Attentive Modeling and Distillation for Out-of-Distribution Generalization of Federated Learning

Zhuang Qi†[1], Weihao He†[1], Xiangxu Meng[1] and Lei Meng*[1,2]
[1] School of Software, Shandong University, Jinan, China
[2] Shandong Research Institute of Industrial Technology, Jinan, China
Email: z_qi@mail.sdu.edu.cn, rogerhe130@gmail.com, {mxx, lmeng}@sdu.edu.cn

***Abstract*—Out-of-distribution issues lead to different optimization directions between clients, which weakens collaborative modeling in federated learning. Existing methods aims to decouple invariant features in the latent space to mitigate attribute bias. However, their performance is limited by suboptimal decoupling capabilities in complex latent spaces. To address this problem, this paper presents a method, termed FedAKD, that adaptively identifies meaningful visual regions in images to guide the model in learning causal features. It includes two main modules, where the attentive modeling module adaptively locates critical regions to mitigate the negative impact of irrelevant elements, which are considered significant contributors to distribution heterogeneity. The attention-guided representation learning module leverages attentive knowledge to guide the local model to pay more attention to important regions, which acts as a soft attention regularizer to mitigate the trade-off between capturing category-relevant information and irrelevant contextual information in images. Experiments were conducted on four datasets, including performance comparison, ablation study, and case study. The results demonstrate that FedAKD can effectively enhance attention to causal features, which leads to superior performance compared with the state-of-the-art methods.**

***Index Terms*—Federated Learning, Knowledge Distillation, Out-of-Distribution, Attentive modeling**

I. Introduction

Federated learning has emerged as a promising distributed learning paradigm, enabling collaborative modeling with multiple data sources while preserving data privacy [1], garnering widespread attention across various fields [2]–[5]. It aggregates the parameters of local models trained on private data from multiple devices to obtain a global model, without involving data sharing [6]–[8]. Despite the advantages in preserving privacy that federated learning presents, it continues to confront significant challenges associated with data heterogeneity, such as out-of-distribution [1], [9], [10]. The huge attributes skew between data sources often harm the effectiveness of collaboration. This is primarily due to the disparate optimization directions among various local models, and the local model presents declining performance in other clients.

To mitigate the out-of-distribution issue, existing methods can roughly be divided into two groups: regularization-based

† indicates equal contributions.
* indicates corresponding author.

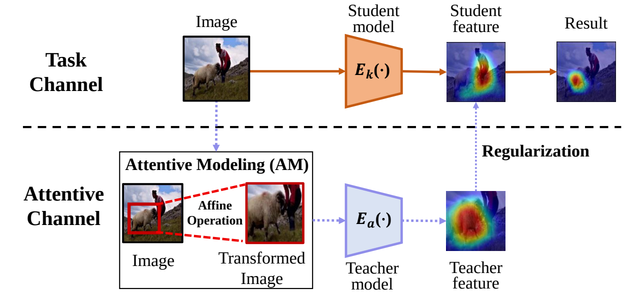

Fig. 1. FedAKD may learn better visual representation. It extracts attentive features from the meaningful regions segmented by the AM module, which provides the instructive knowledge for the task channel. This can reduce the interference of contextual noise.

representation alignment [11], [12] and representation decoupling of invariant attributes [13]–[16]. The former approaches aim to facilitate the learning of consistent knowledge between clients, to mitigate the adverse effects of inconsistent attributes on representation learning. They typically employ prototype-based representation alignment regularization to constrain the local training of clients. For example, FPL constructs unbiased prototypes and employs consistency regularization to align instances with the corresponding unbiased prototypes, which helps to alleviate feature heterogeneity between clients [12]. The latter methods focus on decoupling in the feature space to extract the invariant features, which contributes to the elimination of intervention from irrelevant contextual. For instance, DFL disentangles domain-specific and invariant attributes into two complementary branches, separating domain-specific attributes from model aggregation [14]. These methods offer insights into the efficacy of isolating domain-specific attributes locally to mitigate out-of-distribution issues. However, their performance is hampered by the limited ability to uncover causal relationships.

To address this issue, this paper presents a novel attentive knowledge distillation mechanism, termed FedAKD. As illustrated in Figure 1, compared with conventional methods, the proposed FedAKD effectively identifies the meaningful region in the image, which reduces the interference of irrelevant contextual noise. Specifically, FedAKD has two main modules, the attentive modeling (AM) module and the attention-guided representation learning (AGRL) module. To precisely identify key features and patterns of images, the AM module utilizes self-learned geometric operators to adaptively locate image regions relevant to the class, which provides key guidance information for the model to learn causal features. Subsequently, the AGRL module serves as a soft-attention regularizer to the local model in the task channel by aligning its visual features to the attentive features produced in the attentive channel, which effectively filters out irrelevant distractions. Notably, the AM module, as a plug-and-play component, can be easily integrated into various methods. As observed, FedAKD enhances the generalization across different domains for models.

