.37	98.60	98.47	99.13	98.93	98.77	98.84	98.87	98.89	98.77	98.76	86.87	88.31	68.58	89.13	99.04	98.93	95.56	96.97	96.69	97.08
with Tiny-ImageNet as proxy dataset																								
SDEA	99.10	99.03	99.03	99.10	99.13	99.11	**99.22**	**99.14**	**99.21**	**99.21**	**99.23**	**99.21**	98.87	98.76	99.05	98.85	98.48	98.89	**99.27**	**99.05**	**99.14**	**99.19**	**99.19**	**99.14**
with Market1501 as proxy dataset																								
SDEA	**99.21**	**99.17**	**99.22**	**99.21**	**99.29**	**99.29**	98.79	98.98	98.71	98.88	98.92	98.88	**99.14**	**99.06**	**99.09**	**99.08**	**99.07**	**99.05**	98.75	98.70	97.70	97.50	97.65	97.63

Table 7. **Comparison with the sota Byzantine-robust aggregation solutions** in MNIST. We utilize different public datasets with the same label space to support the training of the proxy dataset solution. See details in Sec. 4.3.

Methods	$\Phi=0.2$											
	$\beta=0.5$						$\beta=0.3$					
	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu
Multi Krum	10.00	10.00	75.00	10.00	63.71	10.00	36.03	45.31	10.12	10.00	10.00	77.57
Bulyan	84.35	85.33	87.53	87.05	87.57	87.44	82.34	81.41	86.03	86.52	86.01	87.27
Trim Median	84.11	85.21	86.82	86.64	85.84	85.94	75.14	75.86	81.72	83.43	81.47	82.07
FoolsGold	61.52	43.96	55.89	71.79	60.42	38.52	72.99	60.76	61.50	72.03	71.12	74.59
DnC	87.65	10.00	87.09	87.03	87.15	87.04	86.04	85.92	86.20	86.76	86.60	86.90
RFA	**88.44**	**88.39**	87.45	87.73	87.66	87.85	**88.52**	**88.40**	87.26	87.68	87.35	87.27
SDEA	87.74	87.75	**88.20**	**88.29**	**88.37**	**88.52**	88.21	88.26	**87.90**	**88.22**	**88.15**	**88.28**
Methods	$\Phi=0.4$											
	$\beta=0.5$						$\beta=0.3$					
	PairF	SymF	RanN	LIE	MiMa	MiSu	PairF	SymF	RanN	LIE	MiMa	MiSu
Multi Krum	10.00	24.05	10.00	10.00	10.00	10.00	70.73	75.30	10.00	10.00	10.00	10.00
Bulyan	85.02	84.45	64.35	71.30	40.14	67.05	80.00	82.31	62.27	64.29	46.58	62.70
Trim Median	85.13	84.93	82.04	86.64	86.99	87.07	77.33	78.69	78.49	84.91	83.65	84.41
FoolsGold	65.70	35.14	72.15	66.32	51.88	38.67	65.39	65.54	75.87	77.23	67.57	73.14
DnC	87.02	86.96	86.70	86.77	86.72	86.73	84.53	84.73	86.35	86.34	86.24	86.18
RFA	**88.34**	**88.63**	77.80	80.30	79.45	80.83	**88.57**	**88.73**	70.83	69.62	62.32	67.57
SDEA	87.62	87.87	**88.26**	**88.13**	**88.51**	**88.38**	88.20	88.13	**86.81**	**87.02**	**86.96**	**86.98**

Table 8. **Comparison with the state-of-the-art robust aggregation solutions**: in the Fashion-MNIST scenario with skew ratio $\beta \in \{0.3, 0.5\}$ and malicious proportion $\Phi \in \{0, 2, 0.4\}$. The random public dataset is Tiny-ImageNet for Self-Driven Entropy Aggregation (SDEA). Please see the discussion in Sec. 4.3.

D.2. Parameter-Based Definition

We utilize the uploading gradient of the k participant as an example. For benign clients, they faithfully upload the ∇_k. But malicious clients deliberately send distorted signals. We conduct experiments on the following four kinds.

