FedCode: Addressing Federated Domain Shift by Contrastive Feature Decoupling

Shaobo Zhang[a,b], Yijie Yin[a,b], Wei Liang[a,b], Fan Wu[c,*], Weizhi Meng[d], Tian Wang[e,f]

[a]*Sanya Institute of Hunan University of Science and Technology, Sanya, 572025, China*
[b]*School of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan, 411201, China*
[c]*School of Computer Science and Engineering, Central South University, Changsha, 410083, China*
[d]*School of Computing and Communications, Lancaster University, LA1 4YW Lancaster, U.K.*
[e]*Institute of Artificial Intelligence and Future Networks, Beijing Normal University and UIC, Zhuhai, 519087, China*
[f]*College of Computer and Data Science, Fuzhou University, Fuzhou, China*

Abstract

As a distributed collaborative technology that enables information fusion while preserving privacy, Federated Learning inherently faces challenges posed by data heterogeneity in multi-source knowledge sharing and integration. Existing research predominantly focuses on non-independent and identically distributed data within single-domain scenarios, yet neglects the critical issue of domain shift in multi-domain settings. This oversight causes client models to overfit local data distributions, thereby severely degrading generalization performance across domains. To address this limitation, we propose a **fed**erated learning framework based on **co**ntrastive feature **de**coupling (FedCode). Its core innovation lies in decoupling domain-specific style features and domain-independent semantic features to eliminate task-irrelevant domain-style noise interference, thereby enhancing local model performance while maximizing cross-domain generalization capabilities. Specifically, we

[*]Corresponding author

Email addresses: `shaobozhang@hnust.edu.cn` (Shaobo Zhang), `yijieyin@mail.hnust.edu.cn` (Yijie Yin), `wliang@hnust.edu.cn` (Wei Liang), `wfwufan@csu.edu.cn` (Fan Wu), `weizhi.meng@ieee.org` (Weizhi Meng), `tianwang@bnu.edu.cn` (Tian Wang)

first implement federated Dual Prototype Learning (DPL) at the server side, where client-submitted style prototypes and semantic prototypes undergo differentiated aggregation to construct global domain-style prototypes and global semantic prototypes. DPL provides rich domain knowledge and a purified semantic target as alignment anchors for local optimization, respectively. Subsequently, we enforce Contrastive Feature Decoupling (CFD) optimization mechanism at client devices, compelling both feature types to align with their corresponding global prototypes while achieving explicit style-semantic decoupling. CFD encourages clients to learn domain-invariant pure semantic features for downstream tasks. Experimental results on multi-domain datasets Digit5 and PACS demonstrate that FedCode achieves the highest local accuracy improvement while maintaining the lowest cross-domain performance degradation. Taking PACS for example, compared to the baseline method FedAvg, FedCode achieves an accuracy improvement of 7.52% and simultaneously reduces the cross-domain performance degradation by 5.02%.

Keywords: Federated learning, Domain shift, Feature decoupling, Contrastive learning

1. Introduction

As a distributed collaborative paradigm, Federated Learning (FL) enables multi-source knowledge sharing without exposing raw data, effectively balancing the trade-off between information sharing and privacy security among distributed nodes [1–3]. In the typical method [4], the central server iteratively constructs a global model by averaging aggregation of client models and then redistributes it for next-round local training, attempting to converge to a high-performance global model that adapts to all clients. However, due to persistent data heterogeneity in real-world scenarios, this method leads to inconsistencies between local and global optimization directions [5], which degrade FL model performance and may even hinder convergence [6].

To address data heterogeneity, numerous methods have attempted to enforce alignment between local optimization directions and global objectives by leveraging global knowledge [6–10] or to enhance local optimization through personalized customization [11–14]. However, these approaches suffer from three critical limitations: **(1) Single-domain assumption**. They predominantly assume that client data is sampled from a single homogeneous domain, ignoring domain shift challenges [15, 16] in multi-domain sce-

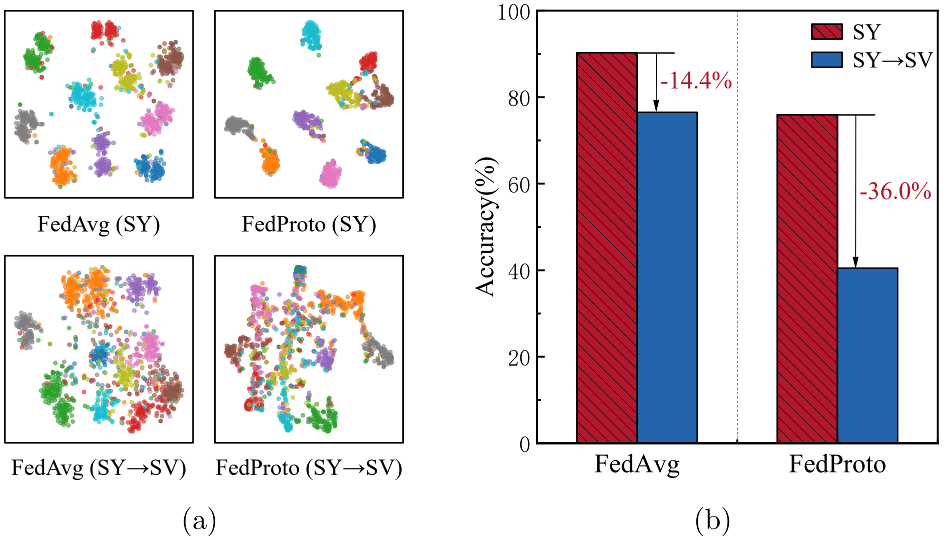


Fig. 1. (a) Feature space misalignment and decision boundary blurring; (b) Cross-domain performance degradation (SY→SV denotes the testing results of client models trained on SYN evaluated on SVHN).

nario—such as stylistic discrepancies between sketches and oil paintings, or weather-related deviations in desert versus rainy urban road imagery. **(2) Naive global knowledge**. They rely on simplistic global knowledge fusion strategies (e.g., linear combinations of model parameters or gradients [10, 14]), which fail to precisely characterize complex cross-domain distributional discrepancies and establish a well-defined semantic optimization objective. **(3) Weak consistency constraints**. They lack robust global consistency constraint mechanisms. Specifically, they exhibit inefficient utilization of global knowledge during local training phases, as exemplified by the L2 regularization terms [9, 12].

Under the three limitations, these methods fail to obtain a domain-invariant global objective and effectively constrain local training to align with it, driving client models to prioritize fitting local data distributions. This is manifested in Fig. 1 as cross-domain feature space misalignment, decision boundary ambiguity, and ultimately severe cross-domain performance degradation. Consequently, existing methods struggle to simultaneously boost client models' local performance and cross-domain generalization capabilities, critically undermining the practical utility of FL. For example, when clients encounter transient out-of-distribution data, poorly generalized models fail to sustain stable inference quality. Thus, devising sophisticated global knowledge representation methods coupled with efficient local optimization mechanisms to jointly enhance the local adaptability and cross-domain generalization capacity of client models has emerged as a critical challenge.

For the purpose of enriching the expression of global knowledge, we consider employing prototypes [9] as additional information carriers. Prototypes, defined as the mean feature vectors of samples within the same class, encode compact yet discriminative class-wise reference knowledge while incurring negligible communication overhead compared to transmitting full model parameters. Classical prototype-based methods like FedProto [9] reduce communication costs by replacing model collaboration with global prototypes aggregated through client prototypes averaging. However, such an averaging operation diminishes the diversity of domain-specific knowledge. To address this issue, FPL [17] constructs class-wise cluster prototypes to capture richer domain variances. Meanwhile, unbiased prototypes (mean of cluster prototypes) are utilized in FPL as a global objective signal. Nevertheless, these approaches suffer from potential drawbacks: client prototypes inevitably entangle label semantics with domain stylistic information, resulting in entangled global prototypes that interfere with clients' ability to learn domain-agnostic semantic features.

Moreover, the lightweight nature of prototypes inherently implies the incompleteness of client knowledge. Due to its sole reliance on prototypes for global knowledge transfer and the use of only class-matched global prototypes to formulate regularization constraints, FedProto fails to effectively guide clients in learning task-related features that remain invariant across domains and clients. This manifests as the phenomenon of the inability to form unified clusters for the same-label samples from different clients in the feature space (see Fig. 4 in Sec. 4.3 and Fig. 2 in [9]). This phenomenon reveals that the client models have fallen into pronounced inductive bias due to local overfitting, which indicates the necessity of the rigid calibration of model parameter sharing and fully exploiting the guidance potential of prototypes. Therefore, to enforce strong consistency constraints on clients and enhance the generalization of client models, we consider preserving model collaboration while constructing an efficient local optimization mechanism by incorporating contrastive learning [18, 19].

