Fed-DFA: Federated Distillation for Heterogeneous Model Fusion Through the Adversarial Lens

Zichen Wang, Feng Yan, Tianyi Wang, Cong Wang*, Yuanchao Shu, Peng Cheng, Jiming Chen

College of Control Science and Engineering, Zhejiang University
{withnorman, yanfeng555, wty1998, cwang85, ycshu, lunarheart, cjm}@zju.edu.cn

Abstract

Most of the federated learning techniques are limited to homogeneous model fusion. With the rapid growth of smart applications on resource-constrained edge devices, it becomes a barrier to accommodate their heterogeneous computing power and memory in the real world. Federated Distillation is a promising alternative that enables aggregation from heterogeneous models. However, the effectiveness of knowledge transfer still remains elusive under the shadow of distinct representation power from heterogeneous models. In this paper, we approach from an adversarial perspective to characterize the decision boundaries during distillation. By leveraging K-step PGD attacks, we successfully model the dynamics of the closest boundary points and establish a quantitative connection between the predictive uncertainty and boundary margin. Based on these findings, we further propose a new loss function to make the distillation attend to samples close to the decision boundaries, thus learning from more informed logit distributions. The extensive experiments over CIFAR-10/100 and Tiny-ImageNet demonstrate about 0.5-3.5% improvement of accuracy under different IID and non-IID settings, with only a small increment of computational overhead.

Introduction

Today's Federated Learning (FL) framework mainly aggregates knowledge from the *homogeneous* models (McMahan et al. 2017). However, in practice, the misalignment of computation time and memory capacity across *heterogeneous* edge devices often leads to the problems of load imbalance and memory error (Wang, Yang, and Zhou 2021), which makes such one-model-fits-all design incompatible with the real-world demands. Model personalization aims to leverage heterogeneous models to balance the varying capacities and constraints on edge devices. A straightforward solution is to find commonalities in sub-model structures (Diao, Ding, and Tarokh 2020; Horvath et al. 2021; Alam et al. 2022; Wang et al. 2023), but these methods are still confined to the same model family and incompatible with the full-fledged heterogeneous model fusion such as transferring the knowledge between convolution and vision transformers.

Federated Distillation (FD) is a viable way to accommodate heterogeneous models (Li and Wang 2019; Lin et al. 2020; Zhu, Hong, and Zhou 2021; Cho et al. 2022; Liu et al. 2022). Inherited from knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015), FD is model-agnostic — it replaces parameterized model fusion (McMahan et al. 2017) by matching client predictions with the consensus logits and minimizes the Kullback-Leibler (KL) divergence, thereby transferring the "dark knowledge" (additional information embedded in soft probabilities) among the participants. The seminal work of FedDF (Lin et al. 2020) has shown empirically that FD achieves a comparative performance of FedAvg (McMahan et al. 2017) while enabling knowledge transfer among the heterogeneous models.

However, KD typically suffers from the infamous capacity gap problem when the student finds it hard to mimic the predictive distribution of the teacher (Son et al. 2021; Mirzadeh et al. 2020; Zhu and Wang 2021a). Since FD inherits the algorithmic backbone from KD but distills in a distributed, online fashion, *would the capacity gap still persist in FD*, and *in what forms?* Previous works have yet to give sufficient insights under federated settings (Lin et al. 2020).

Motivated by these fundamental quests, we first present empirical findings of a complex, mutual influence between the light and heavyweight models — lightweight models progress at an inevitable cost of degrading the heavyweight models and over-parameterized networks (potential teachers) no longer guarantee high performance in FD. Such disparity becomes more prominent and difficult to rectify when the consensus is aggregated on *unlabeled public data* (Zhu, Hong, and Zhou 2021; Cho et al. 2022). If misclassified by the majority, the consensus would mislead model convergence towards a wrong direction, and degrade the overall performance (Du et al. 2020).

Prior works use logit variance to re-weight sample importance in distillation (Cho et al. 2022). Unfortunately, our implementation implies a weak correlation between logit variance and the true labels with a high false positive rate (wrong predictions could have higher variance as well). Blindly enlarging the weights of these samples would mislead the consensus. Are there other measures that can characterize the heterogeneous model capacity more reliably? In this paper, we orchestrate K-step PGD attack as a proxy (Zhang et al. 2020) and build it into the FD pipeline to quantify instance-specific boundary margins that work for a mixture of convo-

*Corresponding Author.

lution networks and ViTs. In particular, we propose a new framework called Fed-DFA (FedDF through the Adversarial Lens) to make distillation attend to samples in the vicinity of the consensus decision boundary. The main contributions are summarized below:

- ✧ We provide new empirical findings of the capacity gap and decision boundaries of heterogeneous model fusion in FD and unveil a latent correlation between boundary margin and predictive uncertainty. To the best of our knowledge, this is the first work that leverages adversarial learning to improve generalization for FD.
- ✧ We propose Fed-DFA to make distillation attend to samples near the decision boundaries, which enables FD to learn from more informed distributions than the overconfident distributions with less information. We also analyze the generalization bound upon domain adaptations.
- ✧ We demonstrate the efficacy of Fed-DFA over a CIFAR-10/100 and Tiny-ImageNet, by enabling knowledge transfer between convolution and vision transformers. The results indicate a 1.5-3.5% improvement compared to the current SOTA of FedDF with even better generalization capabilities under non-IID data.

Background and Related Works

Knowledge Distillation

The classic knowledge distillation (KD) methods transfer the knowledge from one or an ensemble of pre-trained teachers to small-capacity students via minimizing a weighted combination of the cross-entropy and KL divergence (Hinton, Vinyals, and Dean 2015). A plethora of existing efforts focus on closing the teacher-student capacity gap (Son et al. 2021; Mirzadeh et al. 2020; Zhu and Wang 2021a) such as using teaching assistants as intermediate hubs (Son et al. 2021; Mirzadeh et al. 2020); utilizing gradient similarity to enable knowledge transfer (Zhu and Wang 2021a) and distilling from the model checkpoints (Wang et al. 2022). In (Zhu and Wang 2021b), the intrinsic dimension is used to quantify the capacity gap, and a two-step mutual distillation is proposed. Moreover, the previous works have shown the effectiveness of adopting adaptive and instance-specific temperatures (Li et al. 2022), multi-level logit distillation (Jin, Wang, and Lin 2023), and two-way mutual knowledge transfer (Zhang et al. 2018). Different from the prior efforts, we delve into the dynamics of decision boundaries in FD.

