# 基于特征投影与自适应增强的联邦蒸馏方法

陈宁江[1,2,3,4]，章德华[1]
（1.广西大学计算机与电子信息学院，广西 南宁 530004；2.广西智能数字服务技术创新中心，广西 南宁 530004；3.广西高校并行分布与智能计算重点实验室，广西 南宁 530004；4.广西人工智能学院，广西 南宁 530004）

**摘要：**为了解决现有联邦蒸馏方法难以处理异构模型间表征空间不一致与特征分布不均的问题，提出一种基于特征投影与自适应增强的联邦异构知识蒸馏框架，实现了跨异构客户端模型的知识高效融合。该框架在服务器端通过蒸馏方式整合客户端输出，在客户端设计轻量化多出口分支，将异构模型的中间特征投影到对齐的logits空间，以突破异构蒸馏中的特征对齐瓶颈。实验结果表明，该方法在标准数据集上取得良好效果，减少了通信轮次和数据传输量，同时增强了系统在异构环境下的鲁棒性，为联邦异构知识融合提供了一种有效的新方案。

**关键词：**联邦学习；知识蒸馏；异构模型；特征投影
**中图分类号：**TP309
**文献标志码：**A
**doi：**10.11959/j.issn.1000

# A Federated Distillation Method Based on Feature Projection and Adaptive Enhancement

Chen Ningjiang[1,2,3,4], Zhang Dehua[1]
1. School of Computer and Electronic Information, Guangxi University, Nanning 530004, China
2. Guangxi Center of Technology Innovation for Intelligent Digital Services, Nanning 530004, China
3. Key Laboratory of Parallel, Distributed and Intelligent Computing, Education Department of Guangxi Zhuang Autonomous Region, Nanning 530004, China
4. Guangxi Academy of Artificial Intelligence, Nanning 530004, China

**Abstract:** To address the problem that existing federated distillation methods struggle to handle inconsistent representation spaces and uneven feature distributions across heterogeneous models, a federated heterogeneous knowledge distillation framework based on feature projection and adaptive enhancement was proposed, achieving efficient knowledge fusion across heterogeneous client models. On the server side, client outputs were integrated via distillation; on the client side, lightweight multi-exit branches were employed to project intermediate features into an aligned logits space to overcome feature alignment bottlenecks. Promising performance was demonstrated on benchmark datasets, with reduced communication rounds and data transmission as well as enhanced system robustness, thus providing an effective new solution for heterogeneous model knowledge fusion in federated learning.

**Key words:** federated learning, knowledge distillation, heterogeneous model, feature projection

## 0 引言

随着人工智能技术的快速发展，分布式机器学习[1]和联邦学习[2]在隐私保护与高效协作建模方面展现出重要价值。联邦学习通过让多个客户端在本地保留数据并协同训练共享模型，有效避免了数据

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX
通信作者：章德华，2313394042@st.gxu.edu.cn
基金项目：国家自然科学基金资助项目（No. 62162003）；广西重点研发计划项目（GuikeAB25069258, GuikeAB25069130）
**Foundation Items**：The National Natural Science Foundation of China (No.62162003), The Guangxi Key Research and Development Program Project（GuikeAB25069258, GuikeAB25069130）

集中带来的隐私与合规风险。然而，经典的联邦学习范式[3]主要依赖模型参数的直接聚合，该策略要求各客户端使用相同的模型架构。为突破模型同构的限制，联邦蒸馏（Federated Distillation, FD）[4]被提出用于替代传统的参数聚合方法。

然而，现有联邦蒸馏方法主要面向同构模型设计，假设各客户端模型在特征学习空间上具有可比性。这在异构场景下往往难以满足：不同架构的模型在中间表征形式、特征抽象能力以及 logits 分布上均存在显著差异，使得单纯依赖输出蒸馏难以充分传递异构模型中的有效知识。具体而言，卷积神经网络（Convolutional Neural Network, CNN）的局部卷积特征、Transformer 的全局自注意力特征与多层感知机混合器（Multi-Layer Perceptron Mixer, MLP-Mixer）的令牌-通道混合特征在维度、统计分布及语义密度上并不对齐，直接最小化 logits 层面的库尔贝克-莱布勒（Kullback-Leibler, KL）散度等价于强行对齐本不兼容的表示空间，导致“教师-学生”映射关系模糊，学生网络难以复现教师决策边界。此外，异构模型之间的表征偏置差异会在蒸馏过程中累积，导致知识融合效率下降、全局模型性能不稳定及收敛过程不鲁棒。这些问题表明，如何实现跨架构的特征对齐、去偏置知识整合以及高效的多模型协同，仍是当前联邦蒸馏亟待解决的关键挑战[5-6]。