To address the issue of performance degradation in client models across domains caused by domain shift in federated learning, we present a **fed**erated learning framework based on **co**ntrastive feature **de**coupling (FedCode). The framework simultaneously enhances local performance and cross-domain generalization capability of client models by acquiring more fine-grained global knowledge through Dual Prototype Learning (DPL) and then fully utilizing it for style-agnostic semantic feature learning through Contrastive Feature Decoupling (CFD). **Firstly**, in DFL, a style-aware encoder and a universal semantic encoder are employed in each client to separately extract independent domain-specific style features and cross-domain invariant semantic features. Further, on the server side, global domain style prototypes are generated via client style prototype clustering, whereas global semantic prototypes are formed by averaging client semantic prototypes. The global domain-style prototypes capture rich domain knowledge, whereas the global semantic prototype establishes a global semantic convergent target. **Secondly**, in CFD, we introduce Semantic Contrastive Learning (SemCL) and Style Contrastive Learning (StyCL) on the client side. SemCL enforces feature alignment for same-label samples while pushing apart features of different ones, thereby establishing a globally consistent semantic space. Similarly, StyCL drives style features from the same domain towards uniformity while distancing those from different domains. Additionally, SemCL and StyCL also enforce the separation between style features and semantic features of samples. Concurrently, we incorporate Feature Decoupling Regularization (FDR), which promotes decoupling by reinforcing the orthogonality between style features and semantic features, thereby preventing style information from interfering with semantic feature learning. **Finally**, the synergistic effect of DPL and CFD enables FedCode to learn cross-domain invariant semantic features for downstream tasks. The main contributions of this paper are as follows:

- We propose DPL, a federated dual prototype learning strategy that differentially aggregates decoupled style and semantic features to construct global domain-style prototypes and global semantic prototypes, providing rich domain knowledge and a domain-invariant semantic target, respectively.

- We present CFD, a contrastive feature decoupling optimization mechanism. By fully leveraging global knowledge through SemCL and StyCL and enforcing orthogonality between style and semantic features through FDR, the mechanism guides the formation of a feature space characterized by intra-class/intra-domain compactness, inter-class/inter-domain separation, and style-semantic independence. CFD encourages clients to learn purified style-agnostic semantic features, thereby enhancing generalization performance while improving local semantic discriminability of client models.

- Extensive experiments on Digit5 [20] and PACS [21] demonstrate that FedCode surpasses baseline methods in both local data adaptability and cross-domain generalization capability. For instance, on PACS, FedCode achieves a maximum accuracy improvement of 7.52% while reducing the cross-domain performance degradation by 5.02%.

The remainder of this paper is organized as follows: Sec. 2 reviews the most related work. Sec. 3 elaborates on the proposed FedCode framework. Sec. 4 validates the effectiveness of FedCode through a series of experiments and analyses. Finally, Sec. 5 concludes the paper.

2. Related Work

In this section, we briefly review the works most relevant to this paper, including data heterogeneous federated learning, federated prototype learning, and disentangled representation learning.

2.1. Data Heterogeneous Federated Learning

Data heterogeneity, which refers to significant discrepancies in the distribution, structure, or attributes of client data, severely limits the performance of classical federated learning methods like FedAvg [4]. To address this issue, traditional approaches such as FedProx [6] and FedDyn [7] suppress local optimization deviations through global penalty terms, while methods like FedLAW [10] and FedDisco [22] pursue a globally optimal solution via reweighted model aggregation. Nevertheless, these methods struggle to obtain a single strongly generalizable global model that adapts to severely heterogeneous clients, trapping client models in suboptimality. To enhance local adaptability, personalized federated learning has been proposed. Methods such as FedAMP [12] and FedALA [23] employ personalized aggregation to tailor client-specific models better adapted to local data, while methods like FedRep [13] and FedBABU [24] balance global knowledge sharing with local personalization by partially sharing decoupled model components.

However, the aforementioned methods predominantly rely on an in-domain heterogeneity assumption (i.e., heterogeneity occurs within a single data domain), and thus fail to address the prevalent domain shift issues in real-world scenarios. Under domain shift, traditional methods struggle to acquire high-quality global knowledge to provide precise semantic optimization constraints, while personalized approaches, due to their overemphasis on local adaptation, exacerbate cross-domain performance degradation. The underlying rationale is that client models in these methods entangle domain-specific style information with domain-agnostic semantic knowledge, making it challenging to reconcile local performance and cross-domain generalization. To tackle this fundamental limitation, we construct orthogonal style and semantic spaces through DPL and CFD, which forces clients to learn domain-invariant pure semantic features to counteract domain shift.

2.2. Federated Prototype Learning

In federated learning, prototypes serves as compact information carriers encoding class-wise reference knowledge. For instance, FedProto [9] replaces the traditional model aggregation paradigm with prototype aggregation, reducing communication costs by an order of magnitude. However, it merely employs L2-distance constraints to align local sample features with class-matched global prototypes, failing to leverage the valuable knowledge of the other class's global prototypes. To fully exploit global knowledge, methods like FedPCL [25], FedFM [26], and FCPLN [27] introduce contrastive loss terms [28] to guide local sample features to align with class-matched global prototypes while repelling class-mismatched ones, effectively enhancing feature space discriminability. Nevertheless, under domain shift, these methods are constrained by the prototype averaging strategy, rendering them incapable of capturing cross-domain style biases. To address the issue, FPL [17] and FGGP [29] construct multiple clustered prototypes for each class to preserve rich domain knowledge. However, by forcing features to align with multiple cluster prototypes via contrastive learning, these methods inherently fail to resolve the implicit coupling between domain-agnostic semantic knowledge and domain-specific style information. To overcome this, our work adopts the DPL strategy to separately learn global semantic prototypes and global domain-style prototypes, providing explicit feature alignment baselines for local optimization.

2.3. Disentangled Representation Learning

Disentangled Representation Learning (DRL) [30], also known as feature decoupling, aims to enhance model explainability and generalizability by separeting task-relevant independent latent representations from data. For example, typical DRL approaches are often based on generative models [31, 32] and have shown significant potential in learning interpretable representations for visual data. Another effective DRL paradigm minimizes Mutual Information (MI) [33] to eliminate statistical dependencies between features. It is yet challenging to accurately estimate MI between high-dimensional continuous variables in deep neural networks. Therefore, recent methods employ neural networks to construct variational MI estimators for differentiable and scalable MI estimation. For instance, Belghazi et al. [34] proposed the Mutual Information Neural Estimator (MINE), which approximates MI through neural estimation of the KL divergence between joint and marginal distributions. Nonetheless, MINE provides only a lower-bound estimate of MI and fails to guarantee its strictly decreasing property. To overcome this drawback, Cheng et al. [35] introduced the Contrastive Log-ratio Upper Bound (CLUB), enabling differentiable upper-bound MI estimation via conditional probability differences between positive and negative sample pairs. However, the CLUB estimator not only introduces additional computational overhead but also suffers from dimensional sensitivity, failing to accurately track the true MI in high-dimensional scenarios (e.g., when dimensions exceed 200). Therefore, we implement feature decoupling through a computationally friendly regularization loss based on normalized inner product, enforcing orthogonal constraints between sample style features and semantic features, which collaborates with contrastive losses to achieve style-semantic disentanglement.

3. Methodology

In this section, we first delineate the preliminaries, subsequently present an overview of the proposed FedCode framework, and finally delve into its two core modules—DPL and CFD. To facilitate readability, Table 1 encapsulates frequently employed notations.

3.1. Preliminaries

Consider a federated learning system that comprises N clients with their private data $D_n = \{(x_i, y_i)\}_{i=1}^{|D_n|}$, where $|D_n|$ denotes the local data size of client$_n$. These clients exhibit domain shift characterized by $P_j(x \mid y) \neq P_k(x \mid y)\,(P_j(y) = P_k(y))$, meaning participants from different domains j and k share identical label space $P(y)$, but possess distinct feature distributions $P(x \mid y)$.

