Cooperative Pseudo Labeling for Unsupervised Federated Classification

Supplementary Material

1. Algorithm Flow of FedCoPL

We summarize the procedure of FedCoPL in Algorithm 1.

Algorithm 1 FedCoPL

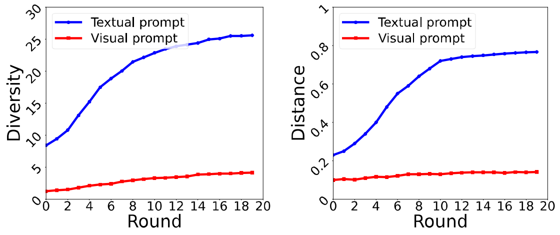
```
Input: number of communication round $T$, client number $K$, unlabeled dataset $\{D_k\}_{k=1}^{K}$, client participating rate $R$, number of local update epochs $E$, batch size $B$, learning rate $\eta$, pseudo labels update interval $Q$.
Output: the global visual prompt $P^v$ and personalized textual prompts $\{P_k^t\}_{k=1}^{K}$.
Initialize $P^v$ , $\{P_k^t\}_{k=1}^{K}$
$m \leftarrow \max(\lfloor R \cdot K \rfloor, 1)$
for communication round $r = 1, 2, \cdots, T$ do
    if r % Q = 0 then
        # cooperative pseudo labeling
        for $k = 1, ..., K$ do
            Obtain the estimated set $D_k^{est}$ with Eq. (2).
            Obtain the estimated statistics $\widetilde{U}_k$ with Eq. (3).
            Obtain $\widetilde{D}_k$ by selecting the most confident samples according to the capacity $\widetilde{U}_k$.
        end for
    end if
    $M \leftarrow$ Randomly select a subset containing $m$ clients.
    # local update
    for each client $k \in M$ do
        Initialize local visual prompt $P_k^v \leftarrow P^v$
        for each batch $\mathcal{B}_i = \{x, \hat{y}\} \in \widetilde{D}_k$ do
            $P_k^v \leftarrow P_k^v - \eta \nabla \mathcal{L}(P_k^v; \mathcal{B}_i)$
            $P_k^t \leftarrow P_k^t - \eta \nabla \mathcal{L}(P_k^t; \mathcal{B}_i)$ # $\mathcal{L}$ is cross-entropy loss
        end for
    end for
    Obtain aggregated visual prompt $P^v$ with Eq. (4).
end for
```

2. Drift Diversity

Following [5], we employ drift diversity to assess magnitude differences, which is defined as follows:

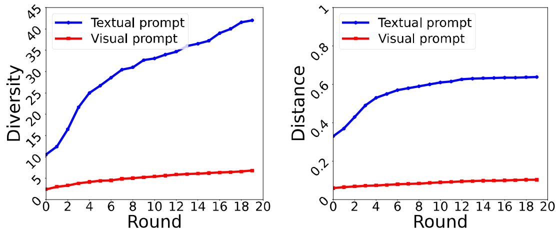
$$\xi^r := \frac{\sum_{k=1}^{K} \|m_k^r\|^2}{\|\sum_{k=1}^{K} m_k^r\|^2} \quad \text{with} \quad m_k^r = P_k^r - P^{r-1} \qquad (1)$$

where P_k^r is updated prompt of client k in round r and P^{r-1} is aggregated prompt on the server in round $r-1$.



(a) Drift Diversity. (b) Cosine Distance.

Figure 1. Drift diversity and cosine distance of prompts among clients during training in DTD [2] dataset. The differences observed in textual prompts are significantly greater than those found in visual prompts.



