One-shot Federated Learning on Medical Data using Knowledge Distillation with Image Synthesis and Client Model Adaptation

Myeongkyun Kang[1,3][0000−0002−9165−870X],
Philip Chikontwe[1][0000−0002−6995−2312], Soopil Kim[1,3][0000−0001−8937−6263],
Kyong Hwan Jin[2][0000−0001−7885−4792], Ehsan Adeli[3][0000−0002−0579−7763],
Kilian M. Pohl[3][0000−0001−5416−5159], and
Sang Hyun Park[1⋆][0000−0001−7476−1046]

[1] Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea {mkkang,shpark13135}@dgist.ac.kr
[2] Electrical Engineering and Computer Science, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Korea
[3] Stanford University, Stanford, CA 94305, USA

Abstract. One-shot federated learning (FL) has emerged as a promising solution in scenarios where multiple communication rounds are not practical. Notably, as feature distributions in medical data are less discriminative than those of natural images, robust global model training with FL is non-trivial and can lead to overfitting. To address this issue, we propose a novel one-shot FL framework leveraging Image Synthesis and Client model Adaptation (FedISCA) with knowledge distillation (KD). To prevent overfitting, we generate diverse synthetic images ranging from random noise to realistic images. This approach (i) alleviates data privacy concerns and (ii) facilitates robust global model training using KD with decentralized client models. To mitigate domain disparity in the early stages of synthesis, we design noise-adapted client models where batch normalization statistics on random noise (synthetic images) are updated to enhance KD. Lastly, the global model is trained with both the original and noise-adapted client models via KD and synthetic images. This process is repeated till global model convergence. Extensive evaluation of this design on five small- and three large-scale medical image classification datasets reveals superior accuracy over prior methods. Code is available at https://github.com/myeongkyunkang/FedISCA.

Keywords: One-Shot Federated Learning · Knowledge Distillation · Noise · Image Synthesis · Client Model Adaptation.

1 Introduction

One-shot federated learning (FL) allows a global model to be trained through a single communication round without sharing data between clients [8,15,35,6,33].

⋆ Corresponding author.

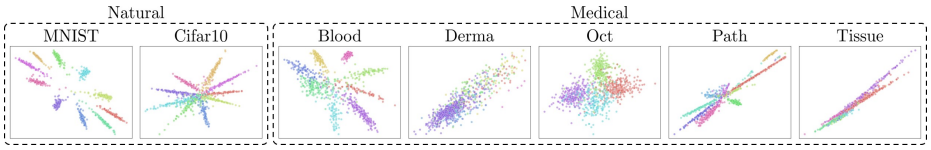

Fig. 1. Feature visualization on natural (MNIST and Cifar10) and medical (Blood, Derma, Oct, Path, and Tissue) images. For visualization, we placed a bottleneck layer before the class prediction layer, reducing the feature dimension to 2. Each color represents a classification label. Notably, the feature distribution in medical data is more complex.

This approach significantly reduces the risk of attack and communication costs compared to FL [21] and allows for decentralized training under extreme conditions. For instance, one-shot FL has emerged as a viable solution for reducing significant transmission costs in scenarios where patient data is only accessible within an isolated network requiring in-person transfer of client models. Since one-shot FL can only access clients' models once during training, recent one-shot FL suggests generating images and using them to transfer knowledge from multiple client models for global model training using knowledge distillation (KD) [33]. However, the lack of diversity in the generated images often leads to overfitting, posing a significant challenge for one-shot FL. To address this issue, [33,22] propose to enhance the transferability of client models by generating diverse natural images near the decision boundary. Compared to natural images, the decision boundaries in medical data are often more complex (*e.g.*, less discriminative as shown in Fig. 1), which limits the applicability of existing one-shot FL approaches to this application. Note, while the challenges in medical data and client heterogeneity can be mitigated through multiple communication rounds [23,12,36,18], the one-shot scenario presents a unique difficulty. Through this study, we reveal the inherent drawbacks of existing one-shot FL methods for medical data (see Table 1), and suggest a more suitable approach to address existing challenges *e.g.*, overfitting.

