

Cross-domain Federated Object Detection

Shangchao Su, Bin Li*, Chengzhi Zhang, Mingzhao Yang, Xiangyang Xue*

School of Computer Science, Fudan University

{scsu20, libin, zhangcz20, mzyang20, xyxue}@fudan.edu.cn

Abstract—Detection models trained by one party (including server) may face severe performance degradation when distributed to other users (clients). Federated learning can enable multi-party collaborative learning without leaking client data. In this paper, we focus on a special cross-domain scenario in which the server has large-scale labeled data and multiple clients only have a small amount of labeled data; meanwhile, there exist differences in data distributions among the clients. In this case, traditional federated learning methods can't help a client learn both the global knowledge of all participants and its own unique knowledge. To make up for this limitation, we propose a cross-domain federated object detection framework, named FedOD. The proposed framework first performs the federated training to obtain a public global aggregated model through multi-teacher distillation, and sends the aggregated model back to each client for fine-tuning its personalized local model. After a few rounds of communication, on each client we can perform weighted ensemble inference on the public global model and the personalized local model. We establish a federated object detection dataset which has significant background differences and instance differences based on multiple public autonomous driving datasets, and then conduct extensive experiments on the dataset. The experimental results validate the effectiveness of the proposed method.

Index Terms—Federated learning, Object detection, Cross-domain

I. INTRODUCTION

In some real-world scenarios, a centrally trained detection model (usually trained on a server) will be sent to multiple users (clients). However, the clients are often in different domains. For example, in autonomous driving scenarios, distribution discrepancy in weather, lighting, and vehicle type among different clients may all affect the prediction performance of the model. Therefore, it is necessary to perform collaborative learning on the datasets of the server and multiple clients. The trained model needs to learn new knowledge in the client domain while still preserving server's knowledge. Moreover, for the purpose of privacy protection, during the collaborative learning, the server cannot collect the raw datasets from the clients, which brings a great challenge.

Federated learning (FL) is a technology that has the potential to address this challenge. FL aims to allow each participant to perform joint machine learning without sharing its local data. The original FL [1] needs to learn a single global model to handle the requirements of multiple clients, i.e., one model fits all. In each round of communication, the server sends a global model to each client. The client uses its local data to update the model and sends the trained model (local

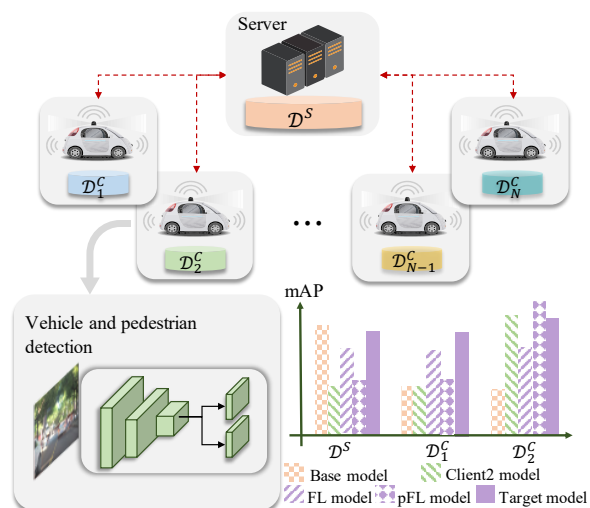


Fig. 1. The problem scenario. The server has a large amount of labeled data, each client has only a small amount of labeled data, and the data distribution of each client is different from one another. The base model trained on \mathcal{D}^S suffers severe performance degradation on client-side datasets.

model) back to the server to perform parameter averaging, and the server uses the averaged model as a new aggregated model. After multiple rounds of communication, the converged aggregated model is able to learn the knowledge from multiple clients. Some studies [2]–[4] have shown that FedAvg [1] can converge to good performance when the datasets of multiple clients are independently identically distributed (iid); but the performance is likely to drop significantly when the data is non-iid, i.e., the data for each client comes from different distributions. Unlike the original FL which aims to improve the performance of the global model, personalized federated learning (pFL) learns a personalized model for each client, thereby improving the performance of all client models and better handling non-iid challenges. For example, FedRep [5] enables each client to compute a personalized classifier and a globally shared feature extractor. FedBN [6] provides personalized batch normalization (BN) layers for different clients. These personalized methods can improve the traditional FL in non-iid scenarios to a certain extent.