针对上述问题，本文提出了一种结合特征投影与自适应目标增强的联邦异构知识蒸馏框架（One-for-All Federated Distillation, FedOFA），旨在实现客户端异构模型间更高效、更稳定的知识融合。该框架不依赖模型参数对齐，而是通过中间特征投影与分布增强的方式改善不同架构之间的知识传递质量，从而提升联邦学习在异构环境中的适应性。本研究的核心贡献包括以下几点：

1) 特征投影：通过为学生模型增设额外出口分支，将异构模型间错位的中间特征表示映射至低架构特异性的 logits 空间，再通过分支输出与教师分类层输出在该空间的匹配，实现了跨架构的中间层知识有效迁移。

2) 自适应目标增强：针对异构模型因归纳偏置差异导致预测分布不同的问题，对传统蒸馏损失进行改进并引入调制参数。依据教师的预测置信度，自适应强化目标类信息，同时降低 logits 中无关信息对学生模型的干扰，提升了异构架构蒸馏的鲁棒性与有效性。

3) 隐私与效率平衡：仅传输公共数据集的 logits，避免私有数据与完整模型参数泄露，同时降低通信开销。

通过将上述方法有机结合，本研究构建了一个既能处理异构模型架构，又能高效融合客户端知识的联邦学习系统。该系统不仅能够提升联邦学习系统的鲁棒性和适应性，还能充分利用异构客户端模型的优势，实现高效的模型融合和知识传递。

在本文中，将详细介绍该框架的设计与实现，并通过一系列实验验证其在不同数据集和模型架构下的有效性。本文其余部分组织如下。第一节简要介绍了联邦学习、知识蒸馏等相关工作，第二节详细描述了算法实现与设计，在第三节中，进行了实验来验证算法的性能，第四节对本文进行总结。

## 1 相关工作

### 1.1 联邦学习

联邦学习作为分布式机器学习范式，通过本地训练、共享模型参数而非原始数据，有效解决了数据隐私与安全问题[2]。经典的联邦学习平均算法通过加权平均客户端模型参数来更新全局模型[3]。算法的全局模型更新公式为：

$$w^{t+1} = \sum_{k=1}^{K} \frac{n_k}{N} w_k^t \tag{1}$$

其中，$w_k^t$是客户端$k$在迭代$t$时的模型参数，$n_k$是客户端$k$的数据量，$N$是所有客户端数据量的总和。

在实际应用中，客户端终端异构导致模型架构难以统一，参数聚合在结构不一致时无法执行，易引发性能劣化甚至训练失败[7]。为应对模型异构问题，已有研究尝试从不同方向提出解决方案。一类方法通过限制模型结构的可变性或引入可对齐的子网络以确保参数空间的一致性[5]。另一类摒弃参数聚合，采用基于输出层的知识蒸馏实现模型融合，绕过结构不匹配问题[4]，为异构联邦学习提供了核心技术路径。

### 1.2 知识蒸馏

知识蒸馏作为一种模型压缩和迁移学习的技术，旨在训练一个较小的学生模型，使其能够模仿更大、更复杂的教师模型[8]。经典的蒸馏损失函数结合了交叉熵损失（Cross Entropy Loss, CE）和

KL 散度损失，公式如下：

$$\mathcal{L}_{\mathrm{KD}} = \alpha \cdot \mathcal{L}_{\mathrm{CE}}(y, \hat{y}_s) + (1 - \alpha) \cdot \mathcal{L}_{\mathrm{KL}}(\mathrm{softmax}(\hat{y}_t/T), \mathrm{softmax}(\hat{y}_s/T)) \quad (2)$$

其中，$\mathcal{L}_{\mathrm{CE}}$是交叉熵损失，$\mathcal{L}_{\mathrm{KL}}$是KL散度，$\hat{y}_s$和$\hat{y}_t$分别是学生模型和教师模型的输出，$T$是温度参数，$\alpha$是平衡参数。

在同构模型环境中，教师与学生共享类似的表示空间，使得 logits 或特征可以直接对齐。然而，当蒸馏用于异构模型时，不同架构的特征表达方式、归纳偏置和抽象层次均不一致，导致直接蒸馏可能出现信息偏移甚至负迁移[9]。因此，在模型结构差异显著的情况下，如何构建稳定可靠的蒸馏机制，是知识迁移研究中的重要课题，为联邦学习中的异构模型融合提供了理论基础[10]。

