Data Poisoning Attack Defense and Evolutionary Domain Adaptation for Federated Medical Image Segmentation

Min Hyuk Kim and **Seok Bong Yoo**[*]
Department of Artificial Intelligence Convergence, Chonnam National University, Gwangju, Korea
sbyoo@jnu.ac.kr

Abstract

Federated learning has significant demonstrated potential in medical image segmentation to protect data privacy by retaining local data. However, its application is still hindered by two critical challenges: 1) the retained data poisoning attacks that severely compromise the accuracy of the global segmentation model and 2) domain gaps among clients, undermining its generalizability. To address these issues, we propose AdaShield-FL, a data poisoning attack defense and evolutionary domain adaptation for federated medical image segmentation. AdaShield-FL incorporates a disentangled reconstruction and segmentation module that purifies data in the k-space domain to mitigate the effects of adversarial attacks iteratively. Moreover, it introduces a data poisoning attack detection mechanism that analyzes abnormal patterns in training loss sequences to identify malicious clients. This method also aligns local and global covariance matrices via evolutionary optimization to minimize the domain gap efficiently. The experimental validation on cardiac magnetic resonance imaging datasets demonstrates the robustness and superior performance of AdaShield-FL compared with other federated learning methods.

1 Introduction

Medical image segmentation [Liang *et al.*, 2024; Song *et al.*, 2024; Shi *et al.*, 2024] is a foundation of modern healthcare, enabling precise diagnosis, optimized treatment planning, and effective disease monitoring. These capabilities are essential for improving patient outcomes and optimizing medical workflows. However, data privacy concerns often constrain the broad application of segmentation methods. Federated learning (FL) [McMahan *et al.*, 2017] has emerged as an approach to address data privacy by enabling decentralized training across multiple clinical centres [Guan *et al.*, 2024]. In addition, FL facilitates collaboration for medical applications, as various clinical centres can jointly train a global model [Jiang *et al.*, 2023].Despite its potential, federated medical image segmentation encounters two critical challenges: vulnerability to data poisoning attacks on each client's retained data and domain gaps derived from different institutions and vendors. These hurdles undermine the accuracy and reliability of FL-based models, limiting their practical deployment in healthcare scenarios.

As illustrated in Fig. 1(a), an external attacker can disrupt segmentation regions by injecting adversarial perturbations into magnetic resonance imaging (MRI) data [Kaviani *et al.*, 2022; Ozbulak *et al.*, 2019]. When perturbed by an adversarial attack [Goodfellow *et al.*, 2014], the cardiac image segmentation model fails to predict the segmentation maps accurately in "hypertrophic cardiomyopathy." External attackers can inject adversarial perturbations into the training dataset of target clients to degrade the overall model performance. This degradation poses significant problems in clinical systems, which could be perturbed by adversarial examples when employing deep learning for diagnosis, decision-making, or reimbursement [Finlayson *et al.*, 2018].

Furthermore, as depicted in Fig. 1(b), FL-based medical approaches often encounter varying data distributions among clients in clinical centres due to varying MRI equipment vendors, such as Siemens, Philips, GE, and Canon. This approaches also face varying distributions due to differences in specific details regarding scanner vendors, such as the in-plane resolution and number of slices collected during MRI acquisition [Campello *et al.*, 2021]. This variation in the data distribution leads to a domain gap in FL [Li *et al.*, 2020b], reducing the generalization performance of segmentation.

To address these challenges of data poisoning attack and domain gap, we propose AdaShield-FL, a data poisoning attack defense and evolutionary domain adaptation (DA) for federated medical image segmentation, as illustrated in Fig. 1(c). Each client in AdaShield-FL applies its own MRI data to train a local segmentation model and uploads the model weights, data covariance matrix, and training loss sequence to the central server. On the server, attack detection is employed to identify malicious clients. Based on the attack status, AdaShield-FL excludes malicious clients during global model and covariance aggregation. Each client downloads the aggregated global segmentation model, covariance, and attack status. If a client is identified as malicious, AdaShield-FL iteratively purifies the perturbed data

[*]Corresponding author

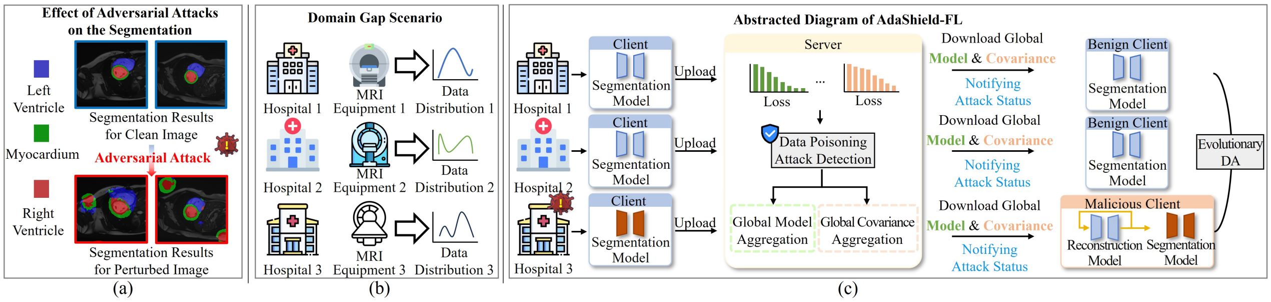


Figure 1: (a) Visualization of the predicted segmentation maps in a patient with "hypertrophic cardiomyopathy" from clean and perturbed data using the FGSM [Goodfellow *et al.*, 2014]. (b) Domain gap scenario in multiple centres with various MRI equipment. (c) Abstracted diagram of the proposed model: detecting malicious clients, purifying, and segmenting regions of interest while aligning the domain gap.

using disentangled reconstruction and segmentation (DRS), and the segmentation model is trained on the purified data with evolutionary DA. For benign clients, the segmentation model is trained on the original data through evolutionary DA. Our source code and appendix are available at https://github.com/alsgur0720/AdaShield. We summarized the contributions below:

- **Federated Purification Framework:** To our knowledge, we are the first to propose an FL framework with a purification method for malicious client's training data, allowing collaboration between different medical centres. It enhances patient data diversity and adversarial robustness in federated medical image segmentation.
- **Disentangled Reconstruction and Segmentation**: We observe that the k-space effectively supports the disentanglement of reconstruction and segmentation features and then propose a purification approach that isolates robust reconstruction features in the k-space, iteratively refines these features, and generates segmentation maps from the purified MRI data.
- **Differential Loss-based Attack Detection:** We observe that the training loss sequence often displays anomalous patterns under adversarial perturbations and then propose a detection of malicious clients by analyzing the loss gradient and curvature, representing abnormal convergence speeds and oscillations, respectively.
- **Evolutionary Domain Adaptation**: We propose an evolutionary adaptive approach that aligns local and global domains by matching covariance matrices and dynamically adjusts the balance between DA and segmentation, enabling automatic and rapid optimization from nonconvex loss functions.

2 Related Works

2.1 Medical Image Segmentation

Medical image segmentation [Liang *et al.*, 2024; Campello *et al.*, 2021; Song *et al.*, 2024; Shi *et al.*, 2024; Sadegheih *et al.*, 2024; Zhou *et al.*, 2021] is essential task in healthcare applications, facilitating accurate diagnosis, treatment planning, and disease monitoring. Specifically, designed explicitly for biomedical image analysis, U-Net introduced an encoder-decoder structure with skip connections, establishing a benchmark for segmentation accuracy. Despite these advances, current methods remain susceptible to data privacy vulnerabilities, underscoring the need for secure and robust segmentation frameworks.

2.2 Federated Learning in Segmentation

Using decentralized data, FL builds a global model focusing on privacy, which has been increasingly adopted for medical image segmentation [Guan *et al.*, 2024; Linardos *et al.*, 2022; Qi *et al.*, 2022; Qiu *et al.*, 2023]. However, these models are susceptible to adversarial attacks that compromise model performance. To overcome this limitation, we propose a purification-based FL framework for malicious clients that refines perturbed patient data.

2.3 Domain Adaptation

In federated medical image segmentation, DA [Zhang *et al.*, 2023; Jiang *et al.*, 2024; Pei *et al.*, 2021] is crucial for mitigating domain gaps arising from heterogeneous data distributions due to variations in clinical centres and MRI equipment. Existing approaches include supervised DA using labeled target domain data with adversarial training and unsupervised methods focusing on translating data between domains, although these methods face challenges, such as high computational costs and instability in convergence. To overcome these obstacles, we propose an evolutionary DA that aligns local and global covariance matrices, enabling rapid optimization and stable convergence from non-convex loss functions.

2.4 Data Poisoning Attack

Adversarial attacks exploit subtle perturbations to undermine neural network robustness. Common attack methods include the fast gradient sign method (FGSM) [Goodfellow *et al.*, 2014], which injects perturbations using the gradient of the loss function. In addition, projected gradient descent (PGD) [Madry *et al.*, 2017] is an extension of the FGSM, that produces stronger attacks by constraining perturbations in a defined boundary. The Carlini and Wagner attack (C&W) [Carlini and Wagner, 2017] also generates adversarial examples as an optimization problem. In this study, we adopt these attacks to demonstrate the adversarial robustness of FL methods designed to defend against such attacks.

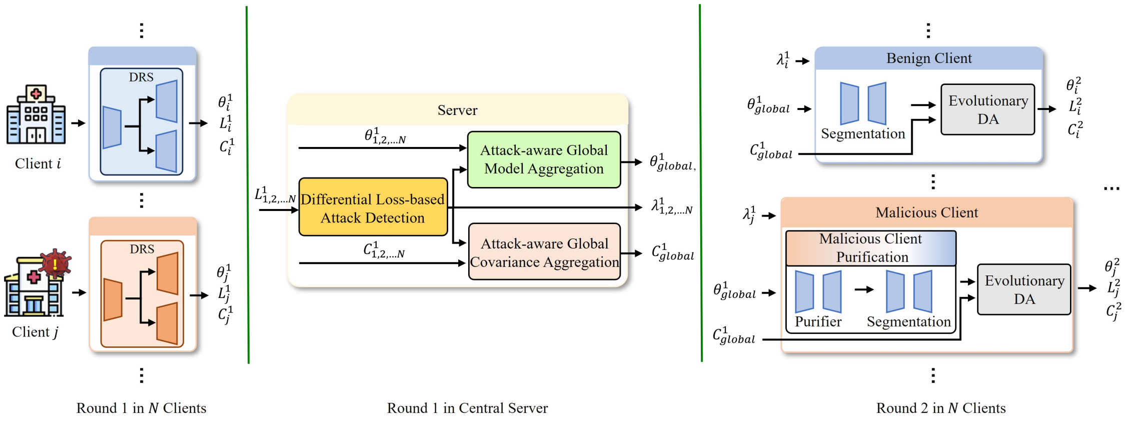


Figure 2: Overall architecture of the AdaShield-FL network.

2.5 Attack Robustness in Federated Learning

A data poisoning attack involves inserting malicious data into training, significantly undermining model accuracy and posing a serious challenge to FL [Kumar *et al.*, 2023]. Recent advancements have introduced adversarial robustness [Errami and Bergou, 2024; Yi *et al.*, 2024; Sun *et al.*, 2024; Cho *et al.*, 2024; Yin *et al.*, 2018; Zhang *et al.*, 2022; Wu *et al.*, 2023b; Li *et al.*, 2020a; Hong *et al.*, 2023]. Specifically, Yi *et al.* [2024] proposed an aggregation method for distributed learning using outlier-resistant one-center and one-mean clustering. In addition, VFLIP [Cho *et al.*, 2024] proposed identification and purification that operate at the inference stage, handling vertically partitioned data on FL participants. However, these approaches are inadequate in handling unseen attacks due to the constraints of adversarial training, or they face reduced patient data diversity by excluding malicious clients during training. To lessen these limitations, AdaShield-FL employs a preemptive strategy that identifies and incorporates purified malicious clients during training.