Extensive experiments are conducted on four datasets in terms of performance comparison, ablation study of the key components, and case study for the effectiveness of key region extraction and recognition. The results verify that modeling meaningful attention in the input space can guide models to learn more robust features, which enhances their generalization capabilities. To summarize, this paper includes two main contributions:

- This paper presents a model-agnostic attentive knowledge distillation mechanism, termed FedAKD. To the best of our knowledge, it is the first method that modeling meaningful attention in the input space to alleviate the attributes skew problem between clients in federated leaning.
- We propose a plug-and-play module, named the AM module, which can be easily integrated into various methods to enhance their performance, significantly improving the quality of representation learning.

II. Related Work

A. Federated Learning with Non-IID Data

To tackle data heterogeneity in federated learning, commonly used methods are generally divided into two groups: one focuses on reducing local biases, while the other aims to improve aggregation efficiency. The former employs regularization or cross-training to help clients acquire comprehensive knowledge. Regularization comes in three types: weight-based [17], [18], feature-based [19], [20], and prediction-based [21], [22]. The latter approach believes that directly averaging local model parameters can harm performance. Instead, they either design better aggregation methods or fine-tune models on the server. For example, Elastic aggregation adaptively combines client models based on how parameter variations affect prediction outputs. Meanwhile, FedFTG [23] generates server-based samples to refine the global model.

B. Out-of-distribution in Federated Learning

Out-of-distribution problems, such as substantial attribute skews between local datasets, often lead to a decline in the performance of federated learning systems. Numerous studies have concentrated on alleviating these challenges [17], [19], [24], [25]. They use regularization to align representations from different sources, learning consistent representations in various contexts, or they decouples invariant features in the latent space to reduce the interference of irrelevant attributes. For instance, FedProc and FPL utilizes prototypes to regulate the representation learning across all clients [12], [26]. FCCL utilizes unlabeled public data to facilitate joint cross-correlation learning, with the goal of maximizing output similarity for identical categories across diverse domains while minimizing output redundancy [27]. DFL decomposes domain-specific attributes and invariant attributes into two complementary branches, which avoids the negative impact of domain-specific information on the aggregation [14]. Despite some achievements, methods based on decoupling in latent space may underperform due to suboptimal causal relationship identification.

III. Methodology

A. Overall Framework

The proposed attentive knowledge distillation scheme in federated learning (FedAKD), outlined in Figure 2, has two main phases: local training and global aggregation. FedAKD places particular emphasis on the local training within its dual-channel framework, where the Attentive Channel utilizes the attentive modeling (AM) module to dynamically extract class-aware region to provide the fine-grained knowledge. The Task Channel employs the attention-guided representation learning (AGRL) module to fully leverages attentive knowledge to help the model optimize the feature extraction and recognition process. Subsequently, FedAKD aggregates all local models to generate the global model in the server.

B. Attentive Modeling

The Attention Modeling (AM) module aims to transform the original image I into one that is more focused on the meaningful object I_{trans}, i.e., $I_{trans} = \mathrm{AM(I)}$. It can provide instructive knowledge to the local student model to mitigate the interference from backgrounds. An intuitive idea is to crop out the main object from the image. Inspired by Spatial Transformer Network (STN) [28], the AM module adaptively locates visual attentive regions based on classification loss. Specifically, it involves a localization network $F_{loc}(\cdot)$, a grid generator $G(\cdot,\cdot)$, and a sampler.

The localization network generates an affine transformation matrix θ for each image I to capture detailed regions in the original image. It can be formulated as follows:

$$\theta = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} = F_{loc}(I) \tag{1}$$

where $\{\theta_{11}, \theta_{22}\}$, $\{\theta_{12}\}$, $\{\theta_{21}\}$ and $\{\theta_{13}, \theta_{23}\}$ represent scaling, rotation, shearing and translation parameters, respectively. $F_{loc}(\cdot)$ is typically a lightweight network.