However, in some practical scenarios, such as the server-client collaborative autonomous driving scenario (Figure 1), there are usually a server with a large-scale labeled dataset (more than ten times the data of the client) and multiple clients, each of which only has a small amount of labeled data. There is a distinct domain shift between each client and the server. In

*Corresponding author.

such a scenario, on one hand, the original FL may be affected by data imbalance and the non-iid issue, making personalized knowledge difficult to learn. On the other hand, while pFL is proposed to improve the client performance, it results in insufficient learning of the global domain (the union of the server and all clients). See Figure 1, the original FL (‘FL model’) focuses on the performance of the global domain, which can hardly obtain high performance on Client 2. pFL (‘pFL model’) focuses on personalized knowledge, but has lower performance on the server and Client 1. In the cross-domain federated object detection scenario considered in this paper, we aim to achieve this target, that for each client it is able to learn both its own personalized knowledge and the knowledge of most other domains.

To achieve this goal, in this paper, we propose a distillation-based federated learning framework for cross-domain Federated Object Detection (FedOD). Each client maintains two models, a public global model and a personalized local model. The public model is the aggregated model received from the server, the personalized model is a model trained based on the aggregated model and the local dataset. In each round of communication, each client’s personalized model is sent to the server, and then the server data is used to perform multi-teacher distillation on these personalized models to obtain the aggregated model. By mimicking the intermediate representations of the personalized models, the aggregated model can distill more global knowledge through using the server data. After a few rounds of communication, the public global model and the personalized local model are used to perform ensemble inference. For moderating the importance of the public global model and the personalized local model, we adopt the weighted bounding boxes (for simplicity, we use ‘bboxes’ in the following) fusion strategy. In this way, each client can maintain both global knowledge and local personalized knowledge. Note that the performance of the ensemble model in the global domain can outperform a single model with the same parameter scale.

In order to validate the proposed framework, we construct a collaborative learning dataset for vehicle pedestrian detection based on three public datasets. We extract annotated images from BDD100K [7] to build the dataset for server, use the annotated images from SODA10M [8] to construct Client 1 and Client 2, and the NuScenes [9] dataset for Client 3 and Client 4. In addition to the distribution differences that these datasets themselves have (regions, lighting, street types, shooting equipment, etc.), in order to introduce differences into instance class distributions, we choose some specific types of vehicles (e.g., motorcycles and trucks) to keep them only in the two NuScenes clients and remove them from the server and the other clients.

Our contributions are summarized as follows: 1) We propose a new realistic federated learning scenario, which is the first attempt to consider server-client collaborative object detection, in which, neither the original FL nor the pFL methods can fully meet the requirements. 2) We propose a federated object detection (FedOD) framework that employs multi-teacher

distillation and weighted bboxes fusion to enable each client to simultaneously obtain global knowledge and personalized knowledge. 3) We establish a federated learning dataset based on public datasets and conduct extensive experiments to validate the effectiveness of our proposed method.

II. METHOD

A. Problem Setting

Suppose we have a server and N clients. The i -th client has its own data distribution \mathcal{P}_i , and its dataset \mathcal{D}_i^C is collected from \mathcal{P}_i ; while the server has dataset \mathcal{D}^S collected from distribution \mathcal{P}^S , where $|\mathcal{D}^S| \gg |\mathcal{D}_i^C|$ for $i = 1, \dots, N$. Note that there exist obvious differences among different data distributions.

The original federated learning is to learn a global model that works well for all clients, a classic description is to minimize the objective $\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{N} \sum_{i=1}^N L_i(\mathbf{w})$, where $L_i(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_i} [l(\mathbf{w}; \mathbf{x})]$, l is the loss function. Based on this objective, in each round, FedAvg [1] samples m clients and collects the trained models from these clients. Then the aggregated model is $\sum_{i=1}^m \lambda_i \mathbf{w}_i$, where λ_i is the importance of the i -th model, usually determined by the amount of data on the client. In addition, $\sum_i \lambda_i = 1$.