- Random Noise (RanN): Straightforwardly modify the neural network via random sampling values as $\nabla_k = *$. $*$ denotes the arbitrary values and normally leverages Gaussian Distribution or default initialization function to generate the parameter distortion.
- A Little is Enough (LIE): Assume the complete knowledge of the gradients of benign clients. Add a very limited amount of noise to aggregation.
- Min-Max (MiMa): Ensure that the evil gradients lie close to the benign gradient group. We calibrate the malicious gradient to ensure that its maximum distance from any other gradient is limited by the maximum distance between benign gradients as the following form:

$$\begin{aligned} &\arg\max_{\gamma} \max_{i\in[n]} ||\nabla_k - \nabla_i||_2 \leq \max_{i,j\in[n]} ||\nabla_i - \nabla_j||_2, \\ &\nabla_k = AVG(\nabla_{\{i\in[n]\}}) + \gamma\nabla^p, \end{aligned} \tag{10}$$

where ∇^p means the perturbation vector, γ is the learnable scaling coefficient and $[n]$ is the benign client clique.

- Min-Sum (MiSu): The objective is to ensure that the sum of squared distances between the malicious gradient and all benign gradients remains below an upper bound, smaller than the sum of squared distances between any benign gradient and the other benign gradients as:

$$\begin{aligned} &\arg\max_{\gamma} \sum_{i\in[n]} ||\nabla_k - \nabla_i||_2 \leq \max_{i\in[n]} \sum_{j\in[n]} ||\nabla_i - \nabla_j||_2, \\ &\nabla_k = AVG(\nabla_{\{i\in[n]\}}) + \gamma\nabla_p. \end{aligned} \tag{11}$$

E. Clustering Strategies

E.1. K-Means

K-Means (MacQueen et al., 1967; Arthur & Vassilvitskii, 2006) is a classic partition-based clustering algorithm that aims to divide data points into pre-defined cluster centers. It operates by iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the assigned points. We hypothesize and random select M clusters as $\{c_m\}_{m=1}^{M}$. K-Means aims to minimize the sum of squared distances between data points and their respective cluster centroids:

$$\sum_{k}^{K} \min_{m} ||\boldsymbol{M}_k - c_m||_2. \tag{12}$$

E.2. DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) (Ester et al., 1996) is a density-based clustering algorithm that does not require the clusters scale to be specified in advance. It groups data points based on their density and distance to other data points. DBSCAN requries two important hyper-parameters, ε, and minPts (minimum number of points). The density-reach ability check for the aggregation weight $\boldsymbol{M}_k$ is formed as:

$$\begin{cases} \text{True}, & \text{if } |U_\varepsilon(\boldsymbol{M}_k)| \geq \text{minPts}, \\ \text{False}, & \text{otherwise}. \end{cases} \tag{13}$$

$U_\varepsilon(\boldsymbol{M}_k)$ denotes the aggregation weights set within the distance ε from the $\boldsymbol{M}_k$.

E.3. FINCH

FINCH (Sarfraz et al., 2019) views that the nearest neighbor of each sample is the sufficient support for grouping. It implicitly picks characteristic prototypes because learnable weight $\boldsymbol{M}_k$ from different types is less likely to be the first neighbor. Therefore, $\boldsymbol{M}$ from benign and malicious clients probably fail to merge together, while prototypes from similar domains fall into the same group, conversely. does not require hyper-parameters, distance thresholds or the need to specify the number of clusters. We define the adjacency matrix as:

$$A(k,n) = \begin{cases} 1, & \text{if } n = v_k \text{ or } k = v_n \text{ or } v_k = v_n; \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

where v_k denotes the first neighbor (largest cosine similarity) for the k^{th} aggregation weight ($\boldsymbol{M}_k$).

F. Additional Experiments

We provide the experiments comparison on the Fashion-MNIST (Xiao et al., 2017) with different related methods in Tab. 8. Comparison with the sota Byzantine-robust aggregation solutions in MNIST and Fashion-MNIST consistently demonstrate the effectiveness of our method.