To prevent global model overfitting, we attempt to leverage random noise as a training source for KD (see Fig. 2). Baradad *et al.* [1] employs diverse types of structured noise for training in order to account for the difference between real images and random noise. However, due to the diversity of medical data [3,13,14], seeking a common noise space is more challenging than in natural images. Hence, we exploit DeepInversion [30], which synthesizes structured proxy noise specific to a task and thus ensures that generated noise matches the properties of medical data. Specifically, we first gather client models on the central server, where each client model is trained on its own dataset. Next, we synthesize images from random noise and store all intermediate samples in memory. Also, as images in the early stages of synthesis (*i.e.*, close to random noise) are different from real images, we design noise-adapted client models that employ adaptive batch normalization (AdaBN) [16]. AdaBN is based on the assumption that domain-related knowledge is represented by the statistics of the batch normalization

(BN) [11] and label-related knowledge is stored in the weight matrix of each layer, ultimately enhancing the KD signal for random noise. Lastly, we train a global model through KD with both the original- and noise-adapted client models using memory-stored images, repeating until global model convergences.

The contributions are as follows: (i) We propose one-shot FL leveraging image synthesis with client model adaptation. This allows to transfer knowledge from client models to the global model with synthesized images ranging from random noise to realistic images and contributes to preventing overfitting. (ii) We employ noise-adapted client models using AdaBN to produce a better KD signal for random noise. (iii) Comprehensive experiments on five small- and three large-scale medical image classification datasets consisting of microscopy, dermatoscopy, oct, histology, x-ray, and retinal images reveal that our method outperforms state-of-the-art one-shot FL methods.

Related Work. Due to the challenges of **one-shot FL**, prior methods were trained on public data [8,15], applying dataset distillation [35], or sharing additional information [6]. However, these assumptions may not hold for several real world scenarios, posing a challenge for their practical application. Recently, Zhang *et al.* [33] proposed the one-shot FL DENSE, which transfers knowledge from an ensemble of client models using KD and generated images. To enhance the transferability of client models, DENSE generates diverse images near the decision boundary to improve its accuracy. However, DENSE does not perform well in one-shot FL for medical data due to the complexity of decision boundaries. While DENSE diversifies generation using a generator, we propose to avoid overfitting by using synthesized images ranging from random noise to realistic images. For **data-free KD** [19], DeepInversion [30] synthesizes images by optimizing RGB pixels with cross-entropy and regularization losses and improves synthesis quality by minimizing feature statistics in BN layers. DAFL [2] uses a generator for image synthesis with a teacher model as a discriminator. To prevent student model overfitting, ZSKT [22] synthesizes images that exhibit mismatch between the student and teacher models. Unlike the methods that choose the best image as a training source for KD, our approach utilizes all intermediate synthesized images to prevent overfitting. Also, while Raikward *et al.* [24] proposed a method for KD that uses random noise as a training source, it requires real images during training and needs to adjust BN layer statistics multiple times iteratively. In contrast, our method performs one-shot FL without requiring real images during training.

2 Method

The overall training processes are shown in Fig. 2 and Algorithm 1. Given K client models $W^c = \{W_1^c, \ldots, W_k^c\}$ with corresponding BN statistics μ_k and σ_k^2 with respect to data D_k, the objective of FL is to train a global model W^g, which represents all data $D = \{D_1, \ldots, D_k\}$. Motivated by [17,34,33], KD enables the transfer of knowledge from client models W^c to the global model W^g. Due to restricted access of D, prior works [33,2,30] use synthetic images $\hat{x}$