Personalized Federated Learning

The existing research takes different directions to accommodate personalized model architectures, which are categorized into *sub-model fusion* (Diao, Ding, and Tarokh 2020; Horvath et al. 2021; Alam et al. 2022; Wang et al. 2023) and *distillated model fusion* (He, Annavaram, and Avestimehr 2020; Lin et al. 2020; Zhu, Hong, and Zhou 2021; Cho et al. 2022; Liu et al. 2022). Sub-model fusion finds a common subset of model structures: HeteroFL (Diao, Ding, and Tarokh 2020) and FjORD (Horvath et al. 2021) perform static extraction of sub-models from the large server model. FedRolex (Alam et al. 2022) extracts the sub-model on a rolling basis for diversified parameter aggregation. FlexiFed (Wang et al. 2023) utilizes the commonalities of architectures within the same network family, e.g., ResNet or VGG. However, these works assume the models to share a backbone structure that is still constrained by the same representational power. There is also a collection of exotic designs outside the FedAvg framework (He, Annavaram, and Avestimehr 2020; Lin et al. 2020; Zhu, Hong, and Zhou 2021; Cho et al. 2022; Liu et al. 2022). However, none of these works have reasoned from the challenges of adopting heterogeneous models with distinct representational power. This work approaches from an adversarial perspective to establish a connection between decision boundary dynamics and heterogeneous model fusion. The closest work to ours is (Nam et al. 2021), in which uniformly random perturbations are introduced to diversify the output logits and avoid overconfidence. Yet, the uniformly perturbed samples might obscure the original consensus without the hard labels. In this work, we leverage the adversarial examples in a non-intrusive manner to guide the distillation process.

Preliminary

Consider a number of N participants with heterogeneous models tailored to their device memory and computational capacity. Each client n first performs gradient descent on his private dataset $\mathbb{D}_n = \{x_i, y_i\}$, where x_i are the data samples and $y_i = \{1, \cdots, C\}$ are the label space for C classes. The goal is to learn from $\mathbb{D} = \{\mathbb{D}_1, \cdots, \mathbb{D}_n\}$ with heterogeneous models $\mathbf{w}_n$. The process consists of local training and global distillation:

$$\textit{Local Training}: \quad \mathbf{w}_n = \mathbf{w}_n - \eta_1 \nabla_{\mathbf{w}_n} \mathcal{L}_{CE}(\mathbf{w}_n, \xi_n), \tag{1}$$

where η_1 is the local learning rate, $\mathcal{L}_{CE}$ is the cross-entropy loss and ξ_n is the mini-batch of data $\mathbb{D}_n$. Once the local training is completed, each participant samples mini-batches of $\mathbf{x}^p = \{x_i^p\} \in \mathbb{D}_p$ from an unlabeled public dataset $\mathbb{D}_p$ to derive the averaged logits of consensus $\frac{1}{N}\sum_{n=1}^{N} f_{\mathbf{w}_n}(\mathbf{x}^p)$. Then each participant transfers the knowledge by aligning their local model outputs with the global consensus.
Global Distillation:

$$\mathbf{w}_n = \mathbf{w}_n - \eta_2 \nabla_{\mathbf{w}_n} \mathcal{L}_{KL}\Big(\sigma\big(\frac{1}{N}\sum_{n=1}^{N}\frac{f_{\mathbf{w}_n}(\mathbf{x}^p)}{R}\big), \sigma\big(\frac{f_{\mathbf{w}_n}(\mathbf{x}^p)}{R}\big)\Big), \tag{2}$$

where $\mathcal{L}_{KL}$ is the KL Divergence, $\sigma(\cdot)$ is the softmax function, η_2 is the learning rate of distillation and R is the distillation temperature. After the knowledge distillation is completed, $\mathbf{w}_n$ is used as the starting point for the next iteration of local training.

Adversarial-Guided Federated Distillation

Understanding Heterogeneous Model Fusion

To gain a deeper understanding of FD, we are interested in answering: 1) Does FD suffer from a similar capacity gap in canonical KD? 2) Can we improve knowledge transfer in the absence of true labels via some latent attributes embedded inside the models? How would other distributional artifacts such as non-IID affect KD? We start with an empirical study

on a vanilla setup of 5 participants, who select the models randomly from the x-axis in Fig. 1.

Observation 1 (Capacity Gaps). The capacity gap still persists as a complex, multi-faceted problem in FD: 1) The same model exhibits perceptible performance variance ($\Delta 3 \sim 6\%$ mAP) while collaborating with different models under various model combinations. This indicates a complex interplay between different models in FD. 2) Even under the same model combination, heterogeneous models have a high-performance variance ($\Delta 10\%$ mAP);

Unlike KD, in which students match their output logits to *pre-trained* teachers in an *offline* fashion, in FD, students with less representative power have a successive impact on their teachers online. As a result, teachers only perform slightly better than the students and over-parameterization cannot help improve the teachers. This makes pre-selection of "experts" (assigning higher weight values (Cho et al. 2022)) difficult when prior knowledge such as model parameters is no longer an accurate measure.

Such mutual influence is mainly attributed to the knowledge transfer when participants minimize the KL loss to match the consensus logits. Reasoned in (Ojha et al. 2024), these low-dimensional logit distributions "encode" the relative positions of samples from the decision boundaries. When the majority of models make wrong decisions, the consensus would mislead the KL loss towards a wrong target, thereby producing a misinformed decision boundary. This process cannot be simply rectified without the true labels of public data, thus leaving us in a paradoxical situation.