### 1.3　联邦蒸馏

联邦蒸馏将知识蒸馏引入联邦学习，既支持客户端个性化异构架构，又能降低通信与计算开销，同时通过教师模型知识引导提升学生模型泛化能力，缓解数据异构问题。

早期研究聚焦于通信效率与基础异构支持。Li 等人[4]提出的 FD 通过在服务器端对客户端 logits 进行平均聚合，并在本地以蒸馏方式替代参数更新，实现无参数共享的联邦优化。Lin 等人[11]引入集成蒸馏，利用公共代理数据融合异构客户端输出，缓解 FedAvg 在异构场景的性能退化。但这类方法依赖输出层蒸馏，默认 logits 可直接对齐，仅适用于架构差异较小的场景，在 CNN、Transformer 等强异构场景下，logits 中的架构偏置会导致知识迁移效率大幅下降。

为进一步提升联邦蒸馏在复杂场景下的适用性，后续研究从通信效率、异构适配、安全防护等方向展开探索。在通信效率优化方面，郑晶晶等人[12]提出 FedASHP 方法，通过自适应稀疏策略实现高性能、低开销的联邦通信，为联邦蒸馏的轻量化部署提供了重要参考；在异构模型适配方面，FedDW[13]基于一致性优化与权重蒸馏实现异构联邦学习，进一步拓展了联邦蒸馏的架构兼容性；在安全防护层面，针对联邦原型学习的特征图中毒攻击与双重防御机制[14]的研究，为联邦蒸馏系统的安全鲁棒性设计提供了理论支撑。

近年来，研究者逐渐认识到：联邦蒸馏在异构场景下的核心瓶颈是模型间表征空间的系统性不一致。尽管部分工作尝试通过原型对齐[15]、表示分支解耦[16]或元表示学习[17]缓解该问题，但多数方法仍侧重于表示一致性约束，缺乏对异构表征偏置的显式消解机制。在客户端架构差异较大时，其蒸馏效果仍显著下降[18]。为此，本文聚焦跨异构架构的稳健知识对齐问题，构建可弥合表征差异的异构联邦蒸馏机制，突破现有方法的性能瓶颈。

## 2　算法设计

### 2.1　算法框架

本文提出的联邦异构知识蒸馏框架如图 1 所示，框架由中央服务器与多个异构客户端构成核心架构，专门面向客户端模型完全异构（兼容 CNN、Transformer、MLP 等多元架构）及数据非独立同分布（Non-Independent and Identically Distributed, Non-IID）[19]的应用场景，通过特征投影与自适应蒸馏的协同设计，实现跨架构高效知识迁移。

客户端可依据本地硬件资源与任务需求，自主选取适配的模型架构（如轻量级 MobileNet、深度卷积网络 ResNet、Vision Transformer 及 MLP-Mixer 等），采用“基础模型+多阶段出口分支”结构，通过极简投影模块将异构中间特征映射至统一 logits 空间。本地训练基于私有数据集完成教师模型训练后，加载服务器下发的公共数据集生成对应 logits 与预测置信度，仅将该部分信息上传至服务器，全程不泄露私有数据及完整模型参数，兼顾隐私安全与通信效率。

服务器端承担公共数据集管理、客户端 logits 聚合及高质量蒸馏目标生成等核心职责。接收各客户端上传的 logits 后，采用数据量加权聚合策略，根据客户端私有数据集样本数量占比分配权重，凸显数据充足客户端的知识贡献；同时结合温度校准与自适应增强损失函数，动态调整目标类信息权重，抑制非目标类噪声干扰，生成兼具可靠性与通用性的蒸馏目标。服务器通过轻量化通信协调，即可实现全局知识的有效统一，指导各客户端完成本地蒸馏优化。