3 Threat Model

The attacker's goal is to degrade the overall model performance via an untargeted attack, resulting in a decreased Dice score across cardiac regions. To achieve this, the attacker uses adversarial examples to disrupt the training process and is assumed to have prior knowledge of the training data and model architecture in a white-box scenario [Nowroozi *et al.*, 2025]. This scenario allows attackers unauthorized access to training data and model parameters and assumes that the attacker can compromise more than 25% of the client population. This access occurs independently of the FL mechanism, which is designed to safeguard data privacy by ensuring that raw data is not transmitted. In this study, benign clients are uncompromised, and malicious clients are altered by the external attacker. The external attacker who takes control of the local client gains access to local data and can manipulate them [Haffar *et al.*, 2023; Kairouz *et al.*, 2021]. Appendix D provides more details on the adversarial scenario.

4 Method

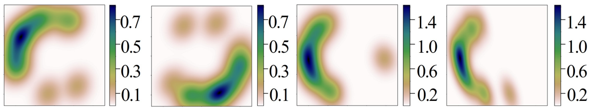
4.1 Overview

As presented in Fig. 2, the Ada-Shield-FL framework maintains patient data diversity and enhances adversarial robustness in federated medical image segmentation. In this framework, each of the N clients trains the local model using its own MRI dataset, whereas a central server performs global model aggregation over R rounds. In the first round, AdaShield-FL trains the DRS, comprising reconstruction and segmentation components, and uploads the trained segmentation model weights $\theta^1_{1,2,...,N}$ to the server. The server also receives training Dice loss sequences $L^1_{1,2,...,N}$ and covariance matrices $C^1_{1,2,...,N}$ from clients. Malicious clients are identified using differential loss-based detection with $L^1_{1,2,...,N}$, resulting in a binary malicious status $\lambda^1_{1,2,...,N}$.

According to the status, the attack-aware global model and covariance aggregations calculate global model weights θ^1_{global} and global covariance matrix C^1_{global}, respectively, by excluding malicious clients. This global weight and matrix are transmitted to clients for the next round. If a malicious status is benign, the client initializes the segmentation model with θ^1_{global} and trains it with evolutionary DA, minimizing the imbalance between segmentation and DA. In contrast, for malicious clients, their local data undergo iterative refinement via the malicious client purification process and are processed via the training process in the same way as benign clients. This overall process iterates across R rounds to enhance adversarial robustness while reducing the domain gap, resulting in a final global segmentation model once all rounds are completed. Appendix A provides overall, malicious client purification, attack detection, and evolutionary DA algorithms.

4.2 Disentangled Reconstruction and Segmentation

Based on class-specific disentangling methods [Yang *et al.*, 2021], we extend this approach to MRI data purification and consider that adversarial attacks primarily disrupt specific tasks. Unlike existing adversarial robustness methods that exclude malicious clients, the proposed purification approach



(a) Clean f_S (b) Perturbed f_S (c) Clean f_R (d) Perturbed f_R

Figure 3: Distributions of the segmentation feature (f_S) and reconstruction feature (f_R) before and after the FGSM attack on the ACDC dataset: (a) f_S extracted from clean data, (b) f_S extracted from perturbed data, (c) f_R extracted from clean data, and (d) f_R extracted from perturbed data.

retains their data to enhance patient data diversity. Specifically, we separate the features for reconstruction and segmentation in the k-space [Sarty *et al.*, 2001], exclude adversarially vulnerable segmentation features, and decode adversarially robust reconstruction features to purify the data for segmentation.

The preliminary study investigated the effect of adversarial attacks targeting segmentation on task-specific disentangled representations in the k-space. To this end, we first extract informative features for reconstruction and segmentation tasks from the k-space by applying the fast Fourier transform (FFT) to the MRI image, as follows:

$$s(k_x(T), k_y(T)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)e^{-i2\pi(xk_x(T)+yk_y(T))}dx\,dy, \tag{1}$$

$$k_x(T) = \frac{\gamma}{2\pi}\int_0^T G_x(\tau)\,d\tau, \quad k_y(T) = \frac{\gamma}{2\pi}\int_0^T G_y(\tau)\,d\tau, \tag{2}$$

where $I(x, y)$ represents the MRI image at position (x, y) and $s(k_x(T), k_y(T))$ denotes the signal value obtained in the k-space coordinate $(k_x(T), k_y(T))$ at a time T, corresponding to the point in time during the MRI acquisition. Moreover, $k_x(T)$ and $k_y(T)$ are proportional to the time integral of the applied magnetic gradient fields, $G_x(\tau)$ and $G_y(\tau)$. The term γ denotes the ratio of the magnetic moment to its angular momentum. The frequency domain is concentrated in the central low-frequencies, whereas high-frequency regions have a sparse distribution [Yang *et al.*, 2014]. Based on this phenomenon, we observed that the sparse k-space, which has the same properties as the frequency domain, is an ideal structure for efficiently disentangling specific tasks.

As depicted in Fig. 3, we use disentangled reconstruction features and segmentation features extracted by a disentangling encoder in the k-space domain and evaluate two types of features under adversarial attack: (1) segmentation features (f_S), and (2) reconstruction features (f_R). Figure 3(a and b) presents the effect of before and after applying the FGSM on the feature distributions of f_S, and Figure 3(c and d) depicts the effect of before and after applying the FGSM on the feature distributions of f_R. Moreover, Figure 3 reveals that perturbations cause a significant shift in the f_S distribution, making it susceptible to attacks, whereas f_R maintains a stable distribution regardless of perturbations.