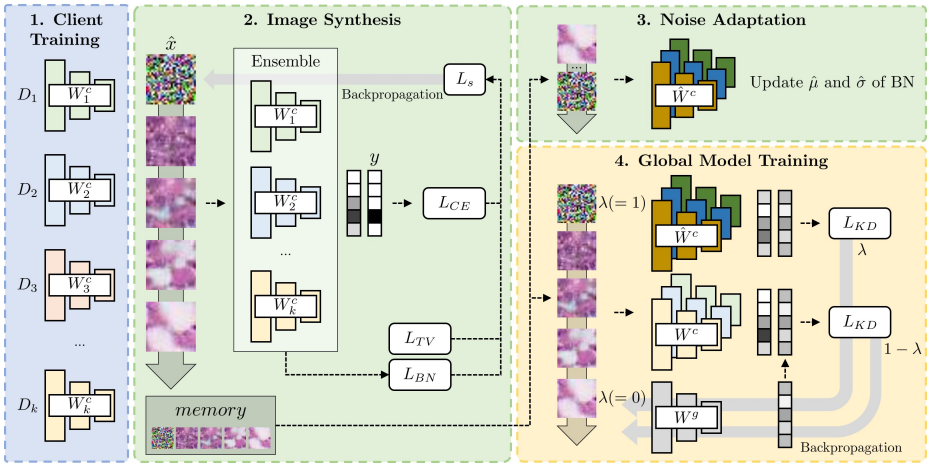

Fig. 2. Illustration of our proposed method. W_k^c denotes a client model with respect to data D_k and W^g denotes a global model. W^c denotes original client models and $\hat{W}^c$ denotes noise-adapted client models. $\hat{x}$ indicates random noise and λ indicates noise level. $\hat{x}$ is optimized to have a property of all D_k using L_{CE}, L_{BN}, and L_{TV}. Afterward, it is used as a training source for KD in global model training.

as a training source for KD. However, since $\hat{x}$ may be monotonous for robust training, overfitting is a significant challenge in one-shot FL. To address this, we employ random Gaussian noise $\mathcal{N}(0,1)$ as a training source for KD [1]. However, in contrast with [1], $\mathcal{N}(0,1)$ does not capture common medical properties. Hence, we employ DeepInversion [30] to ensure random noise retains characteristics of D. Details regarding image synthesis with DeepInversion are described in the following section.

Image Synthesis. Given random noise $\hat{x} \in \mathbb{R}^{H\times W\times C}$ initialized from $\mathcal{N}(0,1)$, where H, W, and C denote height, width, and channels; the objective of image synthesis is to ensure $\hat{x}$ possesses a certain property of D. To achieve this, we optimize RGB pixels of $\hat{x}$ to synthesize a class-conditioned image with respect to a specific label y for I iterations. Formally,

$$L_s(\hat{x}, y; W^c) = L_{CE}(\hat{x}, y; W^c) + \lambda_{BN} L_{BN}(\hat{x}; W^c) + \lambda_{TV} L_{TV}(\hat{x}; W^c), \quad (1)$$

where L_{CE}, L_{BN}, and L_{TV} are cross-entropy, BN, and total variation losses [20]. Hyper-parameters λ_{BN} and λ_{TV} are used to balance the losses. Cross-entropy loss enables the synthesis of an image with respect to the label y, and total variation loss encourages image synthesis consistency. Additionally, $L_{BN}(\hat{x}) = \sum(\|\mu(\hat{x}) - \mu\| + \|\sigma^2(\hat{x}) - \sigma^2\|)$, where $\mu(\hat{x})$ & $\sigma^2(\hat{x})$ are the batch-wise mean & variance features of $\hat{x}$ and μ & σ^2 of the stored statistics of the BN layer. Since BN enforces feature similarity at all levels, this improves the quality of image synthesis significantly.

Recall that our method employs random noise $\hat{x}$ that has D's characteristics for training. In contrast to DeepInversion which selects the best image as a

Algorithm 1 Training process of our proposed method.