### 2.2　客户端操作流程

#### 2.2.1　*客户端操作流程*

客户端可以根据自身设备的计算能力和存储资源选择不同的模型架构。例如，资源受限的移动设

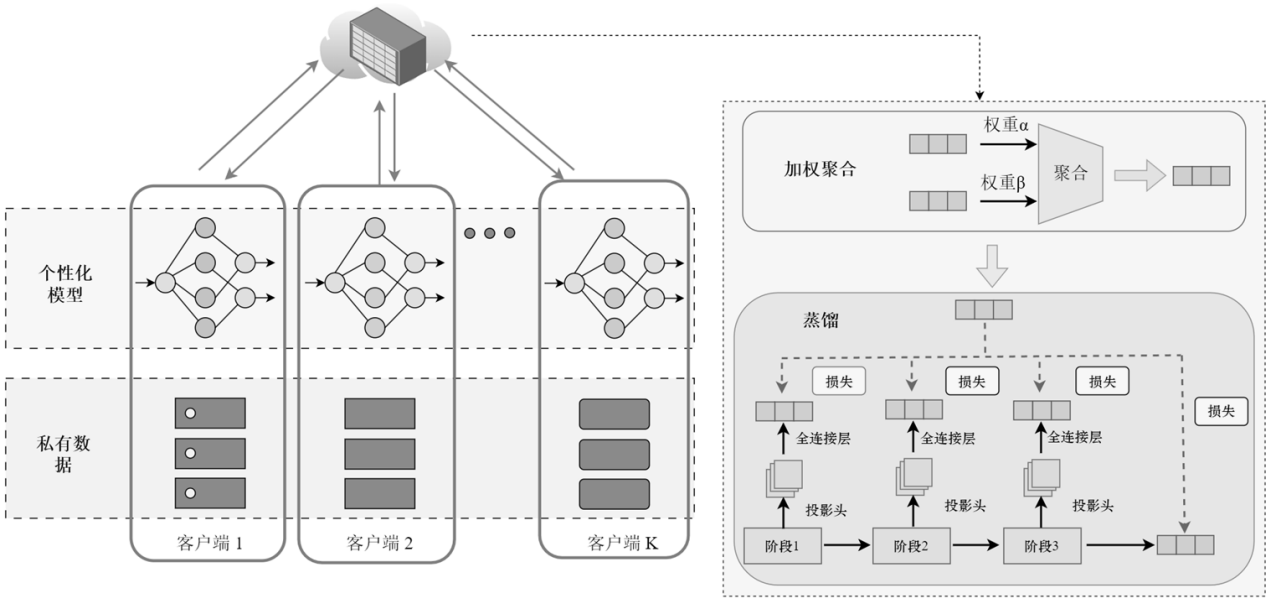



图1　本文方法框架

备可以选择轻量级的模型，如 MobileNet 或 MLP-Mixer，这些模型在保持较高性能的同时，对计算资源的需求较低。而对于计算资源丰富的设备，如服务器或高性能计算设备，可以选择更复杂的模型，如 ResNet 或 Vision Transformer。这种灵活性使得联邦学习系统能够适应各种不同的设备和应用场景，同时最大化利用每个客户端的资源。设预定义模型集合为 $M=\{M_0,M_1,\cdots,M_{m-1}\}$，客户端索引为 $i$，则客户端 $i$ 选择的模型为：

$$M(i)=M_{i \bmod m} \tag{3}$$

其中 $i \bmod m$ 表示对 $m$ 取模的非负余数（范围 $0,\cdots,m-1$）。

#### 2.2.2　客户端本地训练

客户端的本地训练过程是联邦学习的核心环节之一。每个客户端在本地数据上独立训练其模型，这些数据是 Non-IID 的，可能包含不同的类别分布和数据特征。客户端基于私有数据集训练教师模型，优化目标为标准分类损失：

$$\mathcal{L}_{\text{teacher}}=\mathcal{L}_{\text{CE}}(y_t,y) \tag{4}$$

其中，$y_t$ 为教师模型输出，$y$ 为私有数据标签。

客户端使用标准的优化算法[20]对模型进行训练，并在每个训练周期后更新模型参数。此外，客户端还可以根据需要调整训练参数，如学习率和训练周期数，以优化本地模型的性能。

#### 2.2.3　logits 生成与上传

训练完成后，客户端加载公共数据集（与任务同分布，如 CIFAR-100 公共子集、Tiny-ImageNet 验证集子集）。客户端会利用本地已训练完备的教师模型，对该公共数据集进行前向推理，完整生成模型在公共数据上的 logits 输出，也就是模型在全类别上未经 Softmax 归一化的原始预测分数，生成教师模型的 logits 输出：

$$\text{logits}_k=f_k(D_{\text{pub}}) \tag{5}$$

其中，$f_k$ 为客户端 $k$ 的教师模型，$D_{\text{pub}}$ 为公共数据集。客户端仅上传 $\text{logits}_k$ 至服务器，不共享私有数据或模型参数，保障数据隐私。

### 2.3　服务器端操作流程

#### 2.3.1　模型融合

由于各客户端教师模型的架构异构性，其输出 logits 不仅包含任务相关的类别预测信息，还隐含着各自架构专属的归纳偏置特征，如 CNN 的局部纹理依赖、Transformer 的全局依赖编码。直接将异构 logits 用于蒸馏会导致学生模型学习到冗余的架构特异性信息，降低知识迁移效率。因此，服务器端需先对上传的 logits 进行一致性聚合，消除架构差异带来的干扰，生成统一、纯净的蒸馏目标。服务器通过计算客户端模型在公共数据上的 logits 输出，并将其与全局模型的输出进行对比，从而更新全局模型的参数。这一过程通过多次迭代优化，