The grid generator $G(\cdot,\cdot)$ utilizes the affine transformation operator $\mathcal{T}_\theta$ and transformation parameter θ to generate a transformed coordinate for image I, defined as

$$\begin{bmatrix} x_j^s \\ y_j^s \end{bmatrix} = G(\mathcal{T}_\theta, x) = \mathcal{T}_\theta(x) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \end{bmatrix} \begin{bmatrix} x_j^t \\ y_j^t \\ 1 \end{bmatrix} \tag{2}$$

where (x_j^s,y_j^s) and (x_j^t,y_j^t) denote coordinates of each pixel in the input image I and transformed image I_{trans}. j denotes the index of pixel.

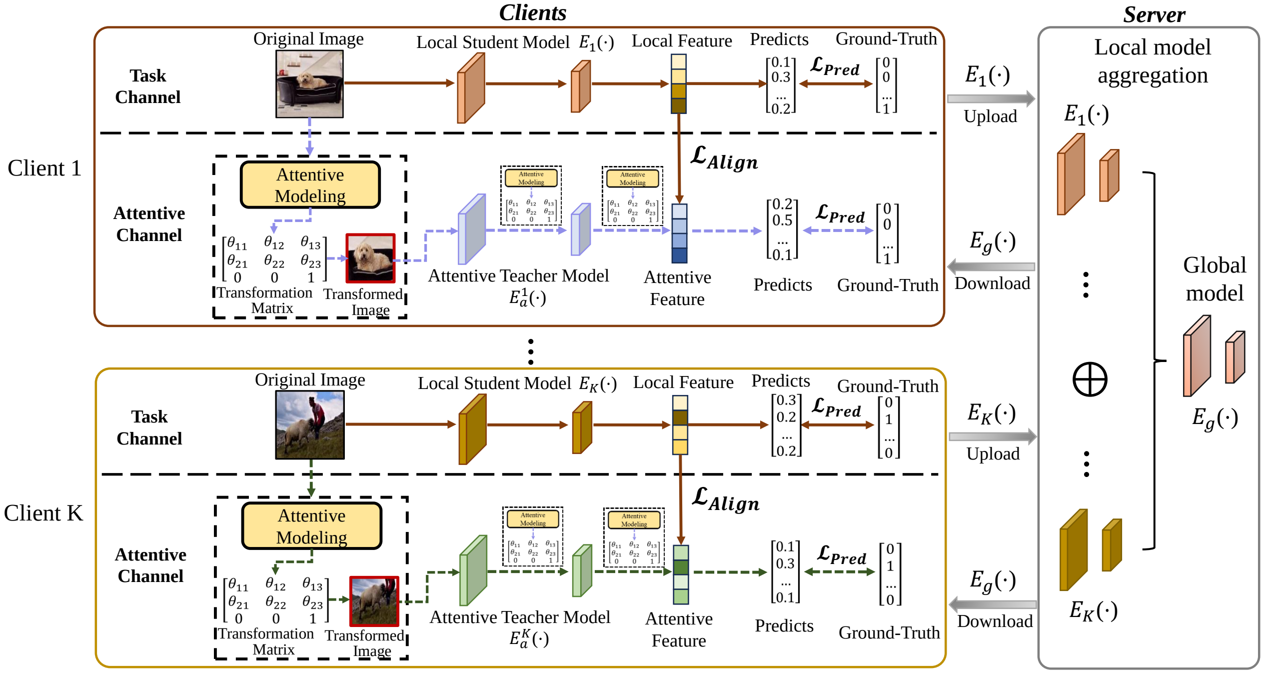

Fig. 2. Illustration of the framework of FedAKD. It utilizes self-learned geometric operators in the attentive channel to locate regions related to categories, which can provide guidance information for the student model to learn causal features. Subsequently, FedAKD aligns the original features in the task channel to the attentive features in the attentive channel.

The sampler extracts pixels from the input image I, and produces the output image I_{trans} through bilinear interpolation, enabling differentiable spatial transformations, defined by

$$I_{trans}(x_j^t, y_j^t) = \sum_n \sum_m I(n,m) \cdot \max(0, 1 - |x_j^s - n|) \cdot \max(0, 1 - |y_j^s - m|) \quad (3)$$

This shows that bilinear interpolation calculates the output pixel value $I_{trans}(x_j^t, y_j^t)$ at the transformed coordinates by taking a weighted average of the input image pixel values $I(n, m)$, where the weights decrease with the distance from the input image pixels to the transformed coordinate point.

Meanwhile, the classification loss is used to optimize the model of the attentive channel, i.e.,

$$\mathcal{L}_{cls}^{att} = \mathcal{L}_{CE}(\hat{y}_{trans}, y) \quad (4)$$

where $\hat{y}_{trans} = F_a(E_a(\text{AM}(I)))$ is the prediction, F_a and E_a are the classifier and the feature extractor in the attentive channel. $\mathcal{L}_{CE}$ denotes the cross-entropy loss. y is the label of the original image I and the transformed image I_{trans}.