According to this observation, as displayed in Fig. 4, the proposed approach disentangles f_S and f_R via an encoder trained to minimize the correlation between the two features. In the two distinct feature sets, f_R is optimized for reconstruction using the mean absolute error $\mathcal{L}_{rec}$ between the reconstructed and input MRI data for training reconstruction decoder, and f_S is optimized for segmentation using the Dice loss $\mathcal{L}_{dice}$ for training of segmentation decoder. To enforce independence between f_R and f_S, we introduce an independence loss $\mathcal{L}_{ind}$, is designed to minimize correlation between f_R and f_S for training of disentangling encoder as follows:

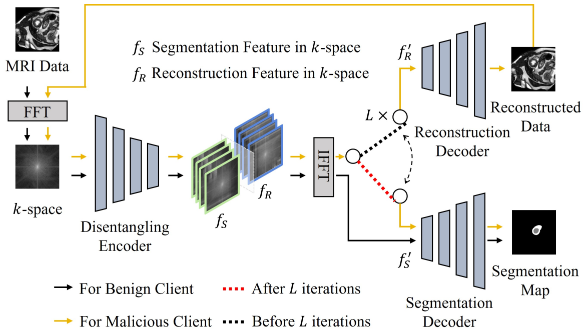


Figure 4: Architecture of disentangled reconstruction and segmentation.

$$\mathcal{L}_{ind} = \frac{1}{h^2}\sum_{k=1}^{h}\sum_{l=1}^{h}\frac{\left(\mathrm{Cov}(f_R, f_S)_{k,l}\right)^2}{\sigma_{f_{R,k}} \cdot \sigma_{f_{S,l}} + \epsilon}, \tag{3}$$

where $\mathrm{Cov}(f_R, f_S)_{k,l}$ denotes the (k, l)-th element of the covariance matrix between f_R and f_S. The terms $\sigma_{f_{R,k}}$ and $\sigma_{f_{S,l}}$ denote the standard deviations of f_R and f_S in the k- and l-th dimensions, respectively. The term h indicates the dimension of the vectors f_R and f_S, and ϵ denotes the stability constant (set to 10^{-9}). This formulation penalizes the correlation between f_R and f_S, promoting their independence. The features f_R and f_S are transformed via the inverse FFT (IFFT), resulting in $f_R^{'}$ and $f_S^{'}$, respectively.

In the malicious process, $f_R^{'}$ is iteratively purified using the reconstruction decoder for L iterations to generate the segmentation map using the purified data and segmentation decoder. For the benign process, $f_S^{'}$ is directly input into the segmentation decoder to generate a segmentation map. Appendix B provides the detailed architecture of the disentangling encoder and decoders, as well as visualizations of f_R and f_S. The trained segmentation weights and covariance matrix obtained using the disentangled feature f_S are uploaded to the central server to aggregate the global model and covariance. Moreover, the training Dice loss sequence is uploaded to the server to detect malicious clients.

4.3 Differential Loss-based Attack Detection and Aggregation

To protect FL systems from data poisoning attacks, we propose a differential loss-based attack detection method analyzes anomalous patterns in training Dice loss sequences. As depicted in Fig. 5, the uploaded training loss sequence of malicious clients typically has a lower convergence speed and more oscillations than benign clients.

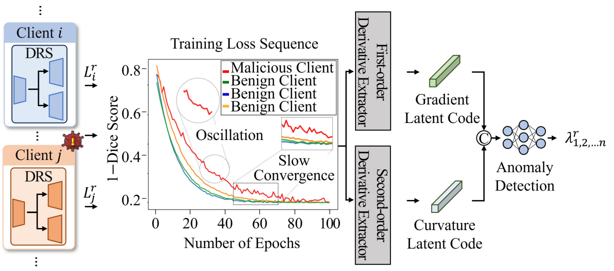


Figure 5: Architecture for the differential loss-based attack detection.

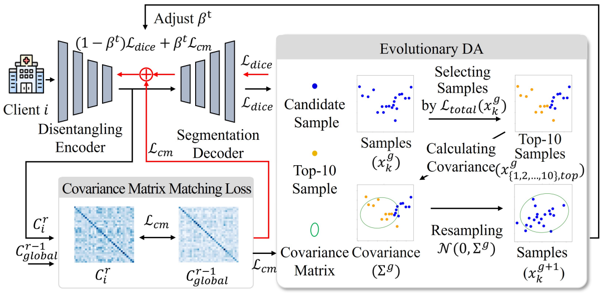


Figure 6: Architecture for the evolutionary domain adaptation.

Two feature extractors were designed to compute the first and second derivatives and detect anomalies using a multi layer perceptron with three layers to detect malicious clients based on abnormal patterns. The gradient latent code $\Delta L_{1,2,\ldots,N}^r$, representing convergence speed, is extracted using the first-order derivative extractor in the r-th round. The second-order derivative extractor creates the curvature latent code $\Delta^2 L_{1,2,\ldots,N}^r$, representing the loss oscillation. The two resulting differential loss patterns, $\Delta L_{1,2,\ldots,N}^r$ and $\Delta^2 L_{1,2,\ldots,N}^r$, are concatenated to serve as input for the anomaly detection model, outputting the malicious status.

If the i-th client is detected as malicious in the r-th round, $\lambda_i^r = 1$; otherwise, $\lambda_i^r = 0$. The server aggregates the model weights of clients using attack-aware global model aggregation (Fig. 2) as follows:

$$\theta^{r+1} \leftarrow \theta^r - \eta \frac{1}{N - \sum_{i=1}^N \lambda_i^r} \times \sum_{i=1}^N \frac{g_i^r \cdot (1-\lambda_i^r)}{\sqrt{G^r} + \epsilon}, \quad (4)$$

where θ^r denotes the weights of the global segmentation model on the r-th round, and g_i^r represents the gradient of the local model in the i-th client of the r-th round. Furthermore, G^r indicates the cumulative sum of squares of the gradient (momentum) $G^r = G^{r-1} + \sum_{i=1}^N (g_i^r)^2$, and η denotes the learning rate (set to 0.01). In Eq. (4), malicious clients ($\lambda_i^r = 1$) are excluded from the aggregation.