逐步提升全局模型的性能。

聚合策略采用数据量加权平均，核心思想是让数据量更充足的客户端贡献更高权重——数据量越大的客户端，其教师模型在私有数据集上的训练越充分，logits 中包含的任务相关知识更可靠。具体公式如下：

$$\text{logits}_{\text{agg}} = \sum_{k \in S_t} \frac{n_k}{N_t} \cdot \text{logits}_k \tag{6}$$

其中，$S_t$ 为第 $t$ 轮通信中参与训练的客户端集合，$n_k$ 表示客户端 $k$ 的私有数据集样本数量，$N_t = \sum_{k \in S_t} n_k$ 为所有参与客户端的总样本数量，$\text{logits}_k$ 为客户端 $k$ 的教师模型在公共数据集上生成的输出 logits。通过该加权聚合操作，能够融合多个异构教师的互补知识，生成兼具可靠性与通用性的聚合 $\text{logits}_{\text{agg}}$，为后续全局学生模型的蒸馏提供高质量目标。

### 2.3.2　异构架构蒸馏

为了应对客户端模型架构的异构性，服务器端采用了全适配知识蒸馏（One-for-All Knowledge Distillation, OFA-KD）框架。该框架通过将客户端模型的中间特征投影到对齐的 logits 空间，解决了不同架构模型之间的特征对齐问题。具体来说，客户端模型中引入了额外的退出分支，这些分支将中间特征映射到 logits 空间，从而去除了架构特定的信息。此外，为了减少无关信息的干扰，服务器端还引入了自适应目标增强机制。通过在蒸馏损失函数中加入调节参数 $\gamma$，服务器能够根据教师模型的预测置信度动态调整目标信息的权重，从而增强目标类别的信息，提高蒸馏效果。

（1）特征投影分支：异构特征空间对齐

投影分支采用“1 层全连接层+激活函数”的极简设计，在保障对齐效果的同时控制模型复杂度与计算开销。其核心功能是将学生模型各阶段的中间特征映射至与聚合 logits 维度一致的空间，剔除特征中潜在的架构特异性冗余信息，保留任务相关的共性知识。具体映射过程如下：

$$\text{feat}^{\,i}_{\text{proj}} = \text{ReLU}\left(\text{FC}_i\left(\text{feat}^{\,i}_{\text{stu}}\right)\right) \tag{7}$$

其中，$\text{feat}^{\,i}_{\text{stu}}$ 表示全局学生模型第 $i$ 阶段的中间特征（维度为 $B \times D_i$），$B$ 为批次大小，$D_i$ 为第 $i$ 阶段特征维度），$\text{FC}_i$ 为第 $i$ 个投影分支的全连接层（参数维度为 $D_i$，$C$ 为任务类别数，与聚合 logits 维度一致），$\text{feat}^{\,i}_{\text{proj}}$ 为投影后的对齐特征（维度为 $B \times C$）。通过多阶段投影设计，能够让学生模型在不同特征抽象层级均受到聚合 logits 的监督，避免仅依赖最终输出导致的中间层知识缺失。

（2）OFA 蒸馏损失：自适应目标信息增强

聚合 logits 虽经过加权融合，但由于各客户端教师模型的归纳偏置差异，其预测分布中仍可能包含少量无关噪声。传统蒸馏损失对所有类别信息一视同仁，会导致学生模型学习到冗余噪声，降低泛化性能。为此，基于 OFA-KD 的核心思想，设计自适应增强损失，聚焦目标类信息的精准迁移，抑制非目标类噪声干扰。

OFA 蒸馏损失的核心逻辑是：根据聚合 logits 对目标类的预测置信度，动态调整目标类信息的蒸馏权重——教师对目标类的置信度越高，该类信息的可靠性越强，蒸馏权重越大；反之则自动降低权重，避免不可靠信息的误导。损失函数具体定义如下：

$$\mathcal{L}_{\text{OFA}} = -(1 + p_{\hat{c}}^{agg})^{\gamma} \log\left(p_{\hat{c}}^{\text{stu}}\right) - \mathbb{E}_{c \neq \hat{c}}\left(p_c^{\text{agg}} \log\left(p_c^{\text{stu}}\right)\right) \tag{8}$$

其中各参数说明如下：

$\hat{c}$ 为样本的真实目标类别（由公共数据集标签确定），确保蒸馏过程聚焦于任务核心目标；

$p_{\hat{c}}^{\text{agg}}$ 为聚合 logits 对目标类 $\hat{c}$ 的预测置信度（即 $\text{softmax}\left(\text{logits}_{\text{agg}}\right)_{\hat{c}}$），反映异构教师对目标类的共识程度；

$p_{\hat{c}}^{stu}$ 为全局学生模型对目标类 $\hat{c}$ 的预测概率（即 $\text{softmax}\left(\text{logits}_{\text{stu}}\right)_{\hat{c}}$），需与聚合置信度保持一致；