Input: Client models W^c with corresponding μ and σ^2, a global model W^g, a iteration I, a learning rate of image synthesis η_s, a learning rate of KD η_d, a momentum α.
$\hat{W}^c \leftarrow W^c, \hat{\mu} \leftarrow \mu, \hat{\sigma}^2 \leftarrow \sigma^2$ `// Initialize noise-adapted client models`
Repeat
  Initialize a batch of random noise $\hat{x}$ and arbitrary labels y.
  $memory \leftarrow [\ ]$
  for $i = 1, \cdots, I$ **do**
    $\hat{x} \leftarrow \hat{x} - \eta_s \nabla L_s(\hat{x}, y; W^c)$ `// Synthesize image`
    $memory.append((\hat{x},\ 1 - i/I))$
  end for
  for $i = 1, \cdots, I$ **do**
    $\hat{x},\ \lambda \leftarrow memory[I - i]$
    $\hat{\mu} \leftarrow \alpha\hat{\mu} + (1-\alpha)\mu(\hat{x}),\ \hat{\sigma}^2 \leftarrow \alpha\hat{\sigma}^2 + (1-\alpha)\sigma^2(\hat{x})$ `// Adapt noise for` $\hat{W}^c$
  end for
  for $i = 1, \cdots, I$ **do**
    $\hat{x},\ \lambda \leftarrow memory[i]$
    $W^g \leftarrow W^g - \eta_d \nabla L_d(\hat{x}, \lambda; W^c, \hat{W}^c, W^g)$ `// Train global model`
  end for
until convergence.
Output: Trained global model W^g.

training source, our method employs all intermediate synthesized samples for KD. Thus we store all intermediate samples and the corresponding noise level λ (*e.g.*, $1 - i/I$ for i steps) in *memory* during I iterations. Due to the visual difference between $\mathcal{N}(0,1)$ and D, we design noise-adapted client models using AdaBN [16] to provide better KD signals for $\hat{x}$. The following section will describe more details regarding noise-adapted client models.

Noise Adaptation. BN [11] was proposed to mitigate internal covariate shifts, allowing to provide consistent input distributions to subsequent layers. Due to the existing discrepancy between $\mathcal{N}(0,1)$ and D, there is no guarantee BN will provide consistent input to subsequent parameters and may lead to poor model predictions. Thus we adapt $\mathcal{N}(0,1)$ by iteratively adjusting the running statistics of BN using AdaBN [16], producing better logit signals for KD. Formally,

$$\hat{\mu} = \alpha\hat{\mu} + (1-\alpha)\mu(\hat{x}),\ \hat{\sigma}^2 = \alpha\hat{\sigma}^2 + (1-\alpha)\sigma^2(\hat{x}), \tag{2}$$

where α represents momentum and $\hat{x}$ is a sample stored in *memory*. Initially, $\hat{\mu}$ and $\hat{\sigma}^2$ are set to μ and σ^2. The samples in *memory* ranging from characteristic images for D to $\mathcal{N}(0,1)$ by gradually adjusting $\hat{\mu}$ and $\hat{\sigma}^2$ towards $\mathcal{N}(0,1)$ through Eq. 2 for I steps. With this in mind, we now describe how to train the global model.

Global Model Training. KD allows to train a global model with multiple client models [33,17,34]. We denote W^c with original μ and σ^2 as W^c, and denote W^c with $\hat{\mu}$ and $\hat{\sigma}^2$ as $\hat{W}^c$. Since $\hat{x}$, W^c, and $\hat{W}^c$ are used for KD, this enables the model to avoid overfitting without being negatively impacted during global model training. Formally,

$$L_d(\hat{x}, \lambda; W^c, \hat{W}^c, W^g) = \lambda L_{KD}(\hat{x}; \hat{W}^c, W^g) + (1-\lambda) L_{KD}(\hat{x}; W^c, W^g), \tag{3}$$

where λ denotes a noise level stored in *memory*. $L_{KD}(\hat{x}; W^c, W^g)$ denotes the Kullback-Leibler divergence between $p(\hat{x}; W^c)$ and $p(\hat{x}; W^g)$ where $p(\cdot)$ is an

ensemble (averaging) prediction of given models with a temperature on softmax inputs [10]. Overall, W^g is trained for I steps. To clarify, random noise contributes to avoiding overfitting, while noise-adapted client models help to produce a better KD signal for random noise, improving robust global model training. These processes *i.e.*, Image Synthesis, Noise Adaptation, and Global Model Training are repeated until the global model W^g converges.