In addition, global covariance aggregation excludes malicious clients using $\lambda_{1,\cdots,N}^r$ to calculate the covariance matrices of exclusively benign clients. The attack-aware global covariance aggregation is formulated as follows:

$$C_{global}^r = \frac{1}{\sum_{i=1}^N \rho_i (1-\lambda_i^r)} \Bigg[\sum_{i=1}^N \rho_i C_i^r (1-\lambda_i^r) + \sum_{i=1}^N \rho_i (\mu_i^r - \mu^r)(\mu_i^r - \mu^r)^T (1-\lambda_i^r) \Bigg], \quad (5)$$

where C_i^r represents the covariance matrix of the i-th client in the r-th round obtained by calculating the mean and covariance based on the disentangled segmentation feature f_s extracted by the disentangling encoder. Moreover, μ_i^r and ρ_i represent the mean and number of samples for the i-th client in the r-th round, respectively, and μ^r represents the mean for all clients in the r-th round.

In addition, the server transmits $\lambda_{1,\cdots,N}^r$ to clients to notify them of the attack status, C_{global}^r to align global and local covariances, and θ^r to initialize the segmentation model for the next training round.

4.4 Evolutionary Domain Adaptation

In Fig. 1(b), data from various equipment across clinical centres create a domain gap in FL due to variations in scanner types, in-plane resolution, and slice thickness. Existing DA methods use complete data, including features irrelevant to segmentation; hence, they tend not to guarantee the ideal alignment for the optimal segmentation performance. This work focuses on the primary segmentation task by employing disentangled segmentation features f_S instead of full data in covariance matrix matching loss $\mathcal{L}_{cm}$, as illustrated in Fig. 6.

Specifically, $\mathcal{L}_{cm}$ minimizes the data heterogeneity between clients by applying the covariance matching from disentangled segmentation features f_s, as follows:

$$\mathcal{L}_{cm} = \left\| \frac{C_i^{r+1}}{\rho_i} - \frac{C_{global}^r}{\rho} \right\|_F^2, \quad (6)$$

where C_{global}^r denotes the global covariance matrix in the r-th round, $\|\cdot\|_F^2$ represents the Frobenius norm and ρ indicates the number of samples for all clients. The loss function minimizes heterogeneity by matching between the local and global covariance matrices. In this approach, because the covariance matrix does not include personal information, it is suitable for FL, which is crucial for preserving privacy.

In addition, when conventional DA is integrated with the target task, the balancing parameters of the loss function are typically determined empirically. Thus, finding the optimal balancing parameters for each client is labor-intensive, as it needs to be performed individually for every client. The following evolutionary DA approach is applied to optimize the balance automatically by dynamically combining $\mathcal{L}_{cm}$ and the segmentation Dice loss $\mathcal{L}_{dice}$ to address this problem:

$$\mathcal{L}_{total} = (1-\beta^t)\mathcal{L}_{dice} + \beta^t \mathcal{L}_{cm}, \quad (7)$$

where β^t denotes the dynamic balancing parameter for prioritizing $\mathcal{L}_{dice}$ and $\mathcal{L}_{cm}$, with t representing the training iteration. We adaptively prioritizes the two tasks by dynamically adjusting β^t, extending the covariance matrix adaptation-evolution strategy (CMA-ES) [Hansen *et al.*, 2003] to its new application to DA. The balancing parameter β^t is iteratively

Model	Dice Score (%)			
	FGSM	PGD	C&W	Benign (No Attack)
FedAvg [McMahan *et al.*, 2017]	69.3	68.8	68.1	86.4
Yi *et al.* [2024]	71.3	70.9	71.0	86.4
IOS [Wu *et al.*, 2023b]	71.4	70.7	71.5	86.9
Karimireddy *et al.* [2021]	71.0	71.4	70.9	87.0
Li *et al.* [2020a]	73.2	72.9	72.1	86.8
FedRBN [Hong *et al.*, 2023]	72.6	72.3	72.5	87.5
AdaShield-FL	**82.7**	**82.9**	**82.9**	**88.4**

Table 1: Comparison with prior methods on the M&Ms validation set in terms of the Dice score. Each federated learning method is trained on the perturbed data generated by each adversarial attack.

Model	Dice Score (%)			
	FGSM	PGD	C&W	Benign (No Attack)
FedAvg [McMahan *et al.*, 2017]	67.3	67.0	67.1	85.9
Yi *et al.* [2024]	70.3	71.8	70.6	86.1
IOS [Wu *et al.*, 2023b]	69.8	69.2	69.3	86.3
Karimireddy *et al.* [2021]	69.3	69.6	69.4	86.0
Li *et al.* [2020a]	70.0	70.5	70.2	85.9
FedRBN [Hong *et al.*, 2023]	70.7	71.7	70.3	86.3
AdaShield-FL	**81.0**	**82.6**	**81.7**	**87.3**

Table 2: Comparison with prior methods on the ACDC validation set in terms of the Dice score. Each federated learning method is trained on the perturbed data generated by each adversarial attack.