C. Attention-Guided Representation Learning

The Attention-Guided Representation Learning (AGRL) module aims to use the attentive knowledge from the output of AM module as a soft regularizer to guide the training of the local student model. It effectively guides the student model to focus on key information by utilizing features from the most task-relevant regions in the image. Specifically, the AGRL module aligns the visual features output by the student model in the task channel with the fine-grained features generated by the teacher model in the attentive channel. To achieve this, the AGRL module utilizes KL Divergence (Kullback-Leibler Divergence) [29] as a measure to encourages the student model to adapt its feature representation to align more closely with that of the teacher model. Therefore, the alignment loss function can be defined as

$$\mathcal{L}_{align} = D_{KL}(f_t || f_a) \quad (5)$$

where $f_t = E_t(I)$ and $f_a = E_a(I_{trans})$ denote the feature output of local student model and the teacher model, respectively. E_t is a feature extractor in the task channel.

Subsequently, we use the empirical classification loss $\mathcal{L}_{cls}^{task}$ to optimize the local student model, i.e.,

$$\mathcal{L}_{cls}^{task} = \mathcal{L}_{CE}(F_t(f_t), y) \quad (6)$$

where F_t denotes the classifier in the task channel.

D. Training Strategy of FedAKD

FedAKD focuses on optimizing the extraction of meaningful object regions in the attentive channel, its corresponding optimization objective is

$$\mathcal{L}_{att} = E_{(x,y)\sim D_{local}}[\mathcal{L}_{cls}^{att}] \quad (7)$$

where x and y denote the original image and the corresponding label in the local dataset D_{local}. Meanwhile, FedAKD aims to enhance the local student model's ability to focus on important objects within the task channel, the corresponding objective is to minimize:

$$\mathcal{L}_{task} = E_{(x,y)\sim D_{local}}[\mathcal{L}_{cls}^{task} + \lambda_{align} \cdot \mathcal{L}_{align}] \quad (8)$$

where λ_{align} is a weight parameter.

TABLE I
PERFORMANCE COMPARISON BETWEEN FEDAKD WITH BASELINES ON CIFAR100, CIFAR100, COLORMNIST AND NICO-ANIMAL. ALL METHODS WERE EXECUTED ACROSS THREE TRIALS, WITH BOTH THE MEAN AND STANDARD DEVIATION BEING REPORTED.

Methods	**NICO-Animal**	**COLORMNIST**		**CIFAR10**		**CIFAR100**	
		β=0.1	β=0.5	β=0.1	β=0.5	β=0.1	β=0.5
FedAvg (AISTATS'17) [30]	29.84±0.6	57.48±0.4	88.15±0.6	74.29±0.5	83.15±0.2	50.13±0.7	52.86±0.4
Fedprox (MLSys'20) [25]	30.12±0.9	56.93±0.6	89.23±0.4	74.50±0.3	83.96±0.7	51.02±0.9	53.81±0.4
MOON (CVPR'21) [19]	31.97±0.4	58.42±0.5	90.12±0.8	74.77±0.2	84.65±0.4	50.78±0.3	53.75±0.3
FedNTD (NeurIPS'22) [31]	29.08±1.0	57.04±0.5	89.92±0.1	73.88±0.6	83.08±0.5	49.23±0.3	52.01±0.4
Fedproc (FGCS'23) [26]	31.29±0.8	58.15±0.4	88.98±0.5	75.49±0.9	84.83±0.6	51.46±0.2	54.12±0.6
FPL (CVPR'23) [12]	32.51±0.5	58.29±0.7	89.71±0.3	75.16±0.8	85.11±0.9	52.42±0.6	53.77±0.2
DaFKD (CVPR'23) [13]	32.14±0.9	58.22±0.8	89.95±0.5	75.05±0.9	85.40±0.6	52.36±0.6	54.74±0.3
FedIIR (ICML'23) [32]	33.05±0.9	58.10±0.9	90.23±0.8	75.07±0.9	85.33±0.7	51.98±0.2	54.32±0.7
FedAKD$_{\mathrm{FedAvg}}$	35.56±0.4	59.03±0.2	**92.75±0.3**	76.98±0.5	86.67±0.3	53.78±0.2	57.21±0.8
FedAKD$_{\mathrm{MOON}}$	**38.72±0.3**	**61.10±0.9**	92.68±0.3	**78.23±0.8**	**87.85±0.2**	**54.34±0.7**	**58.82±0.2**

IV. EXPERIMENTS

A. Experiment Settings

1) Datasets: We validate the efficacy of the proposed framework through experiments conducted on two out-of-distribution generalization datasets, namely COLORMNIST [33] and NICO-Animal [34]. Additionally, we assess its performance on two well-established datasets commonly employed in federated learning, CIFAR10 and CIFAR100 [35]. The statistical details of these datasets are summarized in Table II.