3 Experiments

Datasets. For evaluation, we use five small-scale (28×28) medical image classification datasets *i.e.*, Blood, Derma, Oct, Path, and Tissue from MedMNIST [29]. Additionally, we use three large-scale (224×224) datasets *i.e.*, RSNA, Diabetic, and ISIC from RSNA Pneumonia Detection [25], Diabetic Retinopathy Detection [7], and ISIC2019-HAM-BCN20000 [4,28,5].

Experimental Settings. We explore three scenarios *i.e.*, (i) data heterogeneity levels, (ii) impact on large-scale datasets, and (iii) model heterogeneity *i.e.*, each client has different architectures. In (i), Blood, Derma, Oct, Path, and Tissue datasets are used with Independent and Identically Distributed (IID) clients and Dirichlet distributed [31] clients with $\alpha = 0.6$ and $\alpha = 0.3$. For (ii), RSNA, Diabetic, and ISIC datasets are used with IID clients, including ISIC′ where each client has a different image acquisition system [27]. For (iii), client models used either ResNet18 [9], ResNet34 [9], WRN-16-2 [32], VGG16(with BN) [26], and VGG8(with BN) [26], respectively.

Comparison Methods. We employ three one-shot FL methods: FedAvg [21] with single communication, DAFL [2], and DENSE [33], each evaluated using global model accuracy obtained on test data. For the upper bound, we report the FedAvg with 100 communications. For ablations, we evaluate (a) without image synthesis (w/o IS), (b) without image synthesis and noise adaptation (w/o IS&Ada) with only $\mathcal{N}(0, 1)$ used for training, (c) without noise adaptation (w/o Ada), and (d) without intermediate random noise (w/o $\mathcal{N}$), this is equivalent to DeepInversion [30] in a one-shot FL scenario. For w/o $\mathcal{N}$, we synthesize all images and perform KD. For a fair comparison, we follow each method's original implementation and matched all training/parameter settings. For DAFL, an ensemble of client models was used as the teacher model following [33,17,34] with KD used for global model training. On large-scale datasets, an ImageNet pre-trained model was used with balanced classification accuracy reported for evaluation as in [27].

Implementation Details. We used ResNet18 [9] for our experiments with five clients by default. Client models were trained for 100 epochs with SGD optimizer using learning rate (LR) 1e-3 and batch size 128. For image synthesis, we used Adam optimizer with LR 5e-2 for 100 epochs with 500 and 1,000 synthesis iterations (*i.e.*, I) for small- and large-scale datasets, with batch sizes 256 and 50, respectively. Following [30], $\lambda_{TV} = 0.000025$ and $\lambda_{BN} = 10$, with KD temperature $T = 20$ and momentum $\alpha = 0.9$.

Table 1. Classification accuracy on five datasets with different heterogeneity levels. The first and second sub-rows show the accuracy of the upper bound and one-shot FL methods. The third sub-row shows ablation performance with IS, Ada, and $\mathcal{N}$ denoting w/o image synthesis, noise adaptation, and random noise, respectively. **Bold** indicates the best accuracy among one-shot FL methods.

	IID					Dirichlet ($\alpha = 0.6$)					Dirichlet ($\alpha = 0.3$)				
	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue
FedAvg[21]	93.51	74.61	75.60	84.54	63.64	93.60	72.72	76.50	81.48	55.61	87.49	69.88	73.50	77.52	53.26
FedAvg(1)	13.74	66.88	25.00	5.86	32.07	18.24	66.88	25.00	5.86	32.07	16.92	10.97	25.00	5.86	32.07
DAFL[2]	7.13	66.43	25.00	7.63	11.55	7.13	66.88	34.40	14.97	39.15	7.13	13.62	25.00	18.64	45.00
DENSE[33]	39.37	66.93	33.80	21.89	21.35	34.52	67.78	39.40	30.31	9.47	30.78	12.77	25.80	19.87	9.33
FedISCA	**87.99**	**70.12**	70.20	**84.18**	**61.90**	**82.90**	**69.83**	**68.60**	**82.92**	**53.04**	**46.59**	**15.91**	**60.50**	**79.25**	**51.00**
w/o IS	9.09	66.88	25.20	24.69	23.70	9.09	66.88	26.10	22.41	9.31	23.27	11.02	27.10	18.70	9.31
w/o IS&Ada	7.13	11.12	35.80	4.72	7.13	7.13	66.88	25.00	14.15	32.07	7.13	11.12	25.00	4.72	7.13
w/o Ada	81.61	68.33	**70.30**	82.08	59.34	63.67	68.18	61.90	78.61	51.99	29.73	14.36	54.30	77.69	50.40
w/o $\mathcal{N}$[30]	87.02	68.73	60.20	77.90	57.86	80.62	69.58	60.30	75.54	49.06	45.69	13.87	49.20	70.53	46.73