In our framework, client$_n$ possesses a local model $M_n = \{E_n^s, E_n^c, H_n\}$ parameterized by $\theta_n = \{\theta_n^s, \theta_n^c, \theta_n^h\}$, where:

Table 1
Summary of Main Notations

Notation	Description
N	Number of clients.
K	Number of classes.
R	Number of local training epochs.
T	Number of global communication rounds.
D_n	The dataset of the n-th client (client$_n$).
M_n	Local model of client$_n$.
θ_n	Local model parameters of client$_n$.
P^j	The global style prototype of the j-th pseudo domain.
G^k	The global semantic prototype of the k-th class (class$_k$).
s_n	The style prototype of client$_n$.
c_n^k	Client$_n$'s semantic prototype of class$_k$.

- $E_n^s : \mathcal{X} \to \mathcal{Z}_s$ is a style encoder, mapping an input sample x to a d-dimensional style feature $z^s = f(\theta_n^s; x) = E_n^s(x) \in \mathbb{R}^d$ in the style space $\mathcal{Z}_s$.

- $E_n^c : \mathcal{X} \to \mathcal{Z}_c$ is a semantic encoder parallel to E_n^s, mapping an input sample x to a d-dimensional semantic feature $z^c = f(\theta_n^c; x) = E_n^c(x) \in \mathbb{R}^d$ in the semantic space $\mathcal{Z}_c$.

- $H_n : \mathcal{Z}_c \to \mathbb{R}^K$ is a classifier, producing logits $l = f(\theta_n^h; x) = H_n(z^c)$ from the semantic features z^c.

The objective is to simultaneously enhance both the local performance and cross-domain generalization of client models, enabling their adaptation to multiple data domains. The federated learning optimization problem is formulated as:

$$\{\theta_1^*, \theta_2^*, \ldots, \theta_n^*\} = \underset{\{\theta_1, \theta_2, \ldots, \theta_n\}}{\arg\min} \sum_{n=1}^{N} \mathcal{L}_n\left(\theta_n; D\right), \tag{1}$$

where $\mathcal{L}_n$ denotes the loss function of client$_n$.

3.2. The Overview of FedCode

As illustrated in Fig. 2, this paper proposes FedCode to address the pervasive domain shift challenge in distributed scenarios. FedCode aims to simultaneously enhance local adaptability and cross-domain generalization of

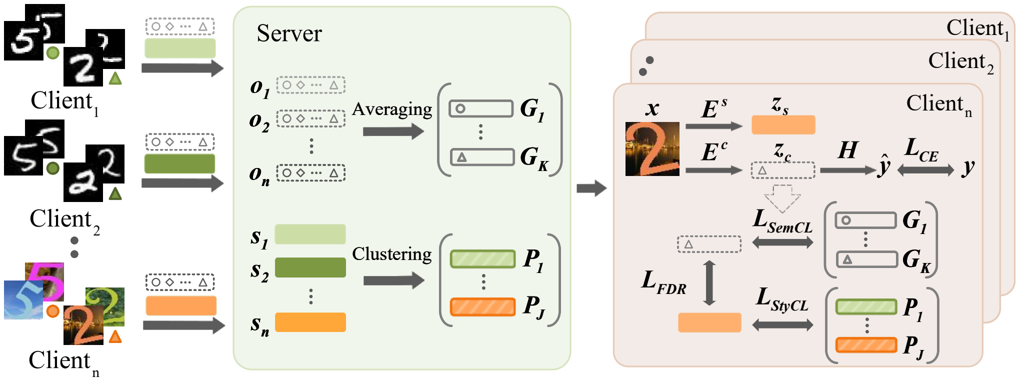


Fig. 2. Illustration of the FedCode model.

client models to mitigate the effects of domain shifts, thereby ensuring stable and accurate processing when the client model encounters out-of-domain data. Specifically, while maintaining baseline performance through semantic encoders sharing, FedCode further improves model capabilities via two core modules: dual prototype learning and contrastive feature decoupling.

The FedCode framework is presented in Algorithm 1, with the workflow outlined as follows.

(1) The server broadcasts the global semantic encoder, global semantic prototype, and global style prototype to clients.

(2) Each client replaces the local semantic encoder with the global semantic encoder and performs local training.

(3) Each client feeds data samples into the style-aware and general semantic encoders in parallel to extract style features and semantic features, respectively.

(4) Each client averages the style features of all its training samples to construct its style prototype, whereas it averages the semantic features by category to derive its semantic prototype.

(5) Each client uploads its local semantic encoder, semantic prototypes, and style prototype to the server.

(6) The server generates the global semantic encoder by aggregating client semantic encoders through data-size-weighted averaging; then com-

Algorithm 1 FedCode

```
Input: Global rounds $T$, local epochs $R$, number of clients $N$, client data sets $\mathcal{D}_n$, client model $\theta_n$, learning rate $\eta$.
Output: The final client models $\{\theta_n\}_{n=1}^{N}$.

Initialize $\theta^0$ and broadcast it to all clients;
for $t = 0, 1, \ldots, T-1$ do
    for $n = 0, 1, \ldots, N-1$ in parallel do
        $c_n^{t+1}, s_n^{t+1}, \theta_n^{c,t+1} \leftarrow$ LocalUpdate$(P^t, G^t, \theta^{c,t})$;
    $G^{t+1} = \{G^{k,t+1} | G^{k,t+1} = \frac{1}{N}\sum_{n=1}^{N} c_n^{k,t+1}\}$;
    $P^{t+1} = \{P^{j,t+1}\}_{j=1}^{J} \xleftarrow{\text{Clustering}} \{s_n^{t+1}\}_{n=1}^{N}$;
    $\theta^{c,t+1} = \frac{1}{N}\sum_{n=1}^{N} \frac{|\mathcal{D}_n|}{|\mathcal{D}|}\theta_n^{c,t}$;

LocalUpdate$(P^t, G^t, \theta^{c,t})$
    $\theta_n^{t+1} \leftarrow \theta_n^t$;
    $\theta_n^{c,t+1} \leftarrow \theta^{c,t}$;
    for $r = 0, 1, \ldots, R-1$ do
        for each batch $b = \{x, y\} \subseteq D_n$ do
            $z^s = f(\theta_n^{s,t+1}; x)$;
            $z^c = f(\theta_n^{c,t+1}; x)$;
            $\mathcal{L}_{SemCL} \leftarrow (z^s, G^t, P^t)$ in Eq. (11);
            $\mathcal{L}_{StyCL} \leftarrow (z^c, G^t, P^t)$ in Eq. (13);
            $\mathcal{L}_{FDR} \leftarrow (z^c, z^s)$ in Eq. (14);
            $\mathcal{L}_{CE} \leftarrow (z^c, y)$ in Eq. (15);
            $\mathcal{L} = \mathcal{L}_{CE} + \alpha(\mathcal{L}_{SemCL} + \mathcal{L}_{StyCL}) + \beta\mathcal{L}_{FDR}$;
            $\theta_n^{t+1} \leftarrow \theta_n^{t+1} - \eta\nabla\mathcal{L}$;
    for $k = 1, 2, \ldots, K$ do
        $c_n^{k,t+1} = \frac{1}{|D_n^k|}\sum_{(x_i,y_i)\in D_n^k} f(\theta_n^{c,t+1}; x_i)$;
    $c_n^{t+1} = [c_n^{1,t+1}, c_n^{2,t+1}, \ldots, c_n^{K,t+1}]$;
    $s_n^{t+1} = \frac{1}{|D_n|}\sum_{(x_i,y_i)\in D_n} f(\theta_n^{s,t+1}; x_i)$;
    return $c_n^{t+1}, s_n^{t+1}, \theta_n^{c,t+1}$.
```

putes the global semantic prototype via mean aggregation of client semantic prototypes; and finally constructs the global style prototype through client style prototypes clustering.

(7) Iterating the above steps until meeting the convergence criteria.

3.3. Federated Dual Prototype Learning

Conventional federated prototype learning methods rely on averaging aggregation of client prototypes to build global prototypes for cross-client knowledge sharing. However, since client data originates from heterogeneous domains, client prototypes inevitably couple label semantics information with domain-specific stylistic information. In this case, on the one hand, the diversity of domain characteristics is diminished by averaging operations; on the other hand, the global class prototypes are contaminated with intertwined stylistic noise, failing to provide pure semantic guidance for local optimization. To overcome these limitations, we decouple client prototypes into client style prototypes and client semantic prototypes. Subsequently, we apply differentiated aggregation strategies to ensure both domain-awareness and semantic discriminability. Specifically, client style prototypes are clustered to preserve domain-specific knowledge, whereas client semantic prototypes are averaging aggregated to construct a cross-domain invariant global semantic optimization target.