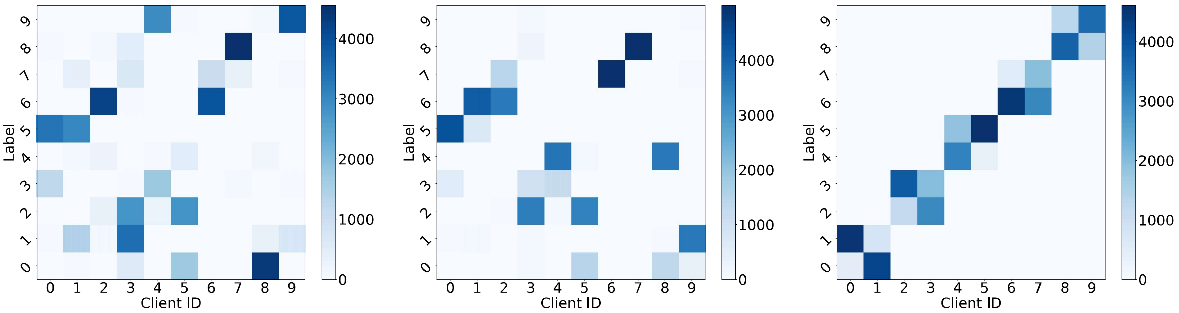
(a) Drift Diversity. (b) Cosine Distance.

Figure 2. Drift diversity and cosine distance of prompts among clients during training in RESISC45 [1] dataset.

Besides, we measure the differences in both textual and visual prompts across all clients using drift diversity [5] and cosine distance in RESISC45 and DTD datasets, which respectively reflect the diversity in the amount and direction of prompts updates among clients, as shown in Figure 1 and Figure 2. These results prove that the differences in textual prompts are significantly greater than those in visual prompts, which confirms our hypothesis that visual prompts tend to be more similar, while textual prompts exhibit greater variability.

3. More Experimental Details

Dataset setup. We evaluate our approach on six diverse visual classification datasets. Table 1 summarizes the key statistics of each dataset, including the original task domain, number of classes, and the number of training and testing samples. We simulate data heterogeneity using both quantity-based and Dirichlet-based label skews. In the quantity-based setting, each client is assigned a fixed number of classes: 10 for DTD, 58 for RESISC45, 66 for CUB, 30 for UCF101, 2 for CIFAR-10, and 20 for CIFAR-100. For the Dirichlet-based label skew, we generate client data



(a) Dirichlet-based ($\beta = 0.1$). (b) Dirichlet-based ($\beta = 0.05$). (c) Quantity-based.

Figure 3. (a) and (b) depict label skews of Dirichlet-based label skews and (c) presents the quantity-based label skew.

Table 1. The detailed statistics of datasets used in experiments.

Dataset	Task	Classes	Training Size	Testing Size
CUB	Image classification	200	5,594	5,794
RESISC45	Scene classification	45	6,300	25,200
UCF101	Action recognition	101	7,639	3,783
DTD	Texture recognition	47	2,820	1,692
CIFAR10	Image classification	10	50,000	10,000
CIFAR100	Image classification	100	50,000	10,000

using Dirichlet distributions with concentration parameters $\beta = \{0.1, 0.05\}$. To illustrate the label distribution under each setting, we visualize the client-level class allocations on CIFAR-10, as shown in Fig. 3.

Implementation details. All input images are resized to 224 × 224 and partitioned into 14 × 14 patches with an embedding dimension of 768. We incorporate deep visual prompts by appending trainable prompts of size 5 × 867 to the output of each transformer layer in the visual encoder. For the text encoder, we use prompts of length 16 with a dimensionality of 512. The batch size is set to 64 for both training and evaluation.

4. More Experiments Results

Results under different image encoder backbones. We further conduct experiments to evaluate the impact of different image encoders on model performance. The comparison results using RN50 are summarized in Table 2. These results reveal that as the zero-shot performance of the pre-trained image encoder declines, the performance of all methods deteriorates sharply. Notably, on the CUB dataset, all baseline methods achieve lower accuracy than the zero-shot baseline, whereas our proposed method surpasses the zero-shot accuracy. Overall, our approach consistently outperforms prior methods, underscoring the effectiveness of our strategy in enhancing the performance of smaller image encoders. These findings highlight the robustness of FedCoPL in real-world federated learning scenarios, particularly under limited computational resources.

Table 2. Experiments using CLIP RN50 as base model under Dirichlet-based label skews ($\beta = 0.1$) across four datasets. FPL [8] is adopted as the baseline pseudo labeling (**PL**) method.