Personalized federated learning is to minimize the following objective function to assign parameter \mathbf{w}_i to the i -th client: $\min_{\{\mathbf{w}_1, \dots, \mathbf{w}_N\}} \frac{1}{N} \sum_{i=1}^N L_i(\mathbf{w}_i)$.

In this paper, we consider a new yet realistic federated learning setting: A large amount of labeled data exists on the server, so it is crucial to make good use of the knowledge on the server side. Our goal is that the client can learn new personalized knowledge from each client while still preserving the knowledge of the global domain. To this end, the following objective is adopted:

$$\min_{\{\mathbf{w}_1, \dots, \mathbf{w}_N\}} \frac{1}{N} \sum_{i=1}^N L_U(\mathbf{w}_i) \quad (1)$$

where $L_U \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{P}^S \cup \mathcal{P}_1 \cup \dots \cup \mathcal{P}_N} [l(\mathbf{w}; \mathbf{x})]$. L_U evaluates the performance of the local model in both its own domain and in other domains. Note that when testing for a certain client, due to the large scale of the server data, $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}^S \cup \mathcal{D}_1^C \cup \dots \cup \mathcal{D}_N^C} [l(\mathbf{w}; \mathbf{x})]$ will be dominated by the server data. Even if the performance on the client testset is greatly improved, a small decline on the server dataset may lead to an observable decline in the performance of the overall union set. Thus the learning effect of personalized knowledge will be covered up. Therefore, in the experimental part, we will show the performance of the model in both the global domain and the local domain.

B. FedOD

Overview. The overall illustration is shown in Figure 2. Suppose we have a base model \mathbf{w}_b trained in the server domain. If \mathbf{w}_b is sent directly to the clients, the performance of \mathbf{w}_b on the client-side will be greatly reduced due to the difference among data distributions of the clients and the server. Our goal is to develop a better model from \mathbf{w}_b within

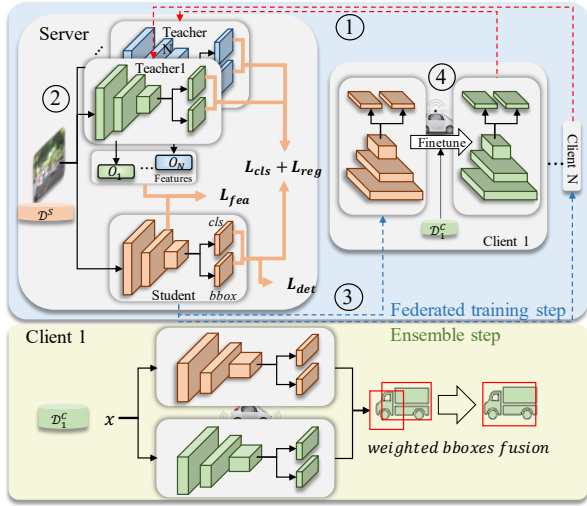


Fig. 2. Overview of the proposed framework for cross-domain federated object detection. We first conduct a few rounds of communication. Second, we apply weighted bboxes fusion to ensemble the outputs of these two models.

few communication rounds to achieve Eq.1. The proposed framework is illustrated in Figure 2: The first step is federated training. In this step, each client always maintains two models, a global public model and a local personalized model. In each communication round, each client sends its personalized model to the server (①), and the server employs multi-teacher distillation to aggregate these personalized models (②). The aggregated model is then sent back to the client as a global public model (③), and the client uses its local data to update the public model (in this paper, we fine-tune the public model) to obtain a personalized model (④). The second step is ensemble. In order to solve the dilemma that the client model can not take into account the global and local knowledge in the traditional FL setting, we adopt a new idea to improve the client’s ability of learning global and local knowledge through ensembles. We make use of the two detection models maintained by each client to ensemble, i.e., perform weighted bounding boxes (bboxes) fusion on the outputs of the global public model and the local personalized model.