3.3.1. Semantic Prototype

Client Semantic Prototype. We employ a common encoder (e.g., ResNet [36]) to extract semantic features from data samples. For client$_n$, the semantic prototype $c_n^k \in \mathbb{R}^d$ for class$_k$ is defined as the mean value of feature vectors belonging to that class:

$$c_n^k = \frac{1}{|D_n^k|} \sum_{(x_i, y_i) \in D_n^k} f(\theta_n^c; x_i), \tag{2}$$

where $D_n^k = \{(x_i, y_i) \mid y_i = k\}_{i=1}^{|D_n^k|} \subset D_n$ denotes the class$_k$ sample set of client$_n$. Semantic prototypes encapsulate class-specific label semantic information to capture discriminative features of each category. We further define the client semantic prototypes as:

$$c_n = \left[c_n^1, c_n^2, \ldots, c_n^K\right] \in \mathbb{R}^{K \times d}. \tag{3}$$

Global Semantic Prototype. Global semantic prototypes are constructed via averaging aggregation of client semantic prototypes on the server side. Owing to the CFD optimization mechanism performed on the client side (Sec. 3.4), the semantic prototypes retain only pure semantic information. This enables them to reliably guide the learning of cross-domain invariant semantic feature during local optimization, thereby improving both generalization and stability of client models. The global semantic prototype G is defined as:

$$\begin{gathered} G^k = \frac{1}{N}\sum_{n=1}^{N} c_n^k \in \mathbb{R}^d, \\ G = \frac{1}{N}\sum_{n=1}^{N} c_n = \left[G^1, \cdots, G^k, \cdots, G^K\right]. \end{gathered} \tag{4}$$

3.3.2. Style Prototype

Style-Aware Instance Normalization Module (SAIN). To amplify the focus on stylistic information, we design a style-aware Instance Normalization (IN) module to strategically replace the commonly used Batch Normalization (BN) layer, thereby constructing the style-aware encoder E_n^s.

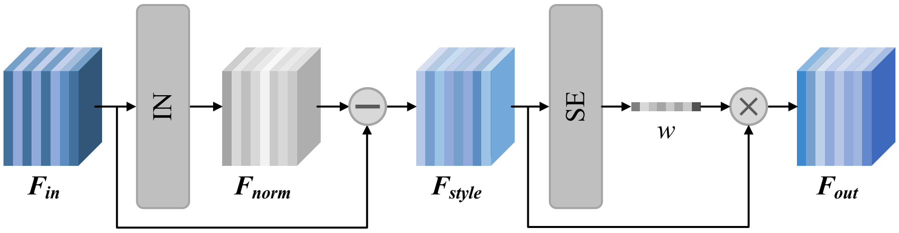


Fig. 3. Illustration of the SAIN module.

As illustrated in Fig. 3, the SAIN module processes the input feature map F_{in} as follows:

Firstly, an IN layer is applied to F_{in} to suppress style-related information (e.g., color, texture), yielding the normalized feature map F_{norm}. Mathematically, for a single sample, let $\mu(F_{in})$ and $\sigma(F_{in})$ denote the channel-wise mean and standard deviation of F_{in}, respectively. The affine parameters γ

and δ are fixed to 1 and 0 in this work. The operations are defined as:

$$F_{norm} = \gamma \cdot \frac{F_{in} - \mu(F_{in})}{\sigma(F_{in})} + \delta. \tag{5}$$

Subsequently, the style feature map F_{style} is derived via residual subtraction:

$$F_{style} = F_{in} - F_{norm}. \tag{6}$$

Finally, we introduce a Squeeze-and-Excitation (SE) module [37], which compresses the feature map into channel descriptors via adaptive average pooling and adaptively learns channel attention weights w through two fully-connected layers. This yields an attention-enhanced style feature map as the output F_{out}:

$$F_{out} = w \cdot F_{style}, \tag{7}$$

where "$\cdot$" denotes channel-wise multiplication.

Client Style Prototype. The client style prototype integrates client-specific stylistic information. For client$_n$, we first employing the style-aware encoder E_n^s to extract style features from data samples. Then, its style prototype $s_n \in \mathbb{R}^d$ is defined as the mean of all local sample style feature vectors:

$$s_n = \frac{1}{|D_n|} \sum_{(x_i, y_i) \in D_n} f\left(\theta_n^s; x_i\right). \tag{8}$$

Global Domain Style Prototype. We employ the FINCH [38] clustering algorithm with cosine distance as the metric to generate several cluster centers as global domain style prototypes, capturing diverse domain-specific styles. Compared to K-Means, FINCH operates in a parameter-free manner, making it particularly suitable for federated learning scenarios where domain labels are unavailable. Let J denote the number of clusters. The global domain style prototype P is defined as:

$$\begin{gathered} \{P^j\}_{j=1}^{J} \in \mathbb{R}^{J \times d} \xleftarrow{\text{FINCH}} \{s_n\}_{n=1}^{N}, \\ P = \{P^1, P^2, \ldots, P^J\}. \end{gathered} \tag{9}$$

where each $P^j \in \mathbb{R}^d$ represents the centroid of the j-th cluster derived from all client style prototypes. According to the clustering results, we assign domain pseudo-labels to corresponding clients for local style contrastive learning (Sec. 3.4.2).

3.4. Contrastive Feature Decoupling Optimization

We propose a CFD optimization mechanism to guide client models in learning cross-domain invariant and purified semantic features, thereby enhancing both their local performance and cross-domain generalization capability. Specifically, we construct three mutually reinforcing loss terms through SemCL, StyCL, and FDR, which fundamentally resolve the coupling problem between label semantics and domain-specific style information.

3.4.1. Semantic Contrastive Learning

SemCL aims to learn style-agnostic, pure semantic features that are both cross-domain invariant and highly generalizable. Specifically, for a sample $(x_i, y_i) \in D_n^k$, we feed it into the semantic encoder E_n^c to obtain its feature vector $z_i^c = f(\theta_n^c; x_i)$. Through contrastive learning, we enforce z_i^c to approach the global semantic prototype G^k of its ground-truth class k and distance from negative prototypes $G^{-k} = G - \{G^k\}$. Furthermore, to ensure that semantic features are devoid of style information, we explicitly incorporate both the global domain style prototype center $\overline{P} = \frac{1}{J}\sum_{j=1}^{J} P^j$ and the corresponding style feature $z_i^s = f(\theta_n^s; x_i)$ as negative samples in contrastive learning. Therefore, the set of negative samples is defined as $\mathcal{N}^c = G^{-k} \cup \{z_i^s, \overline{P}\}$. To construct a contrastive loss term, we first define the temperature-scaled cosine similarity as follows:

$$\mathrm{sim}(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\| \cdot \tau}. \tag{10}$$

Then, the semantic contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{SemCL}} = -\log \frac{\exp\left(\mathrm{sim}(z_i^c, G^k)\right)}{\exp\left(\mathrm{sim}(z_i^c, G^k)\right) + \sum_{n \in \mathcal{N}^c} \exp\left(\mathrm{sim}(z_i^c, n)\right)}. \tag{11}$$

To illustrate the role of $\mathcal{L}_{\mathrm{SemCL}}$ more intuitively, the optimization process for Eq. (11) can be reformulated as:

$$\begin{aligned}
&\min \mathcal{L}_{\text{SemCL}} \\
&= \min \log \left(1 + \frac{\sum_{n \in \mathcal{N}^c} \exp\left(\text{sim}(z_i^c, n)\right)}{\exp\left(\text{sim}(z_i^c, G^k)\right)}\right) \\
&= \min \log \frac{\sum_{n \in \mathcal{N}^c} \exp\left(\text{sim}(z_i^c, n)\right)}{\exp\left(\text{sim}(z_i^c, G^k)\right)} \\
&= \min \left(\log \sum_{n \in \mathcal{N}^c} \exp\left(\text{sim}(z_i^c, n)\right) - \log \exp\left(\text{sim}(z_i^c, G^k)\right)\right) \\
&= \min \sum_{n \in \mathcal{N}^c} \exp\left(\text{sim}(z_i^c, n)\right) \wedge \max \text{ sim}\left(z_i^c, G^k\right).
\end{aligned} \tag{12}$$

As shown in Eq. (12), minimizing $\mathcal{L}_{\text{SemCL}}$ is equivalent to simultaneously maximizing the similarity with the global semantic prototype G^k of the same class, while minimizing the similarities with global semantic prototypes of different classes, global style center $\overline{P}$, and the corresponding style feature. This encourages the semantic feature space to form a well-defined decision boundary and maintain cross-client and cross-domain consistency.