TABLE II
STATISTICS OF CIFAR10, CIFAR100, COLORMNIST AND NICO-ANIMAL DATASETS USED IN EXPERIMENTS.

Datasets	**#Class**	**#Training**	**#Testing**
CIFAR10	10	50000	10000
CIFAR100	100	50000	10000
NICO-Animal	10	10633	2443
COlORMNIST	10	60000	10000

2) Network Architecture: For a fair comparison, all methods share a common network architecture. And in the FedAKD, we maintain consistent architecture for both the student model in the task channel and the teacher model in the attention channel. For COLORMNIST, the architecture involves a convolutional layer serving as an image encoder and a 2-layer MLP as the classifier. Following previous works [19], [36], we employ ResNet-18 [37] as the network backbone for all other datasets. Notably, we adapt the first convolutional kernel size from 7 to 3 for CIFAR10 and CIFAR100, while keeping it at 7 for the NICO-Animal dataset. For the the model used in attentive modeling, it involves two convolutional layers for localisation and a fully connected layer to generate transformation parameter.

3) Hyper-parameter Settings: For all methods, we maintain consistency in hyperparameter settings across experiments. The local training epoch is fixed at 10 for each global round, with the number of clients set to 10 for CIFAR10, CIFAR100, COLORMNIST, and 7 for NICO-Animal, along with a sample fraction of 1.0. The local optimizer employed is the SGD algorithm, and the communication round is set to 100. During local training, we configure the weight decay to 1e-05 and the batch size to 64. The learning rate is initialized at 0.01, and the Dirichlet parameter β is set to 0.1 and 0.5 for CIFAR10, CIFAR100 and COLORMNIST. Furthermore, λ_{align} is fine-tuned from the set $\{0.01, 0.05, 0.1, 0.5\}$. The remaining hyperparameters follow the specifications outlined in the corresponding paper.

B. Performance Comparison

We compare FedAKD with eight SOTA methods, including FedAvg [30], MOON [19], Fedprox [25], Fedproc [26], FedNTD [31], FPL [12], DaFKD [13] and FedIIR [32]. These methods typically use FedAvg as the base algorithm, so we integrate key modules into FedAvg to form FedAKD$_{\mathrm{FedAvg}}$. Additionally, we incorporate the proposed AM module into the global branch of MOON, forming FedAKD$_{\mathrm{MOON}}$. The following results can be derived from Table I.

- **Both FedAKD$_{\mathrm{FedAvg}}$ and FedAKD$_{\mathrm{MOON}}$ demonstrate significant improvements in classification accuracy over their respective baseline models.** This highlights the model-agnostic nature of the FedAKD approach.
- **FedAKD consistently outperforms other methods in terms of classification accuracy.** This is understandable since FedAKD can capture meaningful objects in images while mitigating the negative effects of irrelevant elements.
- **The causal modeling approach focuses on extracting key information in latent spaces and eliminating the interference of irrelevant elements**, which provides a meaningful direction for mitigating the out-of-distribution issues. Notably, a refined causal feature learning could further enhance performance.
- **Incorporating prototypical contrastive learning to guide different clients in learning consistent class-level representations proves to be more effective in CIFAR10 and CIFAR100 than in other datasets** (Fedproc and FPL). This is due to the increased difficulty of learning consistent features under significant attribute variance.

C. Ablation Study

This section further studies the effectiveness of different modules of FedAKD. The results are summarized in Table III.

- **Incorporating the Attentive Modeling (AM) and Attention-Guided Representation Learning (AGRL) modules significantly improve the performance**, indicating their role in enhancing causal discovery.
- **Leveraging different loss functions (KL, L2 and JS) for attention-guided representation learning yielded similar**

TABLE III

ABLATION STUDY ON THE EFFECTIVENESS OF DIFFERENT COMPONENTS OF FEDAKD ON THE NICO-ANIMAL AND COLORMNIST.