The Federated Training Step. We observe that when there are obvious differences in the data distributions of clients, the public global aggregated model itself w_g seems unable to simultaneously extract the knowledge from the server and multiple clients within a few number of communication rounds, so we resort to introducing an additional personalized model w_i on the i -th client at the same time. In the federated training step, the global model and the personalized models of all the clients are updated every round. The global model is aimed to learn as much knowledge of the global domain as possible. The personalized model is expected to learn more personalized knowledge based on the continuously improved global model, because under the assumption of pFL, global knowledge is conducive to the learning of personalized knowledge. By alternately updating these two models, the proposed framework can acquire more comprehensive global

Algorithm 1 FedOD

The Federated Training Step:

Initialize global model $w_g \leftarrow w_b$

for each round $t = 1, 2, \dots$ **do**

$S_t \leftarrow$ Randomly select n clients

for each client $i \in S_t$ **in parallel do**

$w_i \leftarrow$ ClientTrain(D_i^C, w_g)

end for

Initialize the student model w_g by Eq.2.

Update w_g by Eq.3 using $\{w_i\}_{i=1}^n$ as teachers.

end for

The Ensemble Step:

for each client $i = 1, 2, \dots$ **do**

Get the latest w_g from the server

Finetune w_g with local dataset D_i^C

Ensemble w_g and w_i as the final model of the i -th client.

end for

and personalized knowledge for the subsequent ensemble step. Note that w_g is not necessary to perform the best in all client domains, but needs to be able to capture the knowledge of most parties involved, that is, w_g needs to outperform the base model w_b in the global domain $\mathcal{P}^S \cup \mathcal{P}_1 \cup \dots \cup \mathcal{P}_N$.

During federated training, each client first sends its personalized model w_i to the server. Next, the server is responsible for aggregating these local models. The traditional FL methods only conduct simple parameter averaging in the server, and do not fully utilize the server data. Therefore, to significantly improve the model aggregation efficiency, we employ the multi-teacher distillation algorithm for model aggregation on the server. Before distillation, the student model is initialized by averaging the aggregated model of the previous round w_g and the personalized models w_i :

$$w_g \leftarrow \frac{1}{N+1} \left[w_g + \sum_{i=1}^N w_i \right] \quad (2)$$

The feature extraction part of the detection model includes the feature maps of multiple scales, we add the channel-wise distillation loss [10] to the output feature maps of each scale respectively. Suppose that the feature map of the l -th scale of the i -th client (teacher) model is $O_i^{C,l}$ and the feature map of the student model is $O^{S,l}$, we first obtain an attention map by applying softmax to these features: $\mathcal{A}(O_{k,j}) = \frac{\exp(\frac{O_{k,j}}{\mathcal{T}})}{\sum_{j=1}^{W*H} \exp(\frac{O_{k,j}}{\mathcal{T}})}$ where k indexes the channel, W and H are the width and height of the feature map, respectively; \mathcal{T} is the temperature parameter. Then the KL divergence is used to measure the distance between the feature maps of the student and each teacher:

$$KL(O^{S,l}, O_i^{C,l}) = \frac{\mathcal{T}^2}{K} \sum_{k=1}^K \sum_{j=1}^{W \cdot H} w_l \cdot H_l \mathcal{A}(O_{k,j}^{S,l}) \cdot \log \frac{\mathcal{A}(O_{k,j}^{S,l})}{\mathcal{A}(O_{k,j}^{C,l})}$$

The overall feature loss is $\mathcal{L}_{fea} = \frac{1}{N} \sum_l \sum_i KL(O^{S,l}, O_i^{C,l})$. Note that a more common multi-teacher distillation method

is to calculate the KL divergence after averaging the features, that is, use $KL\left(O^S, \frac{1}{N} \sum_i^N O_i^C\right)$. However, under the setting of this paper, the training data of each teacher model is quite different. If the features are directly averaged, their personalized information will be lost.

For detection models, such as RetinaNet [11], the output of the classification head can also be regarded as a feature map, so \mathcal{L}_{cls} has the same formula as \mathcal{L}_{fea} . For the regression head, we use the L2 distance as \mathcal{L}_{reg} . The overall loss for the distillation is:

$$\mathcal{L}_{distillation} = \lambda_1 * \mathcal{L}_{fea} + \lambda_2 * \mathcal{L}_{cls} + \lambda_3 * \mathcal{L}_{reg} \quad (3)$$

Furthermore, the original detection loss \mathcal{L}_{det} is still applied to the student model. After distillation, the student model, as a new aggregated model \mathbf{w}_g , is sent to the clients. The details of the proposed FedOD is shown in Algorithm 1.