optimized over $(g+1)$ generations for each t-th iteration. In dynamic prioritizing, the candidate solutions $x^{g+1}_{1,...,20}$ for each generation over g are sampled as follows:

$$x^{g+1}_k = m^g + \sigma^g \cdot \mathcal{N}(0, \mathbf{\Sigma}^g), \quad k = 1, ..., 20, \tag{8}$$

where $\mathcal{N}(0, \mathbf{\Sigma}^g)$ represents a multivariate normal distribution with a zero mean and covariance matrix $\mathbf{\Sigma}^g$, in the g-th generation. In Fig. 6, the lengths of the major and minor axes of the ellipse are proportional to the eigenvalues of the covariance, and their orientations are determined by eigenvectors of the covariance. Moreover, m^g denotes the mean value in the g-th generation and σ^g indicates the step size in the g-th generation. The mean value m^{g+1} for the next generation is updated as follows:

$$m^{g+1} = \sum_{i=1}^{10} w_i x^{g+1}_{i,top}, \tag{9}$$

where w_i denotes the standard proportional weighting [Hansen *et al.*, 2003], and $x^{g+1}_{i,top}$ denotes the top 10-ranked population determined based on Eq. (7), with lower values assigned to top positions in the ranking. Furthermore, the balancing parameter β^t is set to m^{g+1} if the update magnitude of m^{g+1} is less than ϵ_{cma} (set to 10^{-5}), according to the following condition:

$$\beta^t = m^{g+1}, \quad \text{if } \|m^{g+1} - m^g\| \leq \epsilon_{cma}, \tag{10}$$

Unless the condition in Eq. (10) is met, dynamic prioritizing continues as in Eqs. (8) and (9). This automatic balance enables $\mathcal{L}_{total}$ to achieve efficient optimization and convergence by adaptively prioritizing segmentation and DA.

5 Experiments

5.1 Dataset

The experiments employed a dataset created by combining M&Ms [Campello *et al.*, 2021], a multi-vendor dataset, including Siemens, Philips, GE, and Canon, established from 375 participants, and ACDC [Bernard *et al.*, 2018], a single-vendor (Canon) dataset, established from 150 participants. Both datasets consist of cardiac MRI sequences, with each patient having 21 frames. Moreover, each dataset comprises data from six centres and includes four disease labels: "dilated cardiomyopathy", "hypertrophic cardiomyopathy", "abnormal right ventricle", and "normal". It also provides three target labels for segmentation: "left ventricle", "right ventricle", and "myocardium". Additionally, AdaShield-FL is trained for eight clients (N = 8), with four clients using the M&Ms dataset and the other using the ACDC dataset.

5.2 Implementation Details

The experiments follow a standard federated-by-dataset scenario [McMahan *et al.*, 2017], where each client constructs their own dataset and collaborates in FL with a central server. All experiments were conducted in PyTorch with two Nvidia H100 GPUs. This study employs stochastic gradient descent optimization with a momentum of 0.9 and a learning rate of 0.001 for training. AdaShield-FL was trained for 500 global rounds (R = 500), setting L to 3 in the malicious client purification. Appendix B provides quantitative and qualitative experimental results for the hyperparameter L.

5.3 Results and Analysis

This work compares AdaShield-FL with existing FL methods designed for adversarial robustness, including methods by Yi *et al.* [2024], Karimireddy *et al.* [2021], Li *et al.* [2020a], IOS [Wu *et al.*, 2023b], and FedRBN [Hong *et al.*, 2023] on the M&Ms and ACDC datasets, as listed in Tables 1 and 2. This study adopts widely used adversarial attacks for the image segmentation task, such as FGSM, PGD, and C&W.

In addition, the backbone network for the widely used segmentation proposed by Zhou *et al.* [2021] was employed. For each experiment, two malicious clients were randomly selected. Appendix B provides the comparison results when more clients are perturbed. These models were trained and evaluated from scratch, following the experimental settings outlined by the authors, using their provided open-source codes. In all tables, the best scores are in bold.

In Tables 1 and 2, AdaShield-FL demonstrates superior segmentation performance in terms of the average Dice score coefficient compared to other adversarial robustness-based FL methods. It is attributed to purifying and incorporating malicious clients' data into the training process, enhancing model robustness and patient data diversity. Furthermore, outperforming other FL methods when trained with benign data demonstrates the effectiveness of evolutionary DA.

In addition, Table 3 compares the accuracy of attack detection methods for identifying benign and malicious clients. Compared to existing approaches, such as IOS, Yi *et al.*, Karimireddy *et al.*, and Li *et al.*, AdaShield-FL consistently provides more reliable performance in terms of recall

Method	Metric	M&Ms			ACDC		
		FGSM	PGD	C&W	FGSM	PGD	C&W
Yi *et al.*	Recall	0.85	0.87	0.85	0.81	0.80	0.83
	Precision	0.79	0.80	0.78	0.75	0.77	0.80
IOS	Recall	0.82	0.80	0.81	0.79	0.77	0.79
	Precision	0.74	0.71	0.70	0.69	0.68	0.72
Karimireddy *et al.*	Recall	0.84	0.88	0.85	0.81	0.78	0.77
	Precision	0.84	0.83	0.81	0.79	0.80	0.76
Li *et al.*	Recall	0.73	0.71	0.74	0.72	0.73	0.69
	Precision	0.69	0.73	0.70	0.67	0.71	0.70
AdaShield-FL	Recall	**0.95**	**0.96**	**0.94**	**0.91**	**0.91**	**0.92**
	Precision	**0.98**	**0.99**	**0.98**	**0.95**	**0.94**	**0.97**

Table 3: Detection performance in terms of recall and pre-cision for detecting malicious clients on the M&Ms and ACDC datasets.

Metric		Dice Score		
DA method	FL Method	FGSM	PGD	C&W
IPLC [Wu *et al.*, 2023a]	Karimireddy *et al.*	72.3	72.6	72.5
UPL-SFDA [Zhang *et al.*, 2024]	IOS	74.9	73.4	74.2
UPL-SFDA [Zhang *et al.*, 2024]	Karimireddy *et al.*	74.0	73.9	74.1
IPLC [Wu *et al.*, 2023a]	IOS	72.1	72.8	72.0
AdaShield-FL		**82.7**	**82.9**	**82.9**

Table 4: Performance of FL methods with DA models for perturbed data generated by each adversarial attack on the M&Ms dataset.

Model	FLOPs (G)	Params (M)	Convergence Round	Training Time (hours)
FedAvg	336	**14.4**	554	9.2
Yi *et al.*	336	**14.4**	539	8.9
Karimireddy *et al.*	336	**14.4**	527	8.7
Li *et al.*	336	15.3	521	8.6
AdaShield-FL	**326**	21.9	**470**	**7.8**

Table 5: Computational complexity of FL methods on the M&Ms dataset.