	NICO-Animal	COLORMNIST	
		β=0.1	β=0.5
Base	29.84±0.6	57.48±0.4	88.15±0.6
+ $\mathrm{AM}_{\mathrm{layer}}$ + $\mathrm{AGRL}_{\mathrm{js}}$	**33.12±0.5**	57.98±0.7	88.86±0.9
+ $\mathrm{AM}_{\mathrm{layer}}$ + $\mathrm{AGRL}_{\mathrm{l2}}$	32.61±0.7	58.04±0.9	**89.32±0.4**
+ $\mathrm{AM}_{\mathrm{layer}}$ + $\mathrm{AGRL}_{\mathrm{kl}}$	32.92±0.5	**58.13±0.9**	89.24±0.8
+ $\mathrm{AM}_{\mathrm{input}}$ + $\mathrm{AGRL}_{\mathrm{js}}$	34.97±0.6	58.31±0.3	91.71±0.3
+ $\mathrm{AM}_{\mathrm{input}}$ + $\mathrm{AGRL}_{\mathrm{l2}}$	34.72±0.4	**59.15±0.5**	91.38±0.6
+ $\mathrm{AM}_{\mathrm{input}}$ + $\mathrm{AGRL}_{\mathrm{kl}}$	**35.56±0.4**	59.03±0.2	**92.75±0.3**

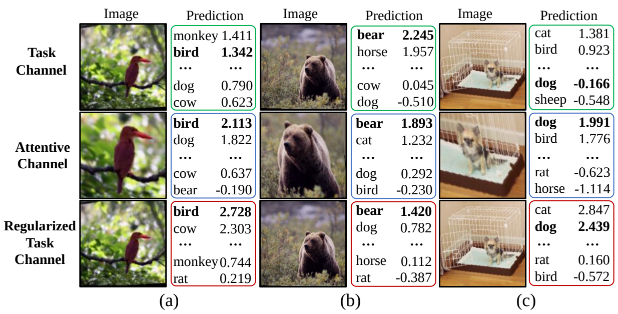

Fig. 3. Visualization of original images in the task channel and transformed images in the attentive channel, along with predictions of the corresponding model. And the predictions from the task channel after applying regularization.

results. This demonstrates FedAKD's insensitivity to the variance in loss function selection, highlighting the pivotal role of its attention-guided mechanism and evidencing the substantial robustness.

- **Attention modeling in the input space ($\mathrm{AM}_{\mathrm{input}}$) generally performs better than in the feature space ($\mathrm{AM}_{\mathrm{layer}}$).** This can be attributed to the AM module captures more key information in the input stage, whereas the feature space inherently loses some meaningful information.

D. Case Study

1) Analysis of the Effectiveness of Attentive Modeling: This section further analyzes the effectiveness of attentive modeling. As shown in Figure 3, we visualize the original images and attentive images of samples. Additionally, predictions for images in the original task channel, attention channel, and regularized task channel are also outputted. Obviously, **objects related to categories appear clearer in the attentive images.** This provides meaningful guidance for the task channel. Specifically, **the AM module can help the local model in the task channel correct prediction errors**, which achieves this by heightening sensitivity to crucial information, as shown in Figure 3(a). Secondly, **the AM module can further assist local model in strengthening the prediction confidence**, which enlarge the gap in predictions between different categories, as illustrated in Figure 3(b). Figure 3(c) shows a case that even though the AM module did not facilitate a correct prediction in the regularized task channel it still lessened the margin to the top-1 prediction. In summary, the AM module can leverage instructive features in input images to aid in boosting the overall accuracy of local models.

2) Error Analysis of FedAKD: In this section, we analyze the working mechanism of FedAKD, focusing on feature attention using GradCAM [38] and model outputs. Figure

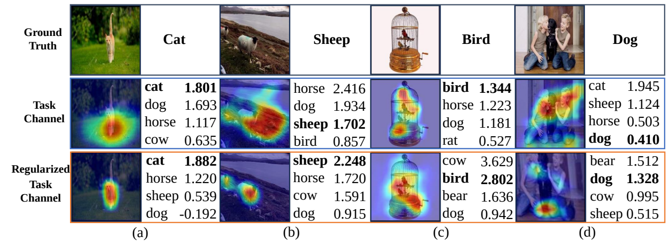

Fig. 4. Error analysis of FedAKD. (a) FedAKD enhances the focus on causal regions as well as the confidence in predictions. (b) FedAKD can utilize attentive modeling and distillation to correct prediction errors. (c) Smaller visual targets may result in a failure of the model to focus on the object. (d) Multiple objects in the image may reduce the model in focusing on the primary object.