The Ensemble Step. After the previous step, for each client, we obtain an aggregated model \mathbf{w}_g , which has better generalization performance compared with the base model \mathbf{w}_b . Also we have a personalized model which fully excavates the personalized knowledge of the client. However, the unified aggregated model itself cannot adapt to the personalized knowledge of all clients, and the personalized model itself cannot well handle the samples from other domains. Therefore, we consider keeping these two models at the same time and ensembling their outputs. Although we introduce an additional model compared to the pFL setting, it will have better performance under the same number of model parameters. We will show this in the experiment section.

The key to ensemble the detection models is how to filter the predicted bboxes. In order to ensemble the models trained from the same domain, a simple method is to use traditional non-maximum suppression (NMS) algorithm. However, under the setting of this paper, the public global model and the personalized local model are not trained from the same domain. Due to domain shift, these two models are likely to have different understandings of the same object, we cannot simply decide which box to discard and which to keep. In this case, standard NMS methods have the possibility to ignore valid predictions, we can't simply drop one and keep the other, but should fuse them. See supplementary materials for more analysis. To this end, we adopt weighted boxes fusion (WBF) [12] algorithm, which uses all the bboxes. The idea of WBF is simple. First, put the output bboxes of the two models into set \mathcal{B} , then cluster the predictions in \mathcal{B} by the preset IoU threshold and fuse the bboxes for each cluster. Suppose a certain cluster has T bboxes, and the confidence of the i -th bbox is C_i , the coordinate vector is \mathbf{b}_i , then the fusion rule is: $C \leftarrow \frac{\min(T, M)}{M * T} \sum_{i=1}^T C_i$, $\mathbf{b} \leftarrow \frac{\sum_{i=1}^T C_i * \mathbf{b}_i}{\sum_{i=1}^T C_i}$, where M is the number of the models to be ensembled, in this paper, $M = 2$. C and \mathbf{b} will be the new confidence and coordinate of the fused bbox. Weighted bboxes fusion makes use of all bboxes at the same time, so it is especially suitable for the ensemble of cross-domain models in this paper. In subsequent experiments, we will compare different ensemble methods.

III. EXPERIMENTS

In this section, we start by setting up our experiments, and then introduce the construction of the dataset, report and analyze the experimental results.

Implementation Details. We implement our framework using MMDetection [13]. The detection model is RetinaNet [11]. The base model \mathbf{w}_b on the server is initialized by COCO2014 [14] pretrained parameters. We use four NVIDIA GeForce RTX 3090 GPUs to do federated training. By default, we set the batch size $B = 16$, learning rate $lr = 0.01$, the local training epoch to be 12, and the distillation training epoch to be 5.

Evaluation Metrics. We use mean Average Precision (mAP) to evaluate the detection performance. Assume that after the federated learning step, the performance of each client on its own testset is r_i^p , on the server testset is r_i^s , and the performance on the global testset (the union of the server testset and all client testsets) is r_i^u . Specifically, r_i^u measures the global generalization ability; r_i^p measures the ability to learn the local personalized knowledge; r_i^s measures the degree of forgetting of the initial knowledge on the server. In short, we finally report the following four indicators: $A_s = \frac{1}{N} \sum_{i=1}^N r_i^s$, $A_p = \frac{1}{N} \sum_{i=1}^N r_i^p$, $A_u = \frac{1}{N} \sum_{i=1}^N r_i^u$ and $A_{com} = \frac{1}{N} \sum_{i=1}^N \alpha * r_i^p + (1 - \alpha) * r_i^u$, where α is a trade-off hyperparameter. Note that for the real dataset, the scale of server data is large, which makes A_u unable to reflect the learning ability of personalized knowledge. Therefore, we should use the weighted sum of A_u and A_p . We report A_{com} with different α in Table I to observe the effect of α .