Prior efforts utilize logit variance as an implicit measure of decision confidence to guide weighted aggregations (Cho et al. 2022). However, due to the fast-growing exponentiation of softmax, small values are magnified and even the wrong decisions become overconfident, approaching a one-hot vector. Our experiments find this phenomenon becoming more prominent under the non-IID data as most of the secondary class probabilities drop to near zero, leaving the output logits with little cues except the one-hot labels. Thus, instead of logit variance, are there other metrics to supplement the distillation process?

Heterogeneous Model Boundaries

To answer this question, we resort to quantitative representations of the heterogeneous model boundaries and examine the possibility of using this latent information.

Definition 1 (Boundary Margin). Define the boundary margin estimate $\tilde{M}_{\mathbf{w}_n}(x_i^P)$ as the distance from a sample $x_i^P \in \mathbb{D}_P$ to the decision boundary of model $\mathbf{w}_n$ in the pixel-space $\mathcal{X}$. The true margin $M_{\mathbf{w}_n} \triangleq \tilde{M}_{\mathbf{w}_n}$ when $\mathbb{D}_P$ and $\mathbb{D}_n$ are identically distributed. Denote $f_{\mathbf{w}_n}(x_i^P)$ as the soft logit predictions and $\hat{y}_i^P = \arg\max_{\hat{y} \in \mathcal{C}} f_{\mathbf{w}_n}(x_i^P)$ as the prediction result with maximum probability. We have,

$$\tilde{M}_{\mathbf{w}_n}(x) = \min_{x_i^P} \|x_i^P - x\|_p, \tag{3}$$

$$f_{\mathbf{w}_n}(x_i^P) - \max_{\hat{y}_i^P \neq y'} f_{\mathbf{w}_n}(x_i^P) = 0, \tag{4}$$

where (4) represents when the boundary margin from soft logits $f_{\mathbf{w}_n}(x_i^P)$ to a point until the output decision has changed. Direct calculation of (3) is computation-intensive for C-class classification. Hence, we relax the problem into finding the closest boundary point from a public data point x_i^P. This allows us to leverage K-step PGD as a proxy to quantify boundary margin by observing when the top-1 probability has changed. This method efficiently approximates (3) based on the gradient information of intermediate model weights $\mathbf{w}_n$ in (1) before distillation.

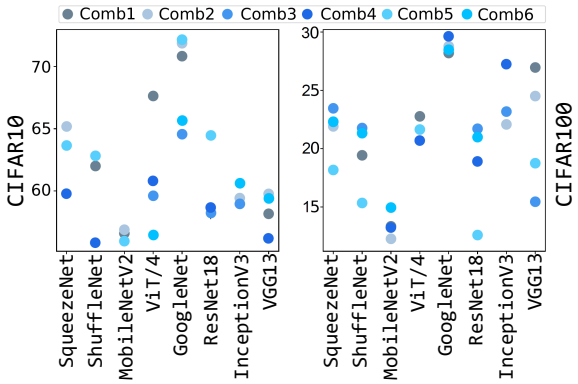


Figure 1: Testing accuracy from 6 model combinations on CIFAR-10/100. Contrary to the common beliefs in KD that heavyweight models often have higher accuracy, we observe that those models are only on par with the average performance in FD.

K-PGD Estimate. Denote x_0' as the starting point x_i^P, we draw x_i^P from a subset of public data to launch K-PGD attacks in the participants model,

$$\text{Repeat: } x_{k+1}' = \Pi_\epsilon(x_k' + \gamma \text{sign}(\nabla_{x_t'} \mathcal{L}_{CE}(f_{\mathbf{w}_n}(x_i^P), \hat{y}_i^P)),$$

$$\text{Until: } \arg\max_{c \in \mathcal{C}} f_{\mathbf{w}^n}(x_k') \neq \hat{y}_i^P, \tag{5}$$

$$\text{where } \|x_i^P - x_k'\|_\infty \leq \epsilon, k = \{1, \cdots, K\}. \tag{6}$$

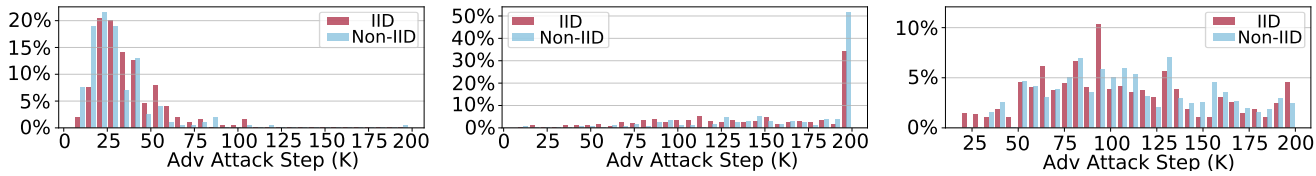
in which Π_ϵ projects the sample into the l_∞ ball, ϵ is noise bound, γ is the step size, $\hat{y}_i^P$ is the argmax label of the prediction from x_i^P. Denote the above process as a function of PGD steps $f_{\mathbf{w}_n}^{\text{PGD}}(x_i^P)$ for public data x_i^P. The boundary margin can be formally estimated by,

$$\tilde{M}_{\mathbf{w}_n}(x_i) \propto f_{\mathbf{w}_n}^{\text{PGD}}(x_i^P). \tag{7}$$

Then we leverage (7) to capture the boundary dynamics in the FD process.

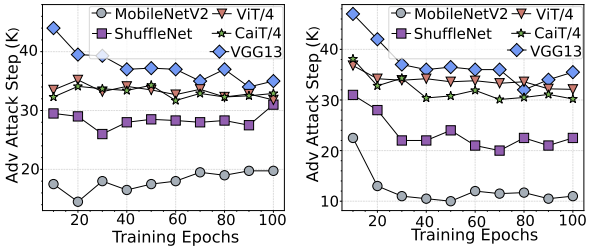
Observation 2 (Boundary Dynamics). We find several intriguing properties empirically:

- ✧ Lightweight models (MobileNetV2) have smaller boundary margins and the decisions are under-confident compared to heavyweight models with much larger margins such as VGG shown in Figs. 2. Vision transformers are less confident compared to VGG, which is consistent with the recent findings from (Kim et al. 2024).
- ✧ Shown in Fig.3(a), as FD progresses, the lightweight models exhibit an upward trend and the opposite is observed for the heavyweight models (VGG) in the IID settings, which echoes with the previous findings of mutual



(a) MobileNetV2 (b) VGG13 (c) ViT/4

Figure 2: Sample-wise distribution of PGD steps K between heterogeneous models (illustrated in Observation2): Lightweight CNN models such as MobileNet are underconfident and heavyweight CNN models are overconfident, whereas vision transformer (ViT/4) is in between.