Table 2. Balanced classification accuracy on large-scale datasets.

	FedAvg[21]	FedAvg(1)	DAFL[2]	DENSE[33]	FedISCA	w/o IS	w/o IS&Ada	w/o Ada	w/o $\mathcal{N}$[30]
RSNA	88.16	78.65	50.55	55.06	**85.34**	50.00	50.00	81.56	50.61
Diabetic	49.04	35.60	22.63	23.51	40.08	20.07	20.02	**40.91**	28.30
ISIC	62.88	38.05	14.51	13.69	**48.39**	12.50	12.50	47.21	25.61
ISIC′	57.15	18.08	18.37	16.46	**22.47**	11.29	12.52	21.72	14.80

3.1 Main Results

Table 1 shows the accuracy on five datasets with different heterogeneity levels. FedISCA outperforms all one-shot FL methods across all datasets regardless of the level of heterogeneity. In Table 2, FedISCA also reports improved performance against the compared methods, validating the viability of our approach on real-world large-scale data. On the contrary, DAFL and DENSE performed poorly on medical data since significant accuracy gaps exist between the upper bound and each competitor (except Derma). Additionally, though FedAvg reports higher accuracy for multiple communication rounds, it shows significantly lower accuracy for single communication (FedAvg(1)). To better explain this phenomenon, we analyzed the accuracy of FedAvg(1) by comparing the variance between client model parameters *i.e.*, client models with high variance *e.g.*, Path IID(=36.10), yield lower accuracy compared to those with low variance *e.g.*, Derma IID(=0.01). This suggests that the variance of client models is correlated with the accuracy of FedAvg(1).

In Fig. 3, we show the synthesized images of FedISCA, DAFL [2], and DENSE [33] on eight datasets. FedISCA generates more realistic images compared to the competitors. Note that DENSE aims to generate a diverse image (*e.g.*, generating highly transferable samples) distributed near the decision boundary, which may not be realistic. Although these methods have achieved higher accuracy on natural data, our experiments reveal that this assumption does not hold in the medical domain. In addition, DENSE outperforms FedAvg(1) on small-scale datasets (except Tissue), but its accuracy is lower than FedAvg(1) on large-scale datasets. This suggests a difficulty in large-scale image generation *i.e.*, the gen-

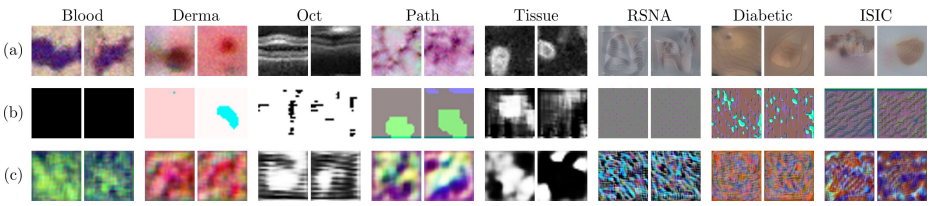

Fig. 3. The synthesized images of (a) FedISCA, (b) DAFL [2], and (c) DENSE [33] on eight datasets. Overall, FedISCA synthesizes more realistic images.