$\gamma$ 为自适应调节参数，控制目标类信息的增强强度，其取值由聚合 logits 的全局置信度均值 $\bar{p}$ 动态确定：

当 $\bar{p} > 0.8$（教师共识性强），取 $\gamma = 1.0$；

当 $\bar{p} < 0.6$（教师存在分歧），取 $\gamma = 1.5$ 以增强目标类约束；

当 $0.6 \leqslant \bar{p} \leqslant 0.8$，采用线性插值实现平滑过渡：

$$\gamma = 1.0 + 0.5 \times \frac{0.8 - \bar{p}}{0.2} \tag{9}$$

其中阈值 0.6、0.8 来自 CIFAR-10、CIFAR-100、Tiny-ImageNet 的多数据集参数敏感性实验，为跨

场景稳定的最优分界值。

第二项 $\mathbb{E}_{c \neq c^*}(p_c^{\text{agg}} \log(p_c^{\text{stu}}))$ 为非目标类的蒸馏损失，按单样本内非目标类别数取平均（对当前样本除真实类别 $c^*$ 以外的所有类别计算均值），保留教师对非目标类的区分性知识，同时避免其权重过大干扰目标类学习。

（3）总损失函数与模型更新

为兼顾“特征空间对齐”与“目标信息增强”两大目标，将 logits 对齐损失与 OFA 蒸馏损失加权融合，构成服务器端的总蒸馏损失：

$$\mathcal{L}_{\text{total}} = \alpha \cdot \sum_{i=1}^{4} (\text{feat}_{\text{proj}}^{i}, \text{logits}_{\text{agg}}) + (1 - \alpha) \cdot \sum_{i=1}^{4} \mathcal{L}_{\text{OFA}}^{i} \tag{10}$$

其中，$\alpha$ 为损失平衡系数，用于调节两类损失的贡献权重；$\mathcal{L}_{\text{OFA}}^{i}$ 为第 $i$ 阶段的 OFA 蒸馏损失，实现各层级目标信息的自适应增强；MSE（均方误差）用于度量客户端特征投影输出与服务器聚合 logits 的对齐程度，为逐样本、逐类别计算的均方误差，具体定义为：

$$\text{MSE}(\text{feat}_{\text{proj}}^{i}, \text{logits}_{\text{agg}}) = \frac{1}{B \cdot C} \sum_{b=1}^{B} \sum_{c=1}^{C} (\text{feat}_{\text{proj},b,c}^{i} - \text{logits}_{\text{agg},b,c})^2 \tag{11}$$

式中 $B$ 表示批次大小，$C$ 为类别总数，$\text{feat}_{\text{proj},b,c}^{i}$ 与 $\text{logits}_{\text{agg},b,c}$ 分别表示第 $i$ 阶段投影特征与聚合 logits 在第 $b$ 个样本、第 $c$ 个类别上的取值。

全局学生模型的参数更新采用 AdamW 优化器，通过梯度反向传播最小化总损失，具体更新公式如下：

$$w_{\text{stu}}^{j} = w_{\text{stu}}^{j-1} - \eta \cdot \nabla_{w_{\text{stu}}} \mathcal{L}_{\text{total}} \tag{12}$$

其中，$w_{\text{stu}}^{j}$ 表示第 $j$ 次蒸馏迭代后的学生模型参数，$\eta$ 为学习率（实验中设为 $5e^{-5}$），$\nabla_{w_{\text{stu}}} \mathcal{L}_{\text{total}}$ 为总损失对学生模型参数的梯度。通过多轮蒸馏迭代，学生模型能够逐步吸收异构教师的共性知识，实现性能提升。

### 2.4 算法流程

本文提出的基于特征投影与自适应增强的联邦蒸馏方法的伪代码如算法 1 所示。

**算法 1** 基于特征投影与自适应增强的联邦蒸馏方法

输入 $w_{\text{stu}}^{0}$、$T$、$N$、$D_{\text{pub}}$、$f_k$

输出 $w_{\text{stu}}^{T}$、$\text{logits}_k$

服务器端流程

1) $w_{\text{stu}} = w_{\text{stu}}^{0}$
2) 下发 $\mathcal{D}_{\text{pub}}$
3) for $t = 1 \rightarrow T$ do
4) $S_t$ = 随机选择($C \cdot K$)
5) for each client $k \in S_t$ in parallel do
6) 接收 $\text{logits}_k$
7) end for
8) 根据式(6)生成 $\text{logits}_{\text{agg}}$
9) for $j = 1 \rightarrow N$ do
10) $d$=采样($D_{\text{pub}}$)
11) proj =投影前向($w_{\text{stu}}, d$)
12) $\mathcal{L}$ =总损失(proj,$\text{logits}_{\text{agg}}$)
13) $w_{\text{stu}}$ =更新($w_{\text{stu}}$,$\mathcal{L}$)
14) end for
15) end for
16) return $w_{\text{stu}}^{T}$