3.4.2. Style Contrastive Learning

StyCL aims to learn semantic-agnostic style features that serve as negative samples for semantic contrastive learning, thereby enhancing feature decoupling effectiveness. In StyCL, we input the sample $(x_i, y_i) \in D_n$ into the style-aware encoder E_n^s to obtain its feature vector $z_i^s = f(\theta_n^s; x_i)$. According to the assigned domain pseudo-label j, we treat the global domain style prototype P^j as the positive sample, while negative samples set is defined as $\mathcal{N}^s = P^{-j} \cup \{z_i^c, \overline{G}\}$, where $P^{-j} = P - P^j$ stands for global domain style prototypes of other domains, $\overline{G} = \frac{1}{K}\sum_{k=1}^{K} G^k$ means global semantic prototype center, and z_i^c the sample's semantic feature. The style contrastive loss is formulated as:

$$\mathcal{L}_{\text{StyCL}} = -\log \frac{\exp\left(\text{sim}(z_i^s, P^j)\right)}{\exp\left(\text{sim}(z_i^s, P^j)\right) + \sum_{n \in \mathcal{N}^s} \exp\left(\text{sim}(z_i^s, n)\right)}. \tag{13}$$

Similarly to Eq. (12), minimizing $\mathcal{L}_{\text{StyCL}}$ encourages the model to learn a style feature space that is orthogonal to the semantic space while enforcing intra-domain compactness and inter-domain dispersion.

3.4.3. Feature Decoupling Regularization

To enforce orthogonality between style and semantic features, we construct a regularization loss based on the normalized inner product (equivalent to cosine similarity) between style feature z_i^s and semantic feature z_i^c for sample $(x_i, y_i) \in D_n$:

$$\mathcal{L}_{\mathrm{FDR}} = \sum_{j}^{d} \frac{\left| z_{i,j}^s \cdot z_{i,j}^c \right|}{\| z_{i,j}^s \| \| z_{i,j}^c \|}, \tag{14}$$

where d denotes the feature dimension. We also employ cross-entropy loss for classification tasks:

$$\mathcal{L}_{\mathrm{CE}} = -\mathbf{1}_{y_i} \log \operatorname{softmax} \left(f \left(\theta_n^h; z_i^c \right) \right). \tag{15}$$

Combining all components, the overall training loss can be written as:

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \alpha \left(\mathcal{L}_{\mathrm{SemCL}} + \mathcal{L}_{\mathrm{StyCL}} \right) + \beta \mathcal{L}_{\mathrm{FDR}}, \tag{16}$$

where α and β are the trade-off hyperparameters. By optimizing $\mathcal{L}$, client models are guided to learn cross-domain invariant semantically purified features, which jointly optimize local task performance and preserve robust generalization capabilities across domains.

4. Experiments

In this section, we first provide a detailed overview of the experimental setup. Subsequently, based on the experimental results, we analyze the local adaptation performance, global generalization performance, ablation study, and hyperparameter impact of the FedCode client model, respectively.

4.1. Experimental Setup

4.1.1. Datasets

We conducted experiments on two multi-domain datasets—Digit5 [20] and PACS [21].

- Digit5 consists of five domains: MNIST, USPS, SVHN, SYN, and MNIST-M, covering 10 classes (digits from 0 to 9). We configured 20 clients: 5 clients for MNIST, 5 for USPS, 3 for SVHN, 3 for SYN, and 4 for MNIST-M. For each domain, we randomly sampled data according to the ratio of the client count of that domain to the total number of clients, and then allocated the samples to clients within the domain randomly (i.e., each client was allocated 1/20 of the specific domain's total data samples on average).

- PACS includes four domains: Photo, Art Painting, Cartoon, and Sketch, containing 7 classes with a total of 9,991 images. We configured 10 clients: 1 client for Photo, 2 for Art Painting, 3 for Cartoon, and 4 for Sketch. For each domain, all data were randomly assigned to the clients within that domain.

4.1.2. BaseLines

We compared the proposed FedCode method with popular baseline approaches, including SOLO (clients train independently without communication or collaboration), FedAvg [4] (AISTATS'17), FedProx [6] (arXiv'18), MOON [8] (CVPR'21), FedProc [39] (FGCS'23), FedProto [9] (AAAI'22), FedPAC [40] (ICLR'23), FedALA [23] (AAAI'23) and FPL [17] (CVPR'23). Among them, FedAvg and FedProx are traditional methods; FedPAC and FedALA are personalized methods; FedProto is a prototype-based approach; while MOON, FedProc, and FPL are contrastive learning-based solutions.

4.1.3. Training Setup

We employed the ResNet-10 [36] model with a feature vector dimension of 512 for all datasets and methods. Experiments were conducted on the Ubuntu 18.04.6 operating system using an NVIDIA GeForce RTX 3090 GPU. We configured the global training rounds to 100, local epochs to 10, and batch size to 64. The SGD optimizer was adopted with the learning rate set to 0.01, the weight decay set to 1×10^{-5}, and the momentum set to 0.9.

4.1.4. Evaluation Metrics

To evaluate the performance of client models, we defined the following metrics (all accuracy mentioned below is the widely used Top-1 accuracy).

- **LTA** (Local Test Accuracy): The mean accuracy across all client models evaluated on their local test sets.

- **ATA** (Average Target Accuracy): The mean test accuracy of client models from clients within a specific domain when evaluated on the test sets of target domain clients (i.e., clients from other domains).

- **GATA** (Global Average Target Accuracy): The average of ATA values, reflecting the average cross-domain generalization performance of all clients.

- **GASA** (Global Average Source Accuracy): The mean test accuracy of each client model evaluated on the test sets of source domain clients (i.e., clients from the same domain as itself, including itself), reflecting the average inner-domain performance of all clients.

- **CPRR** (Cross-domain Performance Retention Ratio): The ratio of GATA to GASA, quantifying the cross-domain performance retention of client models.

According to the above definition, the higher the values of ATA, GATA, GASA and CPRR, the better the overall performance of the algorithm.

4.2. Evaluation of Local Adaptation Performance

Table 2
Comparison of Local Test Accuracy (%).

Method	Digit5		PACS	
	LTA	Δ	LTA	Δ
SOLO	87.42	-4.32	61.66	+2.22
FedAvg	91.74	—	59.44	—
FedProx	91.77	+0.03	59.36	-0.08
MOON	91.97	+0.23	58.91	-0.53
FedProc	92.33	+0.59	59.34	-0.10
FedProto	90.26	-1.48	60.47	+1.03
FedPAC	92.95	+1.21	60.72	+1.28
FedALA	91.63	-0.11	59.32	-0.12
FPL	93.68	+1.94	61.39	+1.95
FedCode	**93.70**	**+1.96**	**66.96**	**+7.52**

We compared the local performance of different algorithms by the average local test accuracy of client models over the last 5 communication rounds. As shown in Table 2, FedCode achieves the best performance. On the Digit5 dataset, FedCode improves accuracy by 1.96% compared to baseline method FedAvg, while on the PACS dataset, the accuracy improvement reaches 7.52%. This indicates that, through the joint effect of DPL and

CFD, FedCode can identify task-relevant semantic information more accurately, thereby enhancing the local adaptation performance of client models. Notably, although SOLO achieved the second-best result on PACS, Table 4 demonstrates that this is already local overfitting.

4.3. Evaluation of Global Generalization Performance

4.3.1. Performance of Client Models

Table 3
Comparison of Generalization Performance on Digit5 (%).