Compared Methods. Two baselines and three competitors are compared as follows. 1) **Baseline1**: the base model \mathbf{w}_b of the server. This represents the performance of doing nothing but using the base model as the local model for each client. 2) **Baseline2**: to show the effectiveness of federated learning, we directly ensemble the base model and the first round finetuned model on the client side, as baseline2 ($E(\mathbf{w}_i^1, \mathbf{w}_b)$), the final ensemble results of FedOD should be better than baseline2. 3) **FedAvg** [1], **FedBN** [6] and **FedRep** [5] are our competitors. For fair comparison (we should compare different methods with the same scale parameters), we compare our aggregation model \mathbf{w}_g with r50 (ResNet50) backbone to the competitors with r50 backbone, and then use the ensemble model to compare with the competitors with r101 backbone. Note that the original setting of these works is different from ours, they do not have the server-side learning task. Therefore, in the implementation of these methods, we use the server as a client to participate in federated learning. 4) The ideal method: merge the data from the server and all clients together to train an ideal model. Note that the ideal model is not available in real scenarios, we cannot fuse all datasets due to privacy protection policy.

Dataset Construction. We extract data from three public datasets: 1) **BDD100K** [7] collected in New York, Berkeley, San Francisco and Bay area. It has 100K labeled images with 10 classes for object detection task. 2) **SODA10M** [8]

TABLE I

PERFORMANCE COMPARISON. w_b IS THE BASE MODEL TRAINED ONLY WITH THE SERVER DATA; w_g^1 IS A MODEL AGGREGATED BY $\{w_i^1\}_{i=1}^N$; w_i^2 IS THE MODEL FINE-TUNED FROM w_g^1 USING THE CLIENT DATASETS; $E(\cdot, \cdot)$ DENOTES THE ENSEMBLE MODEL.

	Ideal model	Without Joint Learning			FedOD 1-th round			FedOD 2-th round			FedOD 3-th round			
		Baseline1 w_b	w_i^1	Baseline2 $E(w_i^1, w_b)$	w_g^1	w_i^2	$E(w_i^2, w_g^1)$	w_g^2	w_i^3	$E(w_i^3, w_g^2)$	w_g^3	w_i^4	$E(w_i^4, w_g^3)$	
A_s	26.90	25.50	18.95	24.23	24.50	18.73	23.13	24.20	18.43	22.90	24.00	18.18	22.53	
A_p	42.83	21.25	39.83	38.65	32.95	42.00	41.18	35.20	42.38	41.93	35.78	42.90	42.45	
A_u	41.30	22.60	29.33	31.80	32.50	30.13	33.93	33.40	30.25	34.53	33.90	30.35	34.88	
A_{com}	$\alpha = 0.1$	41.45	22.47	30.38	32.49	32.55	31.31	34.65	33.58	31.46	35.27	34.09	31.61	35.63
	$\alpha = 0.3$	41.76	22.20	32.48	33.86	32.64	33.69	36.10	33.94	33.89	36.75	34.46	34.12	37.15
	$\alpha = 0.5$	42.06	21.93	34.58	35.23	32.73	36.06	37.55	34.30	36.31	38.23	34.84	36.63	38.66

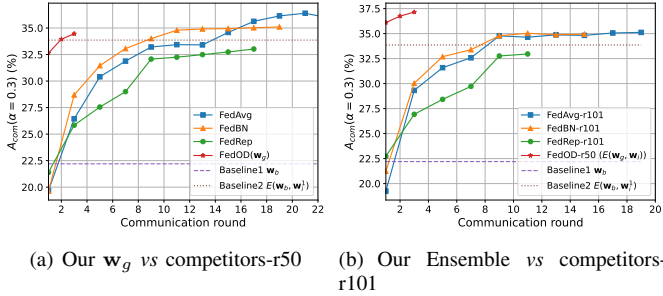


Fig. 3. Performance compared to competitors. FedOD can achieve competitive results with only a few rounds and outperforms the competitors in the global domain.

collected in 32 cities in China. It consists of 10 million unlabeled images and 20K labeled images. 3) **NuScenes** [9] collected in Boston Seaport and Singapore One North, Queenstown and Holland Village. It consists of 14 million unlabeled images and 90K labeled images. In addition to regional differences, the data acquisition equipment and the labeling method also produce endogenous differences for the three datasets. Considering the large number of labeled images of BDD100K, we choose it and extract images including cars, pedestrians and riders to build the dataset for the server. Then we use the labeled 20K images in SODA10M to construct two client datasets. These two clients are independent identically distributed, but they are significantly different from the other clients and the server datasets. We select NuScenes to construct another two client datasets. In order to introduce new vehicle instances compared to the other participants, such as trucks and motorcycles, we keep these types of vehicles only on the two NuScenes clients and remove them from the server and the other clients. The overall dataset details are shown in supplementary materials.