Table 3. Classification accuracy on five datasets with model heterogeneity.

	IID					Dirichlet ($\alpha = 0.6$)					Dirichlet ($\alpha = 0.3$)				
	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue
DAFL[2]	7.13	65.69	25.00	15.72	35.66	7.13	67.23	37.10	28.15	39.54	7.13	13.47	45.30	29.68	19.54
DENSE[33]	46.86	66.88	44.00	33.08	38.28	23.47	67.93	40.70	28.68	36.70	34.67	13.42	44.00	39.37	38.37
FedISCA	**87.96**	**71.17**	**70.00**	**83.02**	**61.74**	**73.43**	**69.23**	**64.80**	**82.73**	**51.95**	**44.20**	**16.61**	**62.00**	**72.26**	**43.80**
w/o $\mathcal{N}$[30]	87.55	69.93	51.20	74.05	57.90	68.78	68.93	61.00	72.79	46.91	43.85	15.61	51.50	64.65	39.89

erator in DENSE deteriorates global model training and leads to lower accuracy, while FedAvg(1) achieves high accuracy due to the low client model variance *e.g.*, RSNA(=0.61), Diabetic(=0.04), and ISIC(=0.09).

Ablations. We report ablation results in Table 1 and 2. In the medical field, generating realistic images is crucial for one-shot FL, as the accuracy of w/o IS, and w/o IS&Ada is significantly lower compared to FedISCA; this validates the need for image synthesis. However, relying on image synthesis alone is not enough to achieve high accuracy, as neither w/o Ada nor w/o $\mathcal{N}$ achieve the best accuracy across all datasets. w/o $\mathcal{N}$ performs worse than w/o Ada in most datasets (except Blood and Derma), showing that solely relying on the best image is not sufficient for robust training. On the contrary, the accuracy of FedISCA suggests that noise-adapted client models alleviate the negative effects of random noise, resulting in high accuracy. Overall, the experimental results support the idea that both components play an essential role in medical one-shot FL. Additionally, we also evaluate the variance in BN statistics between the original and noise-adapted client models. Here, we found that high variance (*e.g.*, RSNA(=0.0018)), yields improved accuracy compared to those with lower variance (*e.g.*, Diabetic(=0.0008)). Finally, Table 3 shows the accuracy of a global model trained on client models with model heterogeneity. The proposed method reports the best accuracy among all competitors, equally demonstrating the effectiveness of our method in one-shot FL with diverse types of model architectures.

4 Conclusion

We present a novel one-shot FL framework that uses image synthesis and client model adaptation with KD. We demonstrate that (i) random noise significantly

reduces the risk of overfitting, resulting in robust global model training; (ii) noise-adapted client models enhance the KD signal leading to high accuracy; and (iii) through experiments on eight datasets, our method outperforms the state-of-the-art one-shot FL methods on medical data. Further investigation into severe heterogeneity in clients will be a topic of future research.

Acknowledgments. This work was supported by funding from the DGIST R&D program of the Ministry of Science and ICT of KOREA (22-KUJoint-02) and the framework of international cooperation program managed by the National Research Foundation of Korea (NRF-2022K2A9A1A01097840) and the NRF grant funded by the Korean Government (MSIT)(No. 2019R1C1C1008727) and the National Institute of Health (MH113406, DA057567, AA021697) and by the Stanford HAI Google Cloud Credit.

References

1. Baradad Jurjo, M., Wulff, J., Wang, T., Isola, P., Torralba, A.: Learning to see by looking at noise. Advances in Neural Information Processing Systems **34**, 2556–2569 (2021)
2. Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q.: Data-free learning of student networks. In: International Conference on Computer Vision. pp. 3514–3522 (2019)
3. Chikontwe, P., Nam, S.J., Go, H., Kim, M., Sung, H.J., Park, S.H.: Feature recalibration based multiple instance learning for whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 420–430. Springer (2022)
4. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
5. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
6. Dennis, D.K., Li, T., Smith, V.: Heterogeneity for the win: One-shot federated clustering. In: International Conference on Machine Learning. pp. 2611–2620. PMLR (2021)
7. EyePACS: Diabetic retinopathy detection (2015)
8. Guha, N., Talwalkar, A., Smith, V.: One-shot federated learning. arXiv preprint arXiv:1902.11175 (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778 (2016)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. PMLR (2015)