4(a) illustrates that both FedAvg and FedAKD made correct predictions, with **FedAKD achieving a more precise focus on the object.** However, as shown in Figure 4(b), FedAvg struggles with complex contexts, while **FedAKD accurately focuses on the causal region.** This can be attributed to the guidance provided by attentive modeling. In the Figure 4(c), FedAKD faces challenges with undue attention to context, making it difficult to distinguish between the object and context. FedAvg also fails to focus on the core object despite a correct prediction. As illustrated in Figure 4(d), both FedAvg and FedAKD encounter challenges in focusing on the object within intricate contexts. However, **FedAKD demonstrates reduce the attention to the irrelevant region**, which decrease the prediction disparity between 'dog' and the top-1 category. This highlights the advantage of FedAKD in federated classification.

V. CONCLUSION

This paper presents a novel attentive modeling and distillation mechanism in federated learning, termed FedAKD, to handle the out-of-distribution issue. It performs attention-guided representation learning to instruct the local models to focus on the meaningful objects within the images. Experimental results show that FedAKD can effectively improve the performance by focusing on the important regions. This enhances the generalization ability of local models and the collaborative effect among them.

There are some directions for further exploration in this study. First, stronger attentive modeling techniques [39], [40] that more accurately identify causal regions can provide the meaningful information. Second, better feature learning methods can further improve the performance [41]–[45]. Third, it is anticipated that applying FedAKD to some challenging tasks would be promising [46]–[55].

REFERENCES

[1] Dashan Gao and et al., "A survey on heterogeneous federated learning," *arXiv preprint arXiv:2210.04505*, 2022.

[2] Dongmin Huang and et al., "Generator-based domain adaptation method with knowledge free for cross-subject eeg emotion recognition," *Cognitive Computation*, vol. 14, no. 4, pp. 1316–1327, 2022.

[3] Hao Guan and et al., "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024.

[4] Qing-Ling Guan, Yuze Zheng, Lei Meng, Li-Quan Dong, and Qun Hao, "Improving the generalization of visual classification models across iot cameras via cross-modal inference and fusion," *IEEE Internet of Things Journal*, 2023.

[5] Haokai Ma and et al., "Plug-in diffusion model for sequential recommendation," *arXiv preprint arXiv:2401.02913*, 2024.

[6] Xin-Chun Li and et al., "Federated learning with position-aware neurons," in *CVPR*, 2022, pp. 10082–10091.

[7] Qiang Yang and et al., "Federated machine learning: Concept and applications," *TIST*, vol. 10, no. 2, pp. 1–19, 2019.

[8] Zhuang Qi and et al., "Clustering-based curriculum construction for sample-balanced federated learning," in *CAAI International Conference on Artificial Intelligence*. Springer, 2022, pp. 155–166.

[9] Tianhan Liu and et al., "Cross-training with prototypical distillation for improving the generalization of federated learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 648–653.

[10] Zhuang Qi and et al., "Cross-silo prototypical calibration for federated learning with non-iid data," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 3099–3107.

[11] Yue Tan and et al., "Fedproto: Federated prototype learning across heterogeneous clients," in *AAAI*, 2022, vol. 36, pp. 8432–8440.

[12] Wenke Huang and et al., "Rethinking federated learning with domain shift: A prototype view," in *CVPR*. IEEE, 2023, pp. 16312–16322.

[13] Haozhao Wang and et al., "Dafkd: Domain-aware federated knowledge distillation," in *CVPR*, 2023, pp. 20412–20421.

[14] Zhengquan Luo and et al., "Disentangled federated learning for tackling attributes skew via invariant aggregation and diversity transferring," *arXiv preprint arXiv:2206.06818*, 2022.

[15] Ruipeng Zhang and et al., "Federated domain generalization with generalization adjustment," in *CVPR*, 2023, pp. 3954–3963.

[16] A Tuan Nguyen and et al., "Fedsr: A simple and effective domain generalization method for federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38831–38843, 2022.

[17] Liang Gao and et al., "Feddc: Federated learning with non-iid data via local drift decoupling and correction," in *CVPR*, 2022, pp. 10112–10121.

[18] Neta Shoham and et al., "Overcoming forgetting in federated learning on non-iid data," *arXiv preprint arXiv:1910.07796*, 2019.

[19] Qinbin Li and et al., "Model-contrastive federated learning," in *CVPR*, 2021, pp. 10713–10722.

[20] Lin Zhang and et al., "Federated learning for non-iid data via unified feature learning and optimization objective alignment," in *CVPR*, 2021, pp. 4420–4428.

[21] Sungwon Han and et al., "Fedx: Unsupervised federated learning with cross knowledge distillation," in *ECCV*. Springer, 2022, pp. 691–707.