Method	**PL**	**DTD**	**RESISC45**	**UCF101**	**CIFAR10**
Zero-shot CLIP	-	41.62	52.12	65.13	74.80
PromptFL [4]	FPL	42.36	56.16	62.90	78.01
PromptProx [7]	FPL	43.12	56.54	64.37	78.77
pFedPrompt [3]	FPL	43.65	57.13	64.97	79.69
FedOPT [6]	FPL	45.52	58.94	65.54	80.13
FedCoPL	**CoPL**	**51.83**	**70.78**	**76.48**	**84.85**

Table 3. Pseudo label accuracy (%) of different methods with Dirichlet-based label skews ($\beta = 0.1$, $\beta = 0.05$) and quantity-based label skew on various datasets.

Method	DTD	RESISC45	CUB	UCF101	CIFAR10	CIFAR100
			Dirichlet-based label skew ($\beta = 0.1$)			
FPL	66.36	77.06	77.97	80.31	91.02	83.05
CPL	69.72	78.21	80.64	81.17	92.18	84.81
CoPL (Ours)	**78.74**	**85.73**	**89.30**	**85.12**	**96.07**	**88.13**
			Dirichlet-based label skew ($\beta = 0.05$)			
FPL	63.54	72.51	76.72	78.19	90.26	81.87
CPL	65.45	76.68	78.24	80.71	90.37	82.08
CoPL (Ours)	**75.82**	**85.09**	**88.62**	**83.90**	**95.26**	**86.97**
			Quantity-based label skew			
FPL	55.18	61.02	62.79	65.44	87.79	76.18
CPL	56.50	64.90	66.92	68.66	88.50	76.91
CoPL (Ours)	**68.29**	**78.87**	**79.21**	**79.76**	**94.24**	**85.03**

Comparison of pseudo-label accuracy. As shown in Table 3, we report the accuracy of various pseudo labeling methods based on CLIP's zero-shot predictions in Dirichlet-based and quantity-based label skews. The proposed pseudo label selection strategy consistently outperforms baseline approaches across multiple datasets. These results underscore the effectiveness of the global pseudo label allocation strategy, which provides a robust foundation for subsequent model training. By explicitly accounting for global class distributions and aggregating client-level pseudo label distributions, our method effectively alleviates label skew across clients and enhances the consistency of assigned

Table 4. Performance (%) of FedCoPL under different values of hyperparameter τ_1 with Dirichlet-based label skews (β = 0.1) on four datasets ($\tau_2 = 0.50$).

τ_1	0.10	0.20	0.30	0.40	0.44	0.46	0.48	0.50	0.52	0.54	0.56	0.60	0.70	0.80	0.90
DTD	55.89	55.92	56.01	56.13	58.90	59.76	60.62	60.89	**60.92**	60.53	60.02	54.85	51.09	47.56	36.16
RESICS45	75.29	75.15	74.92	75.31	75.03	75.84	**76.31**	75.76	75.61	74.88	74.07	71.14	61.27	54.98	47.58
CIFAR10	83.38	80.77	82.84	85.34	92.68	93.83	95.24	95.38	95.52	**95.71**	94.10	90.41	77.56	49.76	55.59
CIFAR100	72.14	72.64	72.12	72.29	72.54	72.89	73.31	**73.59**	73.01	72.65	73.18	72.78	72.21	73.05	68.44

Table 5. Performance (%) of the FedCoPL under different values of hyperparameter τ_2 with Dirichlet-based label skews (β = 0.1) on four datasets ($\tau_1 = 0.50$).

τ_2	0.10	0.20	0.30	0.40	0.44	0.46	0.48	0.50	0.52	0.54	0.56	0.60	0.70	0.80	0.90
DTD	58.04	58.07	58.38	57.98	58.10	58.92	59.93	**60.89**	60.25	59.67	58.99	57.44	56.67	49.22	37.92
RESICS45	71.92	75.12	75.82	75.07	74.85	75.12	75.54	**75.76**	75.04	74.37	73.89	72.43	66.25	64.49	46.80
CIFAR10	81.01	81.92	83.87	82.73	94.08	94.93	95.22	95.38	**95.60**	95.19	94.68	88.36	71.55	42.33	50.92
CIFAR100	72.33	72.39	72.35	71.20	71.33	72.48	72.70	73.59	**73.62**	72.81	71.97	71.13	71.37	71.40	67.20