12. Jiang, M., Yang, H., Li, X., Liu, Q., Heng, P.A., Dou, Q.: Dynamic bank learning for semi-supervised federated image diagnosis with class imbalance. In: Medical Image Computing and Computer Assisted Intervention. pp. 196–206. Springer (2022)
13. Jung, E., Luna, M., Park, S.H.: Conditional gan with 3d discriminator for mri generation of alzheimer's disease progression. Pattern Recognition **133**, 109061 (2023)
14. Kim, S., An, S., Chikontwe, P., Park, S.H.: Bidirectional rnn-based few shot learning for 3d medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1808–1816 (2021)
15. Li, Q., He, B., Song, D.: Practical one-shot federated learning for cross-silo setting. International Joint Conference on Artificial Intelligence (2020)
16. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Adaptive batch normalization for practical domain adaptation. Pattern Recognition **80**, 109–117 (2018)
17. Lin, T., Kong, L., Stich, S.U., Jaggi, M.: Ensemble distillation for robust model fusion in federated learning. Advances in Neural Information Processing Systems **33**, 2351–2363 (2020)
18. Liu, X., Li, W., Yuan, Y.: Intervention & interaction federated abnormality detection with noisy clients. In: Medical Image Computing and Computer Assisted Intervention. pp. 309–319. Springer (2022)
19. Liu, Y., Zhang, W., Wang, J., Wang, J.: Data-free knowledge transfer: A survey. arXiv preprint arXiv:2112.15278 (2021)
20. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Computer Vision and Pattern Recognition. pp. 5188–5196 (2015)
21. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282. PMLR (2017)
22. Micaelli, P., Storkey, A.J.: Zero-shot knowledge transfer via adversarial belief matching. Advances in Neural Information Processing Systems **32** (2019)
23. Qi, X., Yang, G., He, Y., Liu, W., Islam, A., Li, S.: Contrastive re-localization and history distillation in federated cmr segmentation. In: Medical Image Computing and Computer Assisted Intervention. pp. 256–265. Springer (2022)
24. Raikwar, P., Mishra, D.: Discovering and overcoming limitations of noise-engineered data-free knowledge distillation. In: Advances in Neural Information Processing Systems (2022)
25. RSNA: Rsna pneumonia detection challenge (2018)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations (2014)
27. Ogier du Terrail, J., Ayed, S.S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., Muzellec, B., Philippenko, C., Silva, S., Teleńczuk, M., Albarqouni, S., Avestimehr, S., Bellet, A., Dieuleveut, A., Jaggi, M., Karimireddy, S.P., Lorenzi, M., Neglia, G., Tommasi, M., Andreux, M.: Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In: Advances in Neural Information Processing Systems (2022)
28. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
29. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. In: arXiv preprint arXiv:2110.14795 (2021)

30. Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J.: Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Computer Vision and Pattern Recognition. pp. 8715–8724 (2020)
31. Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y.: Bayesian nonparametric federated learning of neural networks. In: International Conference on Machine Learning. pp. 7252–7261. PMLR (2019)
32. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: British Machine Vision Conference (BMVC) (2016)
33. Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., Wu, C.: Dense: Data-free one-shot federated learning. In: Advances in Neural Information Processing Systems (2022)
34. Zhang, S., Liu, M., Yan, J.: The diversified ensemble neural network. Advances in Neural Information Processing Systems **33**, 16001–16011 (2020)
35. Zhou, Y., Pu, G., Ma, X., Li, X., Wu, D.: Distilled one-shot federated learning. arXiv preprint arXiv:2009.07999 (2020)
36. Zhu, W., Luo, J.: Federated medical image analysis with virtual sample synthesis. In: Medical Image Computing and Computer Assisted Intervention. pp. 728–738. Springer (2022)