Method	ATA					GATA	GASA	CPRR
	MNIST	MNIST-M	SYN	USPS	SVHN			
SOLO	36.98	51.03	51.63	30.76	42.36	42.55	78.53	54.18
FedAvg	82.68	88.98	89.17	83.83	89.53	86.84	90.87	95.56
FedProx	83.85	89.39	89.77	84.67	90.04	87.55	90.93	96.28
MOON	83.28	89.20	89.28	84.15	89.55	87.09	91.05	95.65
FedProc	84.87	90.11	89.21	85.70	90.42	88.06	91.27	96.49
FedProto	39.94	56.82	57.43	35.85	47.65	47.54	83.40	57.00
FedPAC	83.55	88.78	89.16	84.76	89.46	87.14	91.67	95.06
FedALA	82.81	89.12	89.37	84.40	89.49	87.04	90.80	95.86
FPL	83.23	86.60	88.23	85.04	86.72	85.96	92.55	92.88
FedCode	**88.68**	**92.37**	**92.26**	**89.27**	**93.85**	**91.53**	**93.16**	**98.21**

Table 4
Comparison of Generalization Performance on PACS (%).

Method	ATA				GATA	GASA	CPRR
	Art-Painting	Cartoon	Photo	Sketch			
SOLO	27.52	25.11	19.96	10.03	20.66	57.09	36.18
FedAvg	44.23	43.60	32.13	21.00	35.24	57.93	60.83
FedProx	44.45	43.46	31.63	20.48	35.00	57.67	60.70
MOON	45.79	42.34	33.21	**22.78**	36.03	57.19	63.01
FedProc	48.57	43.65	36.30	22.74	37.81	58.05	65.15
FedProto	27.48	25.15	20.52	9.81	20.74	55.63	37.28
FedPAC	42.45	39.27	27.15	14.28	30.79	58.51	52.62
FedALA	44.06	43.47	33.37	20.95	35.46	57.36	61.83
FPL	40.14	37.15	28.12	16.62	30.51	58.12	52.49
FedCode	**54.62**	**51.36**	**43.24**	21.48	**42.68**	**64.81**	**65.85**

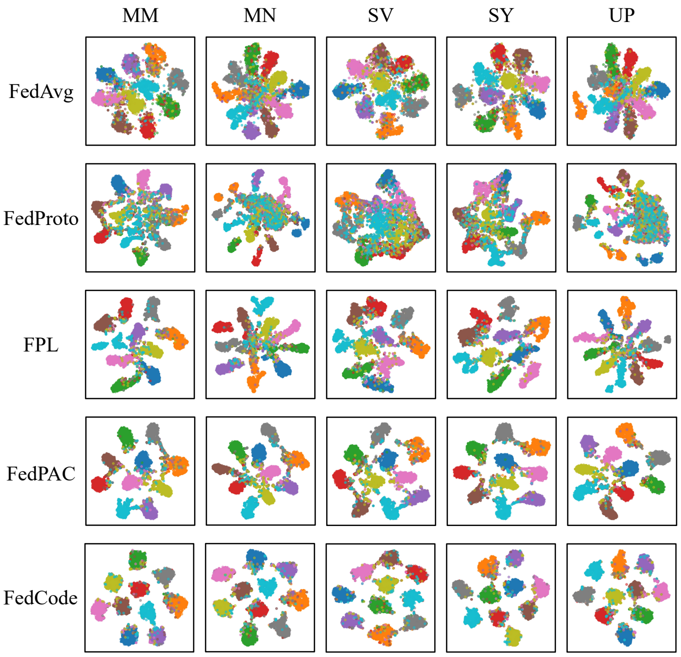


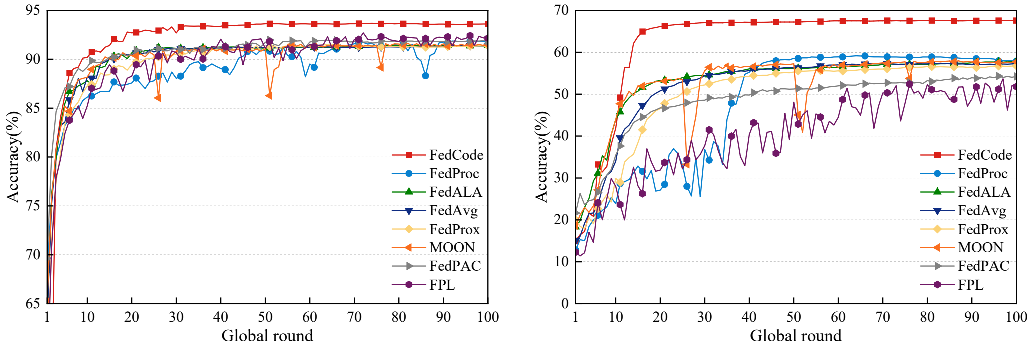
Fig. 4. T-SNE on Digit5.

We calculated the average of the client models' ATA, GATA, GASA, and CPRR metrics over the last 5 communication rounds to compare the generalization performance of client models across different algorithms. As shown in Table 3 and 4, the results demonstrate that FedCode achieves the highest GASA, GATA, and CPRR, significantly outperforming other baselines. Take results on PACS for example, compared to FedAvg, FedCode improves GASA by 6.88% and GATA by 7.44%, and achieves a 5.02% improvement in CPRR (i.e., a 5.02% reduction in cross-domain performance degradation rate). This indicates that FedCode enables client models to handle out-of-domain data more accurately when encountering it temporarily. The reason lies in that DPL provides a cross-domain invariant global semantic target, while CFD strongly constrains each client to optimize towards this target. In contrast, SOLO and FedProto exhibit the most severe cross-domain performance degradation due to their lack of robust global guidance to correct local optimization directions, leading to overfitting on local data. Regarding personalized federated learning methods such as FedALA and FedPAC, they outperform FedAvg in within-domain performance but exhibit lower cross-domain accuracy and performance retention ratios. This stems from their inherent tendency to over-adapt to clients' local data, thereby compromising the global generalization capability of client models under domain shift.

Whether the semantic feature space forms clustered structures with well-defined decision boundaries serves as crucial corroborative evidence in evaluating cross-domain generalization performance. Fig. 4 presents t-SNE visualizations of feature spaces for selected methods. Here, MM, MN, SV, SY, and UP stand for MNIST-M, MNIST, SVHN, SYN, and USPS, respectively. Each column represents the visualization results of the feature spaces generated by client models trained on the corresponding domain when evaluated on all test sets of Digit5. It can be observed that FedeCode exhibits tighter intra-cluster cohesion and sparser inter-cluster separation, with less overlap between different colors (representing features of distinct classes). This further indicates that FedeCode effectively learns pure semantic features that remain invariant across clients and domains.

4.3.2. Performance of Global Model

We present the test accuracy of the global model during training in Fig. 5. For fair comparison, only methods employing a single and complete global model are selected here. It can be observed that FedCode exhibits a stable training process with minimal oscillations, achieves rapid convergence, and attains the highest accuracy.



(a) Digit5. (b) PACS.

Fig. 5. Comparison of Global Model Performance.

4.4. Ablation Study

This experiment evaluates the contributions of each loss component and the SAIN module in FedCode to the model's performance. As shown in Table 5 and Table 6, FedCode exhibits varying degrees of performance degradation when $\mathcal{L}_{\text{SemCL}}$, $\mathcal{L}_{\text{StyCL}}$, or $\mathcal{L}_{\text{FDR}}$ is omitted. As illustrated in Fig. 6, with the SAIN module, FedCode achieves better performance. Overall, the components of FedCode collectively contribute to its optimal performance, which is more pronounced on the more complex dataset PACS. These results indicate that the SAIN module helps capture style information, while the coordination of the three loss terms effectively promotes the learning of semantic features. Furthermore, Fig. 7 confirms that sharing exclusively the semantic feature extractor during global communication is viable, as it can preserve and even promote the performance of the client model.

Table 5
Ablation Study of Loss Terms on Digit5 (%).