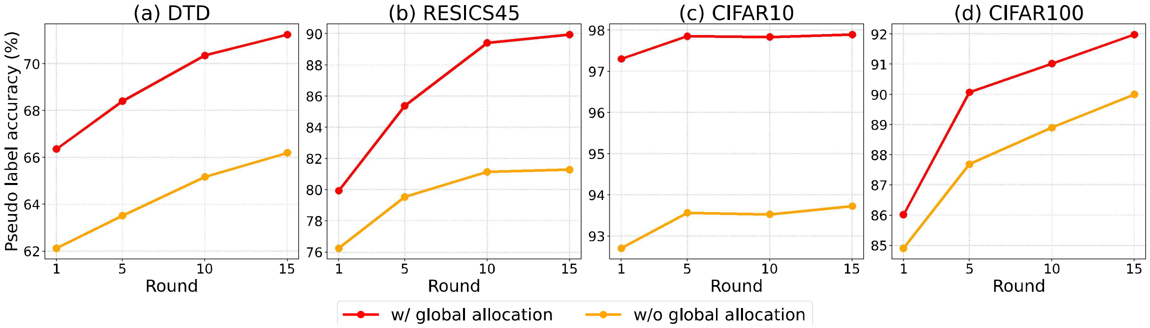


Figure 4. Pseudo label accuracy of our method with and without global pseudo label allocation on four datasets.

pseudo labels.

Sensitivity analysis of hyperparameters τ_1 and τ_2. To demonstrate the robustness of our method with respect to hyperparameter selection, we conduct experiments using a range of values for τ_1 and τ_2 on the DTD, RESISC45, CIFAR10, and CIFAR100 datasets. The results, summarized in Table 4 and Table 5, show that our method exhibits strong insensitivity to the choice of τ_1 and τ_2. Specifically, across $\tau_1 \in \{0.44, 0.46, 0.48, 0.50, 0.52, 0.54, 0.56\}$ (with $\tau_2 = 0.50$) and $\tau_2 \in \{0.44, 0.46, 0.48, 0.50, 0.52, 0.54, 0.56\}$ (with $\tau_1 = 0.50$), our approach consistently achieves approximately 60% on DTD, 76% on RESISC45, 95% on CIFAR10, and 73% on CIFAR100. This stable performance highlights the method's robustness and its capacity to deliver reliable results across varying hyperparameter settings. Notably, in our experiments, we did not perform extensive tuning to identify the optimal values of τ_1 and τ_2. Instead, we simply set both to the default value of 0.5, which may not represent the best possible configuration. This further emphasizes the effectiveness of our method, even without fine-grained hyperparameter selection.

Ablation study on global allocation of pseudo labels. In this subsection, we conduct an ablation study on the global pseudo label allocation strategy to validate its effectiveness. As shown in Fig. 4, we present the accuracy of pseudo labels with and without the global pseudo label allocation. The results show that global allocation not only achieves higher pseudo label accuracy but also leads to more stable and consistent convergence during training. This suggests that global pseudo label allocation among clients helps mitigate the influence of label skews, which are common challenges in federated learning. Moreover, improved pseudo label quality highlights the practical benefits of the proposed global allocation strategy.

References

[1] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 2017. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proc. CVPR*, 2014. 1

[3] Tao Guo, Song Guo, and Junxiao Wang. Pfedprompt: Learning personalized prompt for vision-language models in federated learning. In *Proceedings of the ACM Web Conference 2023*, 2023. 2

[4] Tao Guo, Song Guo, Junxiao Wang, Xueyang Tang, and Wenchao Xu. Promptfl: Let federated participants cooperatively learn prompts instead of models–federated learning in age of foundation model. *IEEE Transactions on Mobile Computing*, 23(5):5179–5194, 2023. 2

[5] Bo Li, Mikkel N Schmidt, Tommy S Alstrøm, and Sebastian U Stich. On the effectiveness of partial variance reduction in federated learning with heterogeneous data. In *Proc. CVPR*, 2023. 1

[6] Hongxia Li, Wei Huang, Jingya Wang, and Ye Shi. Global and local prompts cooperation via optimal transport for federated learning. In *Proc. CVPR*, 2024. 2

[7] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 2

[8] Cristina Menghini, Andrew Delworth, and Stephen Bach. Enhancing clip with clip: Exploring pseudolabeling for limited-label prompt tuning. In *Proc. NeurIPS*, 2023. 2