(a) IID (b) Non-IID

Figure 3: Tracing the closest boundary points in terms of K-PGD during training dynamics: 1) IID settings; 2) non-IID settings.

influence between heterogeneous models. The decision boundary of vision transformers is relatively more stable.

✧ The boundary margins all have downward trends under non-IID data observed from Fig.3(b), indicating that distributional shifts drive samples closer to the decision boundaries. It becomes more difficult to find an optimal boundary to distinguish the non-IID data and knowledge transfer is less effective.

Thus, we confirm that heterogeneous models exhibit distinctive decision boundaries in the learning process.

Connecting Boundary Margin with Predictive Uncertainty

With the new insights of heterogeneous boundary margins, we further look into their connections with predictive uncertainty, an effective measure of logit diversity that corresponds to model generalization (Dubey et al. 2018). We quantify the predictive uncertainty on an instance level by the Shannon Entropy,

$$\mathbf{H}(\mathbf{x}) = -\sum_{c=1}^{C} P(y_c|\mathbf{x}_i^P; \mathbf{w}_n) \log P(y_c|\mathbf{x}_i^P; \mathbf{w}_n), \tag{8}$$

where $P(y_c|\mathbf{x}_i^P; \mathbf{w}_n)$ is the probability of the c-th category. We use the Spearman Correlation Coefficient (Wang, Yan, and Yan 2023) to establish the connection between the entropy and boundary margin. Spearman Correlation is more robust to model non-normal distributed data with a focus on the monotonic relationships. We first rank the entropy $\mathbf{H}$ and PGD step K in an ascending order, using the rank numbers R_{Hi} and R_{Ki}, and define $\bar{R}_H = \frac{1}{N}\sum_{i=1}^{N} R_{Hi}$ and $\bar{R}_K = \frac{1}{N}\sum_{i=1}^{N} R_{Ki}$. The Spearman correlation coefficient ρ is,

$$\rho = \frac{\sum_{i=1}^{N}(R_{Hi} - \bar{R}_H)(R_{Ki} - \bar{R}_K)}{\sqrt{\sum_{i=1}^{N}(R_{Hi} - \bar{R}_H)^2 \sum_{i=1}^{N}(R_{Ki} - \bar{R}_K)^2}}. \tag{9}$$

Observation 3 (Entropy vs. Boundary Margin). As shown in Fig.5(a), there is a strong correlation between the predictive entropy and boundary margin, i.e., samples that lie close to the decision boundaries (small K) tend to have higher predictive uncertainty and vice versa. Since (Cho et al. 2022) use logit variance as a metric to re-weight samples (samples with larger variance are assigned with larger weights in aggregation), we also establish the relation between the logit variance and true labels in Fig. 5(b). It is observed that although the variance increases with a closer l_2 distance to the true labels, there is a large number of false positives with high logit variance, which leads to wrong decisions. Assigning these samples with larger weights could obscure the consensus by misleading the optimization in the wrong direction. This is also due to the paradox from *unlabeled* public data as the re-weighting approach still struggles without effective supervision.

Proposed Method

Based on the discussions above, we see that although knowledge gaps seem inevitable, the heterogeneous models in FD could be better differentiated on the instance level given the boundary margins. We posit that distillation should attend to samples closer to the decision boundaries. This helps distillation match logits with higher entropy and transform knowledge from more informed predictive distributions rather than overconfident ones. Thus, for different models, we calculate the consensus of boundary margins on each mini-batch,

$$\overline{\mathbf{K}} = \frac{1}{NB}\sum_{n=1}^{N}\sum_{i=1}^{B} f_{\mathbf{w}_n}^{\text{PGD}}(x_i^P), K_{\text{th}} = \text{med}\{\overline{\mathbf{K}}\} \tag{10}$$

where $\overline{\mathbf{K}}$ is a vector of averaged PGD steps on a mini-batch B of public data. Then we sort $\overline{\mathbf{K}}$ and set the median as K_{th}. This partitions the public data into $x^+(\overline{\mathbf{K}} \leq K_{th})$ and $x^-(\overline{\mathbf{K}} > K_{th})$, in which we use the $+$ and $-$ signs to denote

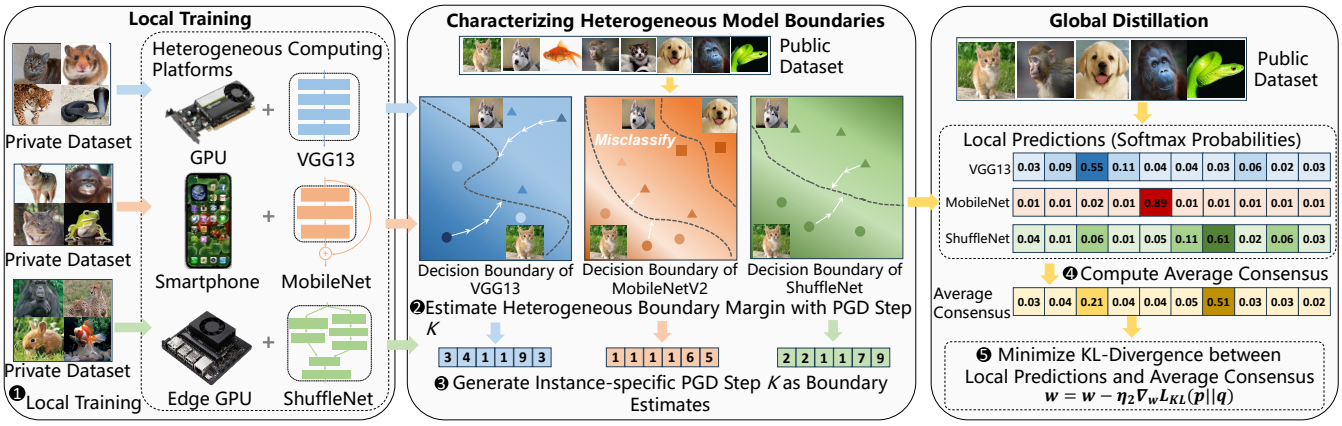


Figure 4: Illustration of Fed-DFA: ❶ Local training on private data;❷ Execute K-PGD attacks to estimate boundary margin; ❸ Enumerate public data to generate instance-specific boundary estimates; ❹ Compute consensus by averaging the prediction probabilities; ❺ Minimize KL divergence between consensus and model predictions.