客户端流程

1) $f_k$ = 选择模型()
2) 加载 $D_k$、$D_{\text{pub}}$
4) for epoch=1$\rightarrow E$ do
5) for batch($x, y$) $\in D_k$ do
6) $\text{logits}_t = f_k(x)$
7) $\mathcal{L}$ = CE($\text{logits}_t, y$)
8) $f_k$ = 更新($f_k$,$\mathcal{L}$)
9) end for
10) end for
11) $\text{logits}_k = f_k(\mathcal{D}_{pub})$
12) 上传 $\text{logits}_k$
13) return $\text{logits}_k$

## 3 实验与分析

### 3.1 实验设置

（1）数据集设置

为了验证该方案的有效性，本文将在多个标准数据集上进行实验，包括 FMNIST、CIFAR10、CIFAR-100、Tiny-ImageNet。

FMNIST 为 MNIST 的服饰类替代数据集，格式相同但含 10 类服饰物品图像，能更有效评估模型的特征区分与泛化能力。

CIFAR10 是 3 通道 RGB 彩色图像数据集，含 6 万张 32×32 图像（5 万训练、1 万测试），涵盖 10 类日常物体，用于低分辨率彩色图像分类任务评估。

CIFAR-100 是 CIFAR10 的扩展，含 100 个细分类别（归为 20 个超类），图像规格相同，用于检验模型在复杂分类任务中的性能。

Tiny-ImageNet 是一个适用于图像分类与迁移学习研究的轻量级数据集，包含 1000 个类别、约 10 万张 64×64 像素彩色图像，是 ImageNet 的简化版本，旨在降低模型训练门槛并支持快速实验验证。

（2）数据异构设置

在客户端和服务器的设置上，本文模拟了一个典型的联邦学习环境。客户端数据分布采用 Dirichlet 分布[21]来控制 Non-IID 的程度，其中 $\alpha$ 值越小，数据分布越不均匀。图 2 是数据集 FMINIST 和 CIFAR-10 进行 Non-IID 之后的可视化展示。使用不同的划分方式和参数来模拟不同的异质性情况。

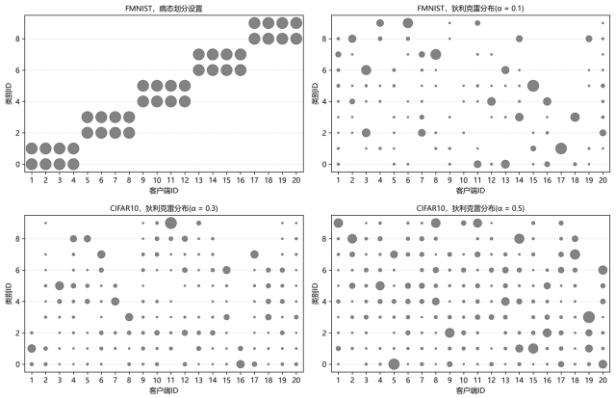



图2　异构数据可视化展示

（3）模型与训练设置

在实验中，将使用多种客户端模型架构，如 ResNet、MobileNet 和 MLP-Mixer，并在不同的客户端数据分布和通信设置下进行测试。性能指标将包括模型在测试集上的准确率、达到目标准确率所需的通信轮次以及整个训练过程的时间开销。通过这些实验，可以全面评估该方案在不同场景下的性能表现。

每个客户端根据自身的资源选择合适的模型架构，并使用 SGD 或 Adam 优化器进行本地训练。训练周期数根据模型和数据集的不同而有所调整，通常在 10 到 40 个 epoch 之间。性能指标将包括模型在测试集上的准确率、达到目标准确率所需的通信轮次以及整个训练过程的时间开销。通过这些实验，可以全面评估该方案在不同场景下的性能表现。