Loss Term	LTA	GASA	GATA	CPRR
$\mathcal{L}_{\text{SemCL}} + \mathcal{L}_{\text{FDR}}$	93.69	92.85	91.25	98.28
$\mathcal{L}_{\text{SemCL}} + \mathcal{L}_{\text{StyCL}}$	93.67	92.87	91.14	98.14
$\mathcal{L}_{\text{StyCL}} + \mathcal{L}_{\text{FDR}}$	92.07	90.17	73.55	81.58
$\mathcal{L}_{\text{SemCL}} + \mathcal{L}_{\text{StyCL}} + \mathcal{L}_{\text{FDR}}$	**93.70**	**92.95**	**91.29**	**98.21**

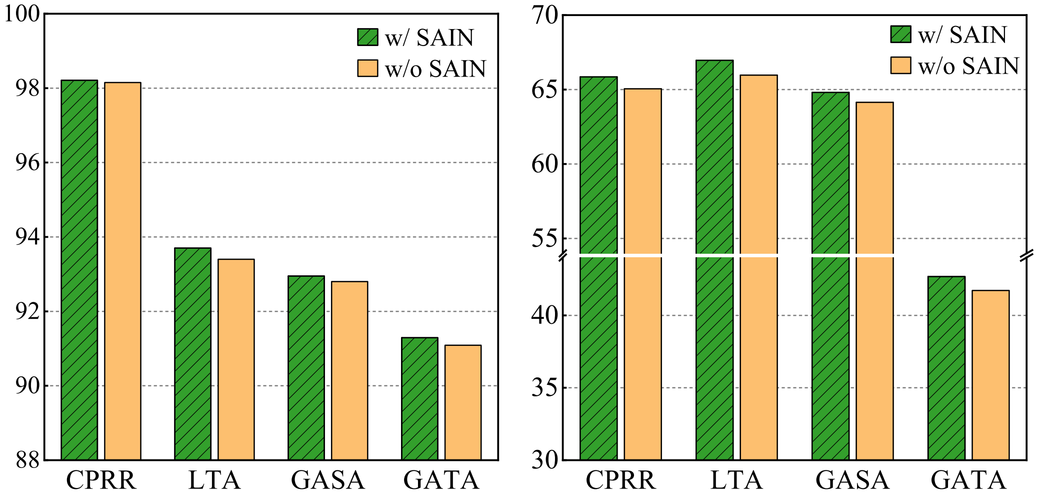
Table 6
Ablation Study of Loss Terms on PACS (%).

Loss Term	LTA	GASA	GATA	CPRR
$\mathcal{L}_{\text{SemCL}} + \mathcal{L}_{\text{FDR}}$	66.50	64.37	41.89	65.08
$\mathcal{L}_{\text{SemCL}} + \mathcal{L}_{\text{StyCL}}$	66.38	64.46	42.37	65.73
$\mathcal{L}_{\text{StyCL}} + \mathcal{L}_{\text{FDR}}$	61.97	59.42	30.42	51.20
$\mathcal{L}_{\text{SemCL}} + \mathcal{L}_{\text{StyCL}} + \mathcal{L}_{\text{FDR}}$	**66.96**	**64.81**	**42.68**	**65.85**

4.5. Impact of the Hyperparameter

4.5.1. Impact of α

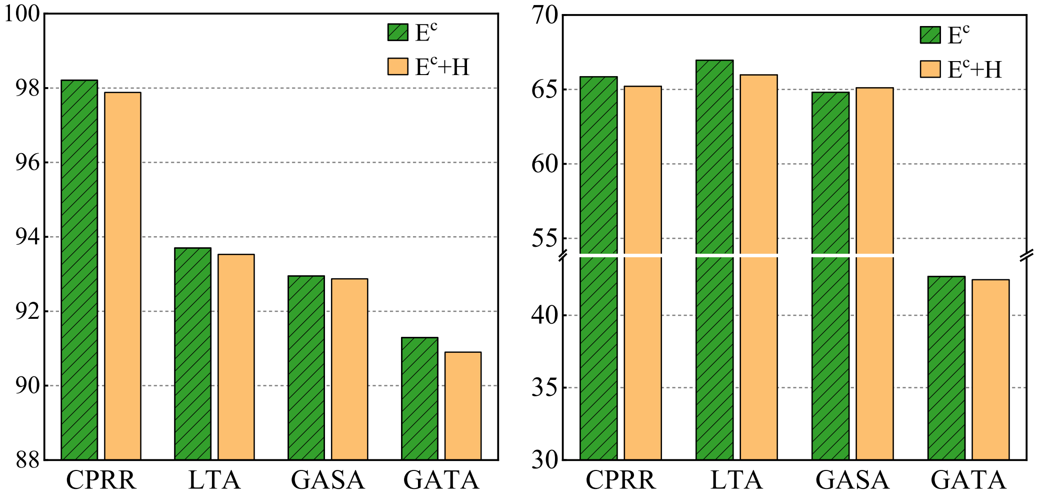
The hyperparameter α controls the strength of both SemCL and StyCL. As α increases, it simultaneously enhances semantic prototype alignment, style alignment, and semantic-style decoupling. We evaluated α's impact across the range $\alpha \in \{0.1, 1, 10, 20, 50, 100, 150\}$. As shown in Fig. 8, Digit5



(a) Digit5.

(b) PACS.

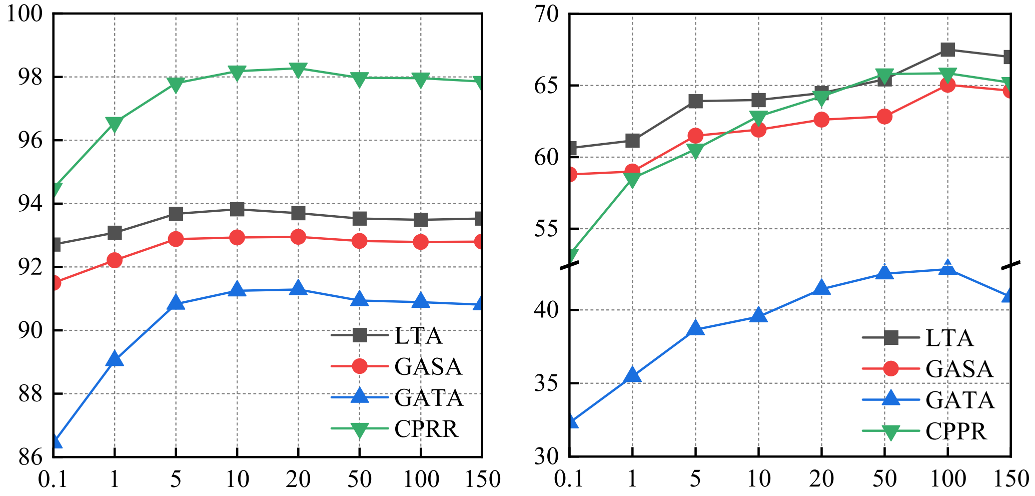
Fig. 6. Effect of the SAIN module.



(a) Digit5.

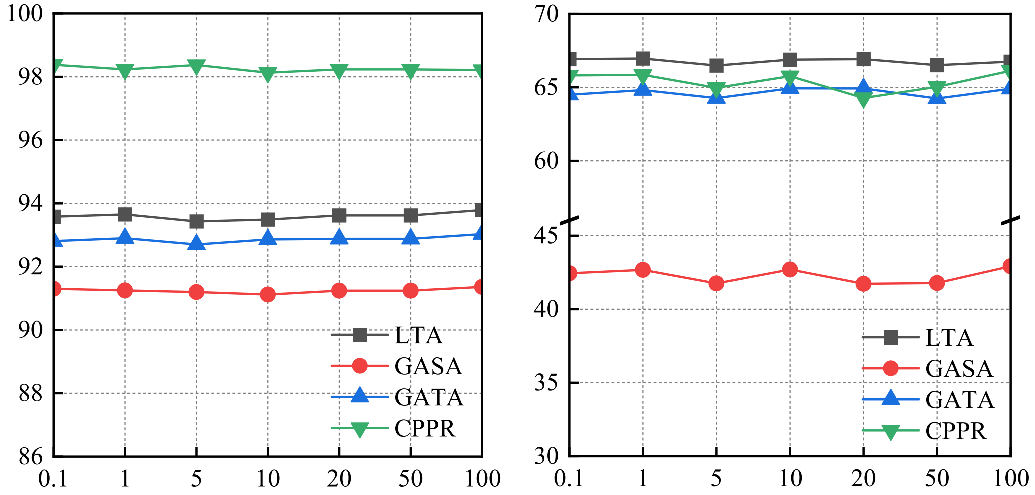
(b) PACS.

Fig. 7. Impact of whether classifier H is shared.



(a) Digit5. (b) PACS.

Fig. 8. Impact of α on client model performance when β is fixed at 1.



(a) Digit5. (b) PACS.

Fig. 9. Impact of β on client model performance when α is fixed at 20 on Digit5 and 100 on PACS.

achieves peak performance when $\alpha > 5$, whereas PACS requires $\alpha = 100$ to reach its optimum. This divergence stems from fundamental differences in inter-domain characteristics and task complexity: Digit5, with its minor style variations and simpler classification tasks, only needs moderate constraints to accomplish cross-domain semantic alignment. In contrast, PACS faces significant domain gaps and high-dimensional visual complexity, necessitating intensive contrastive constraints to suppress style interference and enhance semantic invariance.