whether the distillation should pay more or less attention to. Then, we replace (2) with the new loss function,

$$\mathcal{L}'_{KL} = \mathbb{E}_{x^+\sim\mathbb{D}_P}\left[\mathcal{L}_{KL}\left(\sigma\left(\frac{1}{N}\sum_{n=1}^{N}\frac{f_{\mathbf{w}_n}(x^+)}{R}\right),\sigma\left(\frac{f_{\mathbf{w}_n}(x^+)}{R}\right)\right)\right] + \beta\cdot\mathbb{E}_{x^-\sim\mathbb{D}_P}\left[\mathcal{L}_{KL}\left(\sigma\left(\frac{1}{N}\sum_{n=1}^{N}\frac{f_{\mathbf{w}_n}(x^-)}{R}\right),\sigma\left(\frac{f_{\mathbf{w}_n}(x^-)}{R}\right)\right)\right]. \tag{11}$$

β is a scaling factor with $0 \leq \beta \leq 1$. Next, we derive the generalization bound according to (Ben-David et al. 2010).

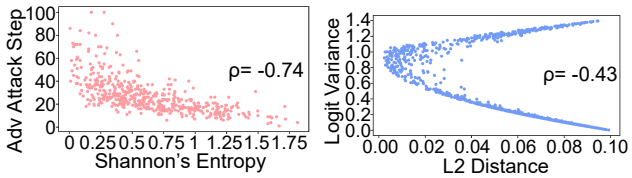
Theorem 1 (Generalization Bound). For N participants with the *true* data distribution $\mathcal{D}_n$ of the n-th local domain and the *true* global distribution as $\mathcal{D}$, denote $\hat{\mathcal{D}}_n$ and $\hat{\mathcal{D}}$ as the empirical distribution with samples of size m each, drawn from $\mathcal{D}_n$ and $\mathcal{D}$, respectively. According to (10), $\mathcal{D}_n$ can be considered a mixture of distributions $\mathcal{D}_n = \mathcal{D}_n^+ \cup \mathcal{D}_n^-$.

Consider hypothesis h, $\mathcal{X} \to \mathcal{Y}$, from the input space $\mathcal{X}$ to label space $\mathcal{Y}$ with hypotheses space $\mathcal{H}$. h_n is the hypothesis learned from $\mathcal{D}_n$, that $h_n = \arg\min_h \mathcal{L}_{\mathcal{D}_n}(h)$ and $\hat{h}_n = \arg\min_h \mathcal{L}_{\hat{\mathcal{D}}_n}(h)$. $d_{\mathcal{H}\Delta\mathcal{H}}(\hat{D}, D)$ is defined as the divergence over $\mathcal{H}$. For any $\tau \in (0,1)$, the following bound holds with probability at least $1-\tau$,

$$\mathcal{L}_{\mathcal{D}}\left(\sum_{n=1}^{N} h_k\right) \leq \sum_{n=1}^{N}\mathcal{L}_{\hat{\mathcal{D}}_n}(h_k) + \frac{1}{2}\sum_{n=1}^{N} d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_n^+, \hat{\mathcal{D}}) + \frac{1}{2}\sum_{n=1}^{N}\beta d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_n^-, \hat{\mathcal{D}}) + \sum_{n=1}^{N}\lambda_n + 4\sqrt{\frac{2d\log 2m + \log\frac{4}{\tau}}{m}} \tag{12}$$

where $\mathcal{L}_{\hat{\mathcal{D}}_n}(h_k)$ is the empirical loss on $\hat{\mathcal{D}}_n$, $\lambda_n = \min_h(\mathcal{L}_{\mathcal{D}}(h) + \mathcal{L}_{\mathcal{D}_n}(h))$ is the combined error of the hypothesis.

Theorem 1 implies that a larger divergence from $d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_n^+, \hat{\mathcal{D}})$ and $d_{\mathcal{H}\Delta\mathcal{H}}(\hat{\mathcal{D}}_n^-, \hat{\mathcal{D}})$ degrades the overall generalization and more samples m reduce the loss at an $\mathcal{O}(\log m/\sqrt{m})$ rate. The impact of such distributional shifts is available in the supplement materials.



(a) Entropy vs. adv steps (b) Logit var. vs. dist. to label

Figure 5: Analysis of predictive uncertainty: a) K-PGD vs. Shannon entropy of logits; b) Logit variance vs. L2 distance between predictive outputs and the true labels, in which a large number of false positives are found. Instead, K-PGD displays a stronger correlation with predictive uncertainty.

Reduce Computational Overhead. To model the dynamics of decision boundary closely, it requires generating adversarial examples in each epoch during training. Hence, the computational overhead scales linearly with the number of participants and distillation data size. From Fig. 3, the boundary dynamics become more stabilized as the model converges, this allows to reduce computation via using a *stale* estimate of the boundary.

Experiments

Experimental Setup

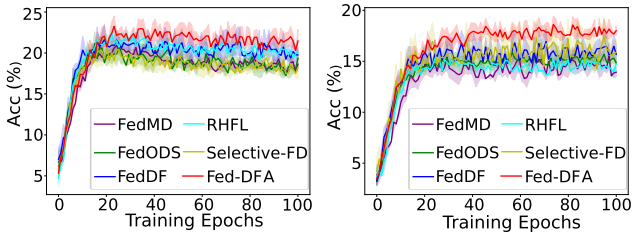
Datasets and Heterogeneous Models. We conduct extensive experiments on the CIFAR-10/100 and Tiny-ImageNet Datasets. To cover a large collection of heterogeneous models, we form different combinations with a mixture of convolution and ViTs as shown in Table 1 and 2. We adopt the Dirichlet distribution to generate non-IID data as in (Lin et al. 2020), which uses α to evaluate different intensities of non-IIDness. A smaller α represents a higher degree of non-IIDness.