[22] Gihun Lee and et al., "Preservation of the global knowledge by not-true distillation in federated learning," *arXiv preprint arXiv:2106.03097*, 2021.

[23] Lin Zhang and et al., "Fine-tuning global model via data-free knowledge distillation for non-iid federated learning," in *CVPR*, 2022, pp. 10174–10183.

[24] Sai Praneeth Karimireddy and et al., "Scaffold: Stochastic controlled averaging for federated learning," in *ICML*. PMLR, 2020, pp. 5132–5143.

[25] Tian et al. Li, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[26] Xutong Mu and et al., "Fedproc: Prototypical contrastive federated learning on non-iid data," *Future Generation Computer Systems*, vol. 143, pp. 93–104, 2023.

[27] Wenke Huang and et al., "Learn from others and be yourself in heterogeneous federated learning," in *CVPR*, 2022, pp. 10143–10153.

[28] Max Jaderberg and et al., "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[29] S Kullback and et al., "On information and sufficiencyannals of mathematical statistics, 22, 79–86," *MathSciNet MATH*, vol. 3, 1951.

[30] Brendan McMahan and et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[31] G Lee and et al., "Preservation of the global knowledge by not-true self knowledge distillation in federated learning. arxiv 2021," *arXiv preprint arXiv:2106.03097*.

[32] Yaming Guo and et al., "Out-of-distribution generalization of federated learning via implicit invariant relationships," in *ICML*. PMLR, 2023, pp. 11905–11933.

[33] Yann LeCun, "The mnist database of handwritten digits," *http://yann.lecun. com/exdb/mnist/*, 1998.

[34] Tan Wang and et al., "Causal attention for unbiased visual recognition," in *CVPR*, 2021, pp. 3091–3100.

[35] Alex Krizhevsky and et al., "Learning multiple layers of features from tiny images," 2009.

[36] Quande Liu and et al., "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *CVPR*, 2021, pp. 1013–1023.

[37] Kaiming He and et al., "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[38] Ramprasaath R Selvaraju and et al., "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.

[39] Yuqing Wang and et al., "Causal inference with sample balancing for out-of-distribution detection in visual classification," in *CAAI*. Springer, 2022, pp. 572–583.

[40] Yuqing Wang and et al., "Meta-causal feature learning for out-of-distribution generalization," in *ECCV*. Springer, 2022, pp. 530–545.

[41] Zitan Chen and et al., "Class-level structural relation modeling and smoothing for visual representation learning," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 2964–2972.

[42] Xiangxian Li and et al., "Cross-modal learning using privileged information for long-tailed image classification," *Proceedings of CVM*, 2023.

[43] Zitan Chen and et al., "Class-aware convolution and attentive aggregation for image classification," in *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, 2023, pp. 1–7.

[44] Lei Meng and et al, "Learning using privileged information for food recognition," in *Proceedings of the 27th ACM international conference on multimedia*, 2019, pp. 557–565.

[45] Xiangxian Li and et al., "Comparative study of adversarial training methods for long-tailed classification," in *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 2021, pp. 1–7.

[46] Yuqing Wang and et al., "Multi-channel attentive weighting of visual frames for multimodal video classification," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[47] Haokai Ma and et al., "Cross-modal content inference and feature enrichment for cold-start recommendation," in *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2023, pp. 1–8.

[48] Lei Meng and et al., "Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 3460–3468.

[49] Jingyu Li and et al, "Unsupervised contrastive masking for visual haze classification," in *Proceedings of the 2022 International Conference on Multimedia Retrieval*, 2022, pp. 426–434.

[50] Xin Qi and et al, "Machine learning empowering drug discovery: Applications, opportunities and challenges," *Molecules*, vol. 29, no. 4, pp. 903, 2024.

[51] Sijin Zhou and et al, "Objectivity meets subjectivity: A subjective and objective feature fused neural network for emotion recognition," *Applied Soft Computing*, vol. 122, pp. 108889, 2022.

[52] Jingyu Li and et al., "Unsupervised segmentation of haze regions as hard attention for haze classification," in *International Conference on Image and Graphics*. Springer, 2023, pp. 346–359.

[53] Ran Wang and et al., "Learning to fuse residual and conditional information for video compression and reconstruction," in *International Conference on Image and Graphics*. Springer, 2023, pp. 360–372.

[54] Haokai Ma and et al., "Triple sequence learning for cross-domain recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1–29, 2024.

[55] Haokai Ma and et al., "Comparative study of adversarial training methods for cold-start recommendation," in *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, 2021, pp. 28–34.