Differential Loss-based Attack Detection	Evolutionary DA	Malicious Client Purification	Dice Score
✓	✓	✓	**82.9**
✓	✓		75.4
✓		✓	80.7
	✓		72.2
✓			74.8
			68.8

Table 6: Ablation study for AdaShield-FL on the M&Ms dataset perturbed by PGD attack in terms of the Dice score.

Dice Score			
$\beta = 0.4$	$\beta = 0.5$	$\beta = 0.6$	Evolutionary Strategy (β^t)
81.2	80.0	81.1	**82.7**

Table 7: Effect of the balancing parameter (β) on Dice score in DRS on the M&Ms dataset under the FGSM attack.

and precision, demonstrating its robustness against data poisoning attacks and effectiveness of Dice loss sequence.

Furthermore, Table 4 reveals the performance of FL adversarial robustness models integrated with DA methods. This experiments employs IPLC [Wu *et al.*, 2023a] and UPL-SFDA [Zhang *et al.*, 2024] for DA. Further, it employs the model by Karimireddy *et al.* and IOS, byzantine aggregation methods that can be integrated with DA models for federated medical image segmentation. As a results, the proposed model consistently outperforms Karimireddy *et al.* and IOS combined with DA methods. These results highlight the effectiveness of covariance matrix matching loss via f_S and evolutionary DA in enhancing segmentation performance.

Table 5 compares FL methods regarding floating-point operations per second (FLOPs), Params, convergence rounds, and training time, with eights clients. In these methods, the common segmentation model [Zhou *et al.*, 2021] with 42 GFLOPs and 1.8M Params is used for each client. Since the disentangling encoder in AdaSheild-FL operates in the k-space, which is inherently sparse, the FLOPs are reduced to 33 GFLOPs compared to the original segmentation encoder. Although AdaShield-FL introduces additional parameters due to its reconstruction decoder (0.9M Params for each client) and attack detection (0.3M Params for central server), it achieves the fastest convergence (470 rounds), the shortest total training time (7.8 h), and the fewest total operating points (326 GFLOPs). This improved computational efficiency is attributed to the use of evolutionary DA, which enables automatic and rapid optimization from non-convex loss functions and the disentangling encoder in the k-space. Moreover, Appendix C offers visual qualitative segmentation results and Appendix E presents limitations and future work.

5.4 Ablation Study

This section analyzes the performance of each component in AdaShield-FL. In Table 6, the checkmark (✓) indicates that a module was activated. The first row shows the performance of AdaShield-FL, incorporating all modules. The second and third rows report the results when malicious client purification and evolutionary DA are excluded, respectively. The fourth and fifth rows report the results when only evolutionary DA and differential loss-based attack detection are included, respectively. Last, the final row displays the backbone performance. Comparing the first and remaining rows reveal that each proposed module improves performance.

Furthermore, Table 7 illustrates the influence of β in the evolutionary DA on the Dice score. The balancing parameter β controls the priority between the Dice and DA loss. In Table 7, the evolutionary strategy outperforms several fixed values of β. This is attributed to the ability of the evolutionary DA to dynamically adjust β during training, ensuring an optimal balance between segmentation and DA. Appendix B includes additional ablation studies, experiments that highlight the effectiveness of $\mathcal{L}_{ind}$, and experiments on another modality, the computed tomography (CT) dataset.

6 Conclusion

This paper addresses the critical challenges posed by the retained data poisoning attacks and data heterogeneity in federated medical image segmentation. We propose AdaShield-FL, a comprehensive framework that integrates k-space disentangled purification, attack detection, and evolutionary DA to address these problems. AdaShield-FL identifies malicious clients, purifies malicious clients, and matches data distributions to improve segmentation accuracy in the FL framework. The experimental results demonstrate that AdaShield-FL outperforms existing FL methods, achieving state-of-the-art performance on the M&Ms and ACDC datasets while achieving computational efficiency and rapid convergence.

Acknowledgments

This work was supported by the IITP grant funded by the Korea government (MSIT) (No.2021-0-02068, RS-2023-00256629, RS-2022-00156287, RS-2024-00437718).

References

[Bernard *et al.*, 2018] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

[Campello *et al.*, 2021] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[Cho *et al.*, 2024] Yungi Cho, Woorim Han, Miseon Yu, Younghan Lee, Ho Bae, and Yunheung Paek. Vflip: A backdoor defense for vertical federated learning via identification and purification. In *European Symposium on Research in Computer Security*, pages 291–312. Springer, 2024.

[Errami and Bergou, 2024] Latifa Errami and El Houcine Bergou. Tolerating outliers: Gradient-based penalties for byzantine robustness and inclusion. In *IJCAI*, 2024.

[Finlayson *et al.*, 2018] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Guan *et al.*, 2024] Hao Guan, Pew-Thian Yap, Andrea Bozoki, and Mingxia Liu. Federated learning for medical image analysis: A survey. *Pattern Recognition*, page 110424, 2024.

[Haffar *et al.*, 2023] Rami Haffar, David Sanchez, and Josep Domingo-Ferrer. Explaining predictions and attacks in federated learning via random forests. *Applied Intelligence*, 53(1):169–185, 2023.

[Hansen *et al.*, 2003] Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11(1):1–18, 2003.

[Hong *et al.*, 2023] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Federated robustness propagation: sharing adversarial robustness in heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7893–7901, 2023.

[Jiang *et al.*, 2023] Meirui Jiang, Holger R Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, and Ziyue Xu. Fair federated medical image segmentation via client contribution estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16302–16311, 2023.

[Jiang *et al.*, 2024] Enyi Jiang, Yibo Jacky Zhang, and Sanmi Koyejo. Principled federated domain adaptation: Gradient projection and auto-weighting. In *The Twelfth International Conference on Learning Representations*, 2024.