Baselines. We compare with the following baselines:

✧ FedMD (Li and Wang 2019): One of the earliest FD methods that only require limited black box access.

		CIFAR-10						CIFAR-100					
		MobileNetV2	ShuffleNet	VGG13	ViT/4	CaiT/4	Average	MobileNetV2	ShuffleNet	VGG13	ViT/4	CaiT/4	Average
IID	FedMD	61.71	64.57	69.92	54.07	47.35	59.52	20.02	26.35	24.23	18.62	19.23	21.69
	FedDF	63.34	**67.13**	70.64	53.42	49.84	60.87	22.19	**28.25**	23.92	19.38	19.27	22.60
	FedODS	63.67	66.35	**73.58**	53.26	49.54	61.28	**22.38**	26.64	26.18	22.36	19.43	23.40
	RHFL	63.36	**67.43**	72.73	**55.22**	49.33	**61.61**	21.32	27.30	**27.48**	**23.02**	20.05	**23.83**
	Selective-FD	**64.54**	66.52	70.98	54.70	**50.58**	61.46	21.88	26.83	25.72	22.66	**20.20**	23.46
	Fed-DFA	**64.01**	66.25	**73.65**	**55.81**	**51.42**	**62.23**	**24.88**	**28.36**	**28.84**	**25.86**	**22.36**	**26.06**
Non-IID ($\alpha = 1$)	FedMD	**57.00**	60.53	65.33	49.82	49.51	56.44	**21.59**	25.74	20.86	20.22	16.59	21.00
	FedDF	53.33	62.45	**72.87**	50.99	50.09	57.95	20.70	22.80	**27.56**	20.91	18.31	22.06
	FedODS	55.71	61.03	**73.62**	47.43	**51.45**	57.85	21.02	24.36	25.72	22.32	**18.94**	22.47
	RHFL	54.12	63.63	72.53	49.92	50.60	58.16	20.93	24.10	**29.07**	20.26	18.77	**22.63**
	Selective-FD	55.47	**64.98**	71.21	**52.49**	49.92	**58.81**	19.99	**26.53**	25.85	**22.89**	16.93	22.44
	Fed-DFA	**57.53**	**64.33**	71.97	**51.79**	**51.53**	**59.43**	**22.39**	**26.56**	27.46	**24.09**	**19.83**	**24.07**
Non-IID ($\alpha = 0.1$)	FedMD	33.61	42.32	**44.89**	35.05	34.77	38.13	14.00	17.74	15.79	15.24	13.72	15.30
	FedDF	**34.32**	44.98	44.15	35.43	34.45	38.67	15.28	17.62	19.84	14.88	14.60	16.44
	FedODS	32.23	45.53	**45.65**	34.99	**35.65**	38.81	15.82	18.72	18.11	**15.98**	14.65	16.66
	RHFL	32.06	**53.50**	38.49	**37.58**	34.58	**39.24**	14.72	**19.62**	**21.85**	15.67	13.75	17.12
	Selective-FD	**35.78**	48.97	44.11	35.25	31.16	39.05	**18.03**	17.22	20.88	15.72	**14.82**	**17.33**
	Fed-DFA	32.98	**54.26**	42.20	**38.47**	**35.61**	**40.70**	**17.94**	**20.73**	**21.72**	**16.02**	**17.66**	**18.81**

Table 1: Comparison of mAP (%) on CIFAR-10/100. Two top numbers are bolded with the best in Red and the second in Blue.



(a) ShuffleNet (b) CaiT/4

Figure 6: Learning curves on CIFAR-100 ($\alpha = 0.1$).

- ✧ FedDF (Lin et al. 2020): A comprehensive and state-of-the-art FD framework.
- ✧ FedODS (Nam et al. 2021): We utilize the method to maximize diversities of output logits in FD by generating public data perturbed with Output Diversification Sampling (ODS) (Tashiro, Song, and Ermon 2020).
- ✧ RHFL (Fang and Ye 2022): A framework that improves the robustness of heterogeneous FD against noisy labels.
- ✧ Selective-FD (Shao, Wu, and Zhang 2023): Selective-FD utilizes a selective knowledge-sharing mechanism to identify knowledge from local and ensemble predictions.

Implementation Details. We use the Adam optimizer for both the local training and global distillation and set the learning rate to 10^{-3}. Our base testing includes 1 local epoch and 10 distillation epochs, and the temperature $R = 1$ unless stated otherwise. We adopt l_∞ PGD attacks with a step size $\gamma = 0.01$, $\epsilon = 0.1$, and $K = 5$ to explore the decision boundary. We set $\beta = 0.1$ in (11) to put more attention on the samples closer to the boundaries.

Performance Comparison of mAP

CIFAR-10/100. Table 1 compares the proposed Fed-DFA on three convolutional models (MobileNetV2, ShuffleNet and VGG13) and two vision transformers (ViT/4 and CaiT/4). Our goal is to not only focus on individual performance but also assess the average performance and the knowledge transfer between models under the IID/non-IID settings. It is observed that Fed-DFA outperforms all the benchmarks in terms of the average mAP across the 5 models. In particular, it outperforms the current SOTA of FedDF by $1.4-2.0\%$ on CIFAR-10 and $2.0-3.5\%$ on CIFAR-100. In comparison, FedMD slightly suffers from the catastrophic forgetting to switch between local training and distillation, thus cannot effectively transfer knowledge between different models. Since boundary attack is adopted, the computation time of FedODS is significantly higher than other methods while its performance is identical to Fed-DFA only on ViT and CaiT. The noise learning/confidence re-weighting mechanism enables RHFL to achieve performance second to Fed-DFA. Selective-FD can identify accurate and precise knowledge during the FD process thus improving FedDF. There are also some interesting details about Fed-DFA in the IID/non-IID data with different datasets. First, Fed-DFA characterizes the decision boundaries learned from non-IID data well as it outperforms all the competitors under different degrees of non-IIDness ($\alpha = 0.1, 1$). Further, the performance gain becomes even larger on CIFAR-100. This is because predictions are less confident with more classes in CIFAR-100, which gives a larger pool of samples with predictive uncertainty that could further boost the performance of Fed-DFA. Fig. 6 provides the convergence of two representative models. We can see that Fed-DFA begins to functionalize after 20 epochs when the testing accuracy of other methods has plateaued.