A. Main Results

We present the results of the compared methods. For the federated learning step of FedOD, we set up three rounds of federated learning. The comparison results with Baseline 1 and 2 are shown in Table I. We highlight the following points:

- 1) **The base model w_b cannot handle the client data.** Furthermore, the base model degrades more severely in NuScenes clients than in the two clients built from SODA,

reflecting the challenges posed by the imbalance of classes and the unseen instances (trucks, motorcycles, etc.).

- 2) **The personalized models w_i cannot handle the server data.** See w_i^1 . A_p increases from 21.15% to 39.83%, but A_s drops from 25.5% to 18.95% which cause that w_i^1 improves slightly in the union testset, only from 22.60% to 29.30%. In comparison, our aggregated model in the 1-th round w_g^1 improves the A_u of w_b by a large margin, from 22.60% to 32.50%. This is because w_g^1 remembers most of the server-side knowledge while improving client-side performance.
- 3) **The ensemble model after federated learning works well.** Both the generalization indicator A_u and the comprehensive indicator A_{com} of our ensemble results outperform the Baseline2. It can be observed that just ensembling the base model w_b with the fine-tuned client-side model w_i^1 is sub-optimal. $A_{com}(\alpha = 0.3)$ can be improved from 33.86% to 37.15% following our framework. In addition, with different α , the ensemble step of FedOD always achieves the best results.
- 4) **Federated learning improves the learning ability of the local model to learn personalized knowledge.** After three rounds of federated learning, the personalized model w_i^4 improves the A_p of w_i^1 from 39.83% to 42.90%, even surpassing the ideal model.

The proposed FedOD is also compared with the traditional FL method FedAvg [1] and pFL methods, FedBN [6] and FedRep [5]. For FedAvg, to be fair, we show the results of our aggregated model w_g and the ensemble model, respectively. The results are shown in Figure 3. It can be seen that the methods within the traditional FL framework converge slowly under the setting of this paper, while FedOD can achieve competitive results with only three rounds of communication. To explore the influence of the amount of parameters on compared methods, we also add the FL methods with ResNet101 backbone as a comparison. We can see our ensemble model outperforms the other compared methods with r101 backbone. By increasing the parameter size, FedAvg-r101 performs better than FedAvg with ResNet50 backbone in the initial stage, but the performance after convergence is mediocre. This observation indicates that simply increasing the parameter size cannot achieve comparable performance to the ensemble. See the supplementary materials for the detailed performance of different methods.

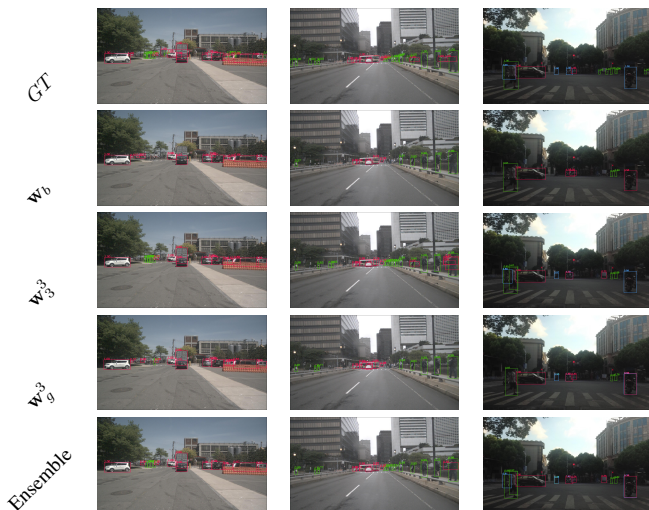


Fig. 4. Visualization. The different effects of different models during the federated learning procedure from the perspective of the testset in client 3. The last column is a failure case.

TABLE II
COMPARISON OF DIFFERENT BBOXES FILTER METHODS.