[Kairouz *et al.*, 2021] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

[Karimireddy *et al.*, 2021] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pages 5311–5319. PMLR, 2021.

[Kaviani *et al.*, 2022] Sara Kaviani, Ki Jin Han, and Insoo Sohn. Adversarial attacks and defenses on ai in medical imaging informatics: A survey. *Expert Systems with Applications*, 198:116815, 2022.

[Kumar *et al.*, 2023] K Naveen Kumar, C Krishna Mohan, and Linga Reddy Cenkeramaddi. The impact of adversarial attacks on federated learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[Li *et al.*, 2020a] Suyi Li, Yong Cheng, Wei Wang, Yang Liu, and Tianjian Chen. Learning to detect malicious clients for robust federated learning. *arXiv preprint arXiv:2002.00211*, 2020.

[Li *et al.*, 2020b] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical image analysis*, 65:101765, 2020.

[Liang *et al.*, 2024] Guoyan Liang, Qin Zhou, Jingyuan Chen, Zhe Wang, and Chang Yao. Advancing medical image segmentation via self-supervised instance-adaptive prototype learning. In *IJCAI*, 2024.

[Linardos *et al.*, 2022] Akis Linardos, Kaisar Kushibar, Sean Walsh, Polyxeni Gkontra, and Karim Lekadir. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Scientific Reports*, 12(1):3551, 2022.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and

Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Nowroozi *et al.*, 2025] Ehsan Nowroozi, Imran Haider, Rahim Taheri, and Mauro Conti. Federated learning under attack: Exposing vulnerabilities through data poisoning attacks in computer networks. *IEEE Transactions on Network and Service Management*, 2025.

[Ozbulak *et al.*, 2019] Utku Ozbulak, Arnout Van Messem, and Wesley De Neve. Impact of adversarial examples on deep learning models for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*, pages 300–308. Springer, 2019.

[Pei *et al.*, 2021] Chenhao Pei, Fuping Wu, Liqin Huang, and Xiahai Zhuang. Disentangle domain features for cross-modality cardiac image segmentation. *Medical Image Analysis*, 71:102078, 2021.

[Qi *et al.*, 2022] Xiaoming Qi, Guanyu Yang, Yuting He, Wangyan Liu, Ali Islam, and Shuo Li. Contrastive re-localization and history distillation in federated cmr segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 256–265. Springer, 2022.

[Qiu *et al.*, 2023] Liang Qiu, Jierong Cheng, Huxin Gao, Wei Xiong, and Hongliang Ren. Federated semi-supervised learning for medical image segmentation via pseudo-label denoising. *IEEE journal of biomedical and health informatics*, 27(10):4672–4683, 2023.

[Sadegheih *et al.*, 2024] Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, and Dorit Merhof. Lhu-net: A light hybrid u-net for cost-efficient, high-performance volumetric medical image segmentation. *arXiv preprint arXiv:2404.05102*, 2024.

[Sarty *et al.*, 2001] Gordon E Sarty, Raoqiong Bennett, and Robert W Cox. Direct reconstruction of non-cartesian k-space data using a nonuniform fast fourier transform. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 45(5):908–915, 2001.

[Shi *et al.*, 2024] Jun Shi, Shulan Ruan, Ziqi Zhu, Minfan Zhao, Hong An, Xudong Xue, and Bing Yan. Predictive accuracy-based active learning for medical image segmentation. In *IJCAI*, 2024.

[Song *et al.*, 2024] Zhengxuan Song, Xun Liu, Wenhao Zhang, Yongyi Gong, Tianyong Hao, and Kun Zeng. Spgnet: A shape-prior guided network for medical image segmentation. In *IJCAI*, 2024.

[Sun *et al.*, 2024] Peng Sun, Xinyang Liu, Zhibo Wang, and Bo Liu. Byzantine-robust decentralized federated learning via dual-domain clustering and trust bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24756–24765, 2024.

[Wu *et al.*, 2023a] Jianghao Wu, Guotai Wang, Ran Gu, Tao Lu, Yinan Chen, Wentao Zhu, Tom Vercauteren, Sébastien Ourselin, and Shaoting Zhang. Upl-sfda: Uncertainty-aware pseudo label guided source-free domain adaptation for medical image segmentation. *IEEE transactions on medical imaging*, 2023.

[Wu *et al.*, 2023b] Zhaoxian Wu, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*, 2023.

[Yang *et al.*, 2014] Yang Yang, Feng Liu, Wenlong Xu, and Stuart Crozier. Compressed sensing mri via two-stage reconstruction. *IEEE Transactions on biomedical engineering*, 62(1):110–118, 2014.

[Yang *et al.*, 2021] Shuo Yang, Tianyu Guo, Yunhe Wang, and Chang Xu. Adversarial robustness through disentangled representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3145–3153, 2021.

[Yi *et al.*, 2024] Yuhao Yi, Ronghui You, Hong Liu, Changxin Liu, Yuan Wang, and Jiancheng Lv. Near-optimal resilient aggregation rules for distributed learning using 1-center and 1-mean clustering with outliers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16469–16477, 2024.

[Yin *et al.*, 2018] Dong Yin, Yudong Chen, Ramchandran Kannan, and Peter Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International conference on machine learning*, pages 5650–5659. Pmlr, 2018.

[Zhang *et al.*, 2022] Zaixi Zhang, Xiaoyu Cao, Jinyuan Jia, and Neil Zhenqiang Gong. Fldetector: Defending federated learning against model poisoning attacks via detecting malicious clients. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2545–2555, 2022.

[Zhang *et al.*, 2023] Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, and Yanfeng Wang. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3954–3963, 2023.

[Zhang *et al.*, 2024] Guoning Zhang, Xiaoran Qi, Bo Yan, and Guotai Wang. Iplc: iterative pseudo label correction guided by sam for source-free domain adaptation in medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 351–360. Springer, 2024.

[Zhou *et al.*, 2021] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.