		VGG16	ResNet50	ViT/16	Average	Std (↓)
IID	FedMD	20.10	28.24	25.69	24.68	**4.16**
	FedDF	20.54	**29.79**	26.26	25.53	4.67
	FedODS	20.30	29.25	26.50	25.35	4.58
	RHFL	20.89	29.13	**27.20**	**25.74**	4.31
	Selective-FD	**20.93**	29.48	25.78	25.40	4.29
	Fed-DFA	**21.61**	**29.82**	**26.81**	**26.08**	**4.15**
Non-IID ($\alpha = 0.1$)	FedMD	10.61	17.75	15.65	14.67	3.67
	FedDF	11.26	17.63	16.60	15.16	3.42
	FedODS	**10.99**	17.78	16.53	15.10	3.61
	RHFL	11.23	**18.22**	16.80	15.42	3.69
	Selective-FD	12.15	18.08	**17.15**	**15.79**	**3.19**
	Fed-DFA	**12.99**	**18.29**	**17.85**	**16.38**	**2.94**

Table 2: Comparison of mAP on Tiny-ImageNet. Top numbers are bolded with the best in Red and the second in Blue.

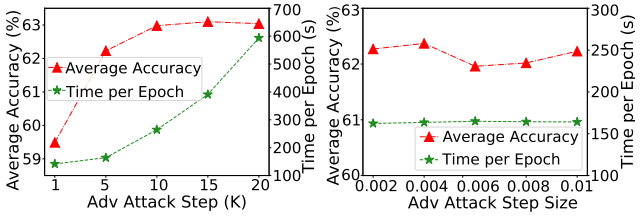
	CIFAR-10		CIFAR-100	
	Acc (↑)	Time (↓)	Acc (↑)	Time (↓)
FedDF-6%	60.87	**114.6**	22.60	**116.1**
FedDF-12%	61.77	138.0	25.21	138.8
FedDF-24%	**62.54**	185.8	**27.10**	188.1
FedDF-Random-6%	61.53	**137.5**	24.96	**136.7**
Fed-DFA-6%	**62.23**	169.1	**26.06**	169.6

Table 3: Comparison of testing accuracy (%) and computation time per epoch (in seconds). FedDF-6% represents 6% of the public distillation data used.

Tiny-ImageNet. Table 2 compares mAP on Tiny-ImageNet over VGG16, ResNet50 and ViT/16. Normally, the decision boundary becomes more difficult to characterize under complex data distributions such as ImageNet. Fed-DFA outperforms the baseline methods under both IID/Non-IID settings. Further, Fed-DFA achieves the lowest performance variation which potentially reduces the capacity gap among the participants on complex classification tasks.

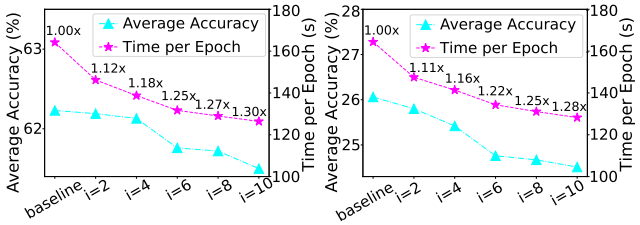
Ablation Studies

Amount of Distillation Data. Theorem 1 states that the loss scales down at $\mathcal{O}(\log m/\sqrt{m})$ regarding the amount of distillation data m. We compare Fed-DFA with several baseline variations: 1) FedDF-6/12/24% represent that a fixed amount of 6/12/24% public distillation data are used; 2) FedDF-Random-6% represents the 6% of distillation data are randomly replaced with new data in each epoch. We compare with Fed-DFA when 6% of the distillation data is used. Table 3 shows that accuracy increases with more distillation data, but the computation time also increases. It is interesting to see that Fed-DFA-6% can achieve the same or higher accuracy than FedDF-12% of double data size. Distilling samples near the decision boundary provides higher performance compared with distilling from a dynamic set of random samples (FedDF-Random-6%). In sum, although our proposed method slightly increases the computational cost due to K-PGD attacks, it utilizes the distillation data more efficiently (less than half of the baseline).



(a) Adversarial step (K) (b) Adversarial step size

Figure 7: Accuracy vs. adversarial setups on CIFAR-10: a) step K; b) step size.



(a) CIFAR-10 (b) CIFAR-100

Figure 8: Accuracy vs. computation cost. The x-axis is how often boundary estimates are conducted.

Impact of Adversarial Step K and Size. K determines how closely the decision boundary is characterized: a larger K brings higher precision at an increasing computational cost. We change K from $1-20$ to examine its impact on accuracy and computation speed in Fig. 7(a). We observe that the accuracy jumps significantly when K increases from 1 to 5, but with marginal improvements over 10. Hence, we set $K = 5$ to achieve a good balance between precision and computation speed. Fig.7(b) illustrates the impact of adversarial step size from $0.002 - 0.01$ when $K = 5$. Both accuracy and computation time are not sensitive to the step size.

Computation Reduction. To reduce computational overhead, we reduce the frequency of boundary estimates (estimate for every $i = 2, 4, 6, 8, 10$ epoch) and trace its accuracy change in Fig. 8, i.e., distilling from a stale estimate of the decision boundary. We observed that accuracy declines as i increases, e.g., $1.3\times$ speed-up causes 1% drop. A good balance occurs at $i = 4$ with $1.1\times$ speed-up and less than 0.5% accuracy drop. This could match FedDF's computation time while achieving higher accuracy.