	NMS	SoftNMS	NWM	WBF
A_s	22.73	21.98	21.85	23.13
A_p	39.25	39.13	37.68	41.18

B. Other Results

Visualization Results. To show the performance changes of different models more concretely during the federated learning, we visualize the detection results in Figure 4. The following phenomenons can be observed: 1) The base model w_b cannot completely detect instances that are not on the server side, such as trucks and buses. 2) Through FedOD, the aggregated model w_g^3 can learn the personalized knowledge of the client. As one can see from the fourth row of the Figure 4, the aggregated model w_g^3 is already able to detect trucks and buses, albeit with some redundancy.

We also show a failure case in the last column of Figure 4, in which, the aggregated model w_g^3 fails to learn the client’s personalised knowledge. Due to the severe imbalance in the number of annotations between ‘Rider’ and other classes in the dataset, and there are few instances of ‘Rider’ on each party except client 0 and client 1. Therefore, the information learned by the aggregated model w_g^3 from the client model w_0^3 is limited. This implies that the class imbalance in the federated learning is a direction worthy of further exploration.

Different bboxes post-processing methods. We discuss the impact of different methods to choose the bboxes on the ensemble step. The results are in Table II.

IV. CONCLUSION

This paper proposes a novel cross-domain federated object detection framework FedOD for server-client collaborative autonomous driving scenarios. This framework includes a federated training step based on multi-teacher distillation and

an ensemble step based on weighted bboxes fusion. Extensive experiments show that compared to the baselines and competitors, our framework achieves significant performance improvements, and can outperform the traditional FL and pFL methods in all domains.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China (No.62176061), National Key R&D Program of China (No.2021ZD0112803), STCSM projects (No.20511100400, No.22511105000), the Shanghai Research and Innovation Functional Program (No.17DZ2260900), and the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning.

REFERENCES

- [1] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *AISTATS*. 2017, vol. 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282, PMLR.
- [2] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang, “On the convergence of fedavg on non-iid data,” in *ICLR*. 2020, OpenReview.net.
- [3] Farzin Haddadpour and Mehrdad Mahdavi, “On the convergence of local descent methods in federated learning,” *CoRR*, vol. abs/1910.14425, 2019.
- [4] Shangchao Su, Bin Li, and Xiangyang Xue, “One-shot federated learning without server-side training,” *Neural Networks*, 2023.
- [5] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai, “Exploiting shared representations for personalized federated learning,” in *ICML*. 2021, vol. 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099, PMLR.
- [6] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou, “Fedbn: Federated learning on non-iid features via local batch normalization,” in *ICLR*. 2021, OpenReview.net.
- [7] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell, “BDD100K: A diverse driving dataset for heterogeneous multitask learning,” in *CVPR*. 2020, pp. 2633–2642, Computer Vision Foundation / IEEE.
- [8] Jianhua Han, Xiwen Liang, Hang Xu, Kai Chen, Lanqing Hong, Jiageng Mao, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Xiaodan Liang, and Chunjing Xu, “SODA10M: A large-scale 2d self/semi-supervised object detection dataset for autonomous driving,” in *NeurIPS Datasets and Benchmarks*, 2021.
- [9] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” in *CVPR*. 2020, pp. 11618–11628, Computer Vision Foundation / IEEE.
- [10] Changyong Shu, Yifan Liu, Jianfei Gao, Zheng Yan, and Chunhua Shen, “Channel-wise knowledge distillation for dense prediction*,” in *ICCV*. 2021, pp. 5291–5300, IEEE.
- [11] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár, “Focal loss for dense object detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020.
- [12] Roman A. Solovyev, Weimin Wang, and Tatiana Gabruseva, “Weighted boxes fusion: Ensembling boxes from different object detection models,” *Image Vis. Comput.*, vol. 107, pp. 104117, 2021.
- [13] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin, “Mmdetection: Open mmlab detection toolbox and benchmark,” *CoRR*, vol. abs/1906.07155, 2019.
- [14] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, “Microsoft COCO: common objects in context,” in *ECCV (5)*. 2014, vol. 8693 of *Lecture Notes in Computer Science*, pp. 740–755, Springer.