4.5.2. Impact of β

β is a hyperparameter controlling the strength of style-semantic orthogonality. We tested the impact of α values on model performance within the range of $\beta \in \{0.1, 1, 5, 10, 20, 50, 100\}$. As shown in Fig. 9, FedCode demonstrates insensitivity to β variations. This is attributed to the established orthogonality between style and semantic features achieved through contrastive learning, coupled with $\mathcal{L}_{\mathrm{FDR}}$ rapidly approaching zero after several training rounds, thereby diminishing the influence of β selection.

5. Conclusion

This paper proposes a federated learning framework based on contrastive feature decoupling (i.e., FedCode) to address the degradation of local performance and cross-domain generalization performance of client models in federated domain shift scenarios. On the server side, FedCode employs dual prototype learning to capture rich domain knowledge while constructing a purified global semantic optimization objective. On the client side, contrastive feature decoupling optimization mechanism fully leverages global knowledge to decouple style and semantic features, thereby eliminating stylistic interference in learning cross-domain invariant semantic features. This approach ultimately enhances local model performance while mitigating generalization degradation. Experimental results validate FedCode's significant advantages in combating domain shifts. While this work focuses on single-modality domain-shift scenarios, systematic exploration of clients with multimodal data remains unexplored. Future efforts will extend the framework to address heterogeneous challenges in federated modality shifts.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Number 62272162, the Hunan Provincial Natural Science Foundation of China under Grant Number 2025JJ50398, and the Postgraduate Scientific Research Innovation Project of Hunan Province, China under Grant Number LXBZZ2024275.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used DeepSeek in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] S. Zhang, Y. Pan, Q. Liu, Z. Yan, K.-K. R. Choo, G. Wang, Backdoor attacks and defenses targeting multi-domain AI models: A comprehensive review. ACM Comput. Surv. 57 (4) (2025), 87.

[2] S. Zhang, W. Chen, X. Li, Q. Liu, G. Wang, APBAM: Adversarial perturbation-driven backdoor attack in multimodal learning, Information Sciences 700 (2025), 121847.

[3] B. Liu, N. Lv, Y. Guo, Y. Li, Recent advances on federated learning: A systematic survey, Neurocomputing 597 (2024), 128019.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.

[5] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, V. Chandra, Federated learning with non-iid data, 2018, arXiv preprint arXiv:1806.00582.

[6] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, V. Smith, Federated optimization in heterogeneous networks, 2020, arXiv preprint arXiv:1812.06127.

[7] D. A. E. Acar, Y. Zhao, R. Matas, M. Mattina, P. Whatmough, V. Saligrama, Federated learning based on dynamic regularization, in: Proceedings of the International Conference on Learning Representations, 2021.

[8] Q. Li, B. He, D. Song, Model-contrastive federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10713-10722.

[9] Y. Tan, G. Long, L. Liu, T. Zhou, Q. Lu, J. Jiang, C. Zhang, Fedproto: Federated prototype learning across heterogeneous clients, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, 2022, pp. 8432–8440.

[10] Z. Li, T. Lin, X. Shang, C. Wu, Revisiting Weighted Aggregation in Federated Learning with Neural Networks," in: Proceedings of the International Conference on Machine Learning, PMLR, 2023, pp. 19767-19788.

[11] C. T. Dinh, N. Tran, J. Nguyen, Personalized federated learning with moreau envelopes, Adv. Neural Inf. Process. Syst. 33 (2020) 21394–21405.

[12] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei and Y. Zhang, Personalized cross-silo federated learning on Non-iid Data, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 7865-7873.

[13] L. Collins, H. Hassani, A. Mokhtari, S. Shakkottai, Exploiting shared representations for personalized federated learning, in: Proceedings of the International Conference on Machine Learning, PMLR, 2021, pp. 2089-2099.

[14] M. Zhang, K. Sapra, S. Fidler, S. Yeung, J. M. Alvarez, Personalized federated learning with first order model optimization, in: Proceedings of the International Conference on Learning Representations, 2021.

[15] W. Huang, M. Ye, B. Du, Learn from others and be yourself in heterogeneous federated learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10143-10153.

[16] L. Zhang, X. Lei, Y. Shi, H. Huang, C. Chen, Federated learning for iot devices with domain generalization, IEEE Internet Things J. 10 (11) (2023) 9622-9633.

[17] W. Huang, M. Ye, Z. Shi, H. Li, B. Du, Rethinking federated learning with domain shift: A prototype view, in: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16312–16322.

[18] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 1597-1607.

[19] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729-9738.

[20] K. Zhou, Y. Yang, T. Hospedales, T. Xiang, Learning to generate novel domains for domain generalization, in: Proceedings of the 16th Computer Vision–ECCV, Springer, 2020, pp. 561–578.

[21] D. Li, Y. Yang, Y.-Z. Song, T. M. Hospedales, Deeper, broader and artier domain generalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5543-5551.

[22] R. Ye, M. Xu, J. Wang, C. Xu, S. Chen, Y. Wang, Feddisco: Federated learning with discrepancy-aware collaboration, in: Proceedings of the International Conference on Machine Learning, PMLR, 2023, pp. 39879-39902.

[23] J. Zhang, Y. Hua, H. Wang, T. Song, Z. Xue, R. Ma, H. Guan, Fedala: Adaptive local aggregation for personalized federated learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, 2023, pp. 11237-11244.

[24] J. Oh, S. Kim, S.-Y. Yun, Fedbabu: Toward enhanced representation for federated image classification, in: Proceedings of the International Conference on Learning Representations, 2022.

[25] Y. Tan, G. Long, J. Ma, L. Liu, T. Zhou, J. Jiang, Federated learning from pre-trained models: A contrastive learning approach, Adv. Neural Inf. Process. Syst. 35 (2022) 19332–19344.

[26] R. Ye, Z. Ni, C. Xu, J. Wang, S. Chen, Y. C. Eldar, Fedfm: Anchor-based feature matching for data heterogeneity in federated learning, IEEE Trans. Signal Process. 71 (2023) 4224-4239.

[27] R. Wang, W. Huang, X. Zhang, J. Wang, C. Ding, C. Shen, Federated contrastive prototype learning: An efficient collaborative fault diagnosis method with data privacy, Knowl. Based Syst 281 (2023), 111093.

[28] F. Wang, H. Liu, Understanding the behaviour of contrastive loss, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2495-2504.

[29] G. Wan, W. Huang, M. Ye, Federated graph learning under domain shift with generalizable prototypes, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, 2024, pp. 15429-15437.

[30] X. Wang, H. Chen, S. Tang, Z. Wu, W. Zhu, Disentangled representation learning, IEEE Trans. Pattern Anal. Mach. Intell. 46 (12) (2024) 9677-9696.

[31] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 2180–2188.

[32] I. Jeon, W. Lee, M. Pyeon, G. Kim, Ib-gan: disentangled representation learning with information bottleneck generative adversarial networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 7926-7934.

[33] X. Hou, Y. Li, S. Wang, Disentangled representation for age-invariant face recognition: A mutual information minimization perspective, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3692-3701.

[34] M. I. Belghazi, A. Baratin, S. Rajeswar, S. Ozair, Y. Bengio, A. Courville, R. D. Hjelm, Mine: Mutual information neural estimation, 2018, arXiv preprint arxiv:1801.04062.

[35] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, L. Carin, Club: A contrastive log-ratio upper bound of mutual information, in: Proceedings of the International Conference on Machine Learning, PMLR, 2020, pp. 1779–1788.

[36] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770-778.

[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132-7141.

[38] S. Sarfraz, V. Sharma, R. Stiefelhagen, Efficient parameter-free clustering using first neighbor relations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8934-8943.

[39] X. Mu, Y. Shen, K. Cheng, X. Geng, J. Fu, T. Zhang, Z. Zhang, Fedproc: Prototypical contrastive federated learning on non-iid data, Future Gener. Comput. Syst. 143 (2023) 93–104.

[40] J. Xu, X. Tong, S.-L. Huang, Personalized federated learning with feature alignment and classifier collaboration, in: Proceedings of the International Conference on Learning Representations, 2023.