Conclusion

In this paper, we propose Fed-DFA for heterogeneous model fusion from an adversarial perspective. We successfully capture decision boundary dynamics characterized by the K-step PGD and integrate this into a new loss function to make distillation attend to samples close to the decision boundaries for better generalization. Our extensive experiments over various datasets demonstrate the effectiveness of the proposed method compared to the benchmarks.

Acknowledgements

This work is supported in part by Natural Science Foundation of Zhejiang Province Z25F020029, National Natural Science Foundation of China 62394341, National Natural Science Foundation of China 62293511, and the Fundamental Research Funds for the Central Universities 226202400182.

References

Alam, S.; Liu, L.; Yan, M.; and Zhang, M. 2022. Fedrolex: Model-heterogeneous federated learning with rolling submodel extraction. *Advances in neural information processing systems*, 35: 29677–29690.

Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. 2010. A theory of learning from different domains. *Machine Learning*, 79: 151–175.

Cho, Y. J.; Manoel, A.; Joshi, G.; Sim, R.; and Dimitriadis, D. 2022. Heterogeneous ensemble knowledge transfer for training large models in federated learning. *arXiv preprint arXiv:2204.12703*.

Diao, E.; Ding, J.; and Tarokh, V. 2020. HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients. In *International Conference on Learning Representations*.

Du, S.; You, S.; Li, X.; Wu, J.; Wang, F.; Qian, C.; and Zhang, C. 2020. Agree to Disagree: Adaptive Ensemble Knowledge Distillation in Gradient Space. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12345–12355. Curran Associates, Inc.

Dubey, A.; Gupta, O.; Raskar, R.; and Naik, N. 2018. Maximum-Entropy Fine Grained Classification. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Fang, X.; and Ye, M. 2022. Robust Federated Learning With Noisy and Heterogeneous Clients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10072–10081.

He, C.; Annavaram, M.; and Avestimehr, S. 2020. Group knowledge transfer: Federated learning of large cnns at the edge. *Advances in Neural Information Processing Systems*, 33: 14068–14080.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *Computer Science*, 14(7): 38–39.

Horvath, S.; Laskaridis, S.; Almeida, M.; Leontiadis, I.; Venieris, S.; and Lane, N. 2021. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems*, 34: 12876–12889.

Jin, Y.; Wang, J.; and Lin, D. 2023. Multi-Level Logit Distillation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24276–24285.

Kim, J.; Park, J.; Kim, S.; and Lee, J.-S. 2024. Curved representation space of vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 13142–13150.

Li, D.; and Wang, J. 2019. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Li, X.-C.; Fan, W.-s.; Song, S.; Li, Y.; Li, b.; Yunfeng, S.; and Zhan, D.-C. 2022. Asymmetric Temperature Scaling Makes Larger Networks Teach Well Again. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 3830–3842. Curran Associates, Inc.

Lin, T.; Kong, L.; Stich, S. U.; and Jaggi, M. 2020. Ensemble distillation for robust model fusion in federated learning. *Advances in Neural Information Processing Systems*, 33: 2351–2363.

Liu, R.; Wu, F.; Wu, C.; Wang, Y.; Lyu, L.; Chen, H.; and Xie, X. 2022. No One Left Behind: Inclusive Federated Learning over Heterogeneous Devices. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 3398–3406. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393850.

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

Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5191–5198.

Nam, G.; Yoon, J.; Lee, Y.; and Lee, J. 2021. Diversity Matters When Learning From Ensembles. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 8367–8377. Curran Associates, Inc.

Ojha, U.; Li, Y.; Sundara Rajan, A.; Liang, Y.; and Lee, Y. J. 2024. What knowledge gets distilled in knowledge distillation? *Advances in Neural Information Processing Systems*, 36.

Shao, J.; Wu, F.; and Zhang, J. 2023. Selective Knowledge Sharing for Privacy-Preserving Federated Distillation without A Good Teacher. arXiv:2304.01731.

Son, W.; Na, J.; Choi, J.; and Hwang, W. 2021. Densely guided knowledge distillation using multiple teacher assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9395–9404.

Tashiro, Y.; Song, Y.; and Ermon, S. 2020. Diversity can be Transferred: Output Diversification for White- and Black-box Attacks. In *Advances in Neural Information Processing Systems*.

Wang, C.; Yang, Q.; Huang, R.; Song, S.; and Huang, G. 2022. Efficient Knowledge Distillation from Model Checkpoints. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 607–619. Curran Associates, Inc.

Wang, C.; Yang, Y.; and Zhou, P. 2021. Towards Efficient Scheduling of Federated Mobile Devices Under Computational and Statistical Heterogeneity. *IEEE Transactions on Parallel and Distributed Systems*, 32(2): 394–410.

Wang, H.; Yan, J.; and Yan, X. 2023. Spearman Rank Correlation Screening for Ultrahigh-Dimensional Censored Data. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, 10104–10112.

Wang, K.; He, Q.; Chen, F.; Chen, C.; Huang, F.; Jin, H.; and Yang, Y. 2023. FlexiFed: Personalized Federated Learning for Edge Clients with Heterogeneous Model Architectures. In *Proceedings of the ACM Web Conference 2023*, WWW '23, 2979–2990. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394161.

Zhang, J.; Xu, X.; Han, B.; Niu, G.; Cui, L.; Sugiyama, M.; and Kankanhalli, M. 2020. Attacks Which Do Not Kill Training Make Adversarial Learning Stronger. In *ICML*.

Zhang, Y.; Xiang, T.; Hospedales, T. M.; and Lu, H. 2018. Deep Mutual Learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4320–4328.

Zhu, Y.; and Wang, Y. 2021a. Student Customized Knowledge Distillation: Bridging the Gap Between Student and Teacher. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5037–5046.

Zhu, Y.; and Wang, Y. 2021b. Student Customized Knowledge Distillation: Bridging the Gap Between Student and Teacher. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5037–5046.

Zhu, Z.; Hong, J.; and Zhou, J. 2021. Data-free knowledge distillation for heterogeneous federated learning. In *International conference on machine learning*, 12878–12889. PMLR.