（4）对比基线

□ FedAvg[3]：经典参数聚合算法，作为基础联邦学习对照。□ FedProx[22]：在 FedAvg 基础上加入近端约束，可在一定程度上缓解异构数据带来的训练不稳定性。□ MOON[23]：基于对比学习的联邦优化方法，通过表示对齐提升异构数据场景下的鲁棒性。□ FedGen[13]：利用生成器辅助蒸馏，实现异构模型间知识迁移，代表生成式联邦蒸馏方法。□ FD[4]：通过交换 logits 进行轻量级蒸馏，可在模型架构不同的情况下完成知识融合。□ FedKD[24]：将知识蒸馏机制融入联邦聚合过程，提升异构客户端间的模型适配能力。□ FedProto[25]：基于原型表示的个性化方案，通过原型对齐实现跨客户端特征一致性。□ FedRoD[26]：面向表示漂移的鲁棒联邦方法，强化异构客户端间的表征稳定性。□ FedMRL[27]：通过元表示学习为每个客户端生成个性化初始权重，提升非 IID 数据下的快速适应与泛化能力。□ FedTGP[28]：利用跨客户端的拓扑图原型对齐，实现图级异构联邦学习中的鲁棒知识聚合。□FedFree[29]：基于无代理数据的分层知识对齐与增益筛选策略，通过高斯伪数据解决跨架构特征错位问题。□FedMKD[30]：采用轻量级代理模型介导的多粒度知识融合蒸馏，通过特征层与类别层的双向对齐缓解模型异构。

这些基线覆盖了联邦蒸馏、生成辅助蒸馏、表示对齐与异构模型协同等主要范式，可全面评估新方法在异构环境下的知识融合能力。

### 3.2　算法性能对比

#### 3.2.1　同构模型蒸馏

（1）实验设计

在 CIFAR-100 和 Tiny-ImageNet 两个核心数据集上，对比本文方法与基线算法在不同 Non-IID 程度下的准确率、收敛速度，所有算法共享 3.1 节的训练参数配置，确保对比公平性。

（2）准确率对比结果

表 1 分别展示了 CIFAR-100 和 Tiny-ImageNet 数据集上各算法在不同数据异构程度下的最终准确率。可以看出，在强数据异构（$\alpha$ = 0.01）条件下，本文方法与最优算法 FedFree 存在一定差距，但显著优于 FedGen、FedAvg、FedProx 等传统联邦学习方法，且与 FD（74.63%）、FedMKD（74.66%）等

经典联邦蒸馏算法性能接近，差距均在 1 个百分点以内。

随着数据异构程度不断变化，所有算法的性能均出现不同程度波动。表 1 中 Per-FedAvg、Ditto 等个性化方法性能波动较大，原因在于训练范式与场景的结构性冲突：数据均匀化会凸显参数聚合的破坏性，同时削弱本地个性化微调的必要性，导致模型退化而非受益。本文方法通过跨架构知识对齐，规避了参数空间的直接冲突，因此在各类数据分布下稳定性均更优。其中，FD 退化最为明显，在 $\alpha = 5$ 时准确率仅为 15.45%；与之相比，本文 FedOFA 方法虽然也下降至 22.21%，但仍高于 FD 等传统蒸馏方法，说明在无极端优化的情况下具备一定稳定性。然而，在数据分布趋于均匀的情况下，本文方法仍明显低于无代理数据的 FedFree（42.18%）与 FedMKD（34.90%），反映出在全局知识趋同的条件下时，依赖公共数据集对齐的蒸馏策略仍存在局限性，也是未来需要重点改进的方向。

在大规模细分类数据集 Tiny-ImageNet 上，本文算法在多数异构设置下保持了具有竞争力的结果。尤其在 $\alpha = 0.1$ 时，FedOFA 准确率达到 34.88%，较 FedGen（19.19%）提升 15.69 个百分点。这是因为传统生成式蒸馏方法的性能受限于生成器容量与数据分布拟合能力，而本文提出的特征投影与自适应增强蒸馏能够更稳定地对齐异构特征，更适用于大规模类别与复杂图像场景。同时可以观察到，FedFree 凭借无代理数据的分层知识迁移机制，在各类数据异构条件下均保持领先，体现出更强的鲁棒性。

（3）收敛速度对比

如图 3 所示，在 CIFAR-100（$\alpha = 0.1$）数据集上，本文方法收敛速度最快，在 20 轮左右就已经达到了较高的准确率和较小的损失，而最终达到收敛也仅需 85 轮通信，而 FD 需要 120 轮，FedGen 需要 150 轮，FedAvg 需要 180 轮。同时可以观察到，FD 的训练损失曲线波动剧烈，难以稳定收敛，核心原因在于 FD 仅依赖输出层 logits 进行直接蒸馏，缺乏中间特征投影的对齐机制与自适应增强模块的去噪支持。本文方法收敛速度快是因为多阶段特征投影让全局模型在不同抽象层级均能吸收异构知识，避免了仅依赖最终 logits 导致的中间层知识缺失；自适应增强损失加快了模型对可靠知识的学习速度，减少了噪声信息带来的训练波动，从而显著提升收敛效率。

**表 1　算法在实际标签偏移场景下不同异构程度的测试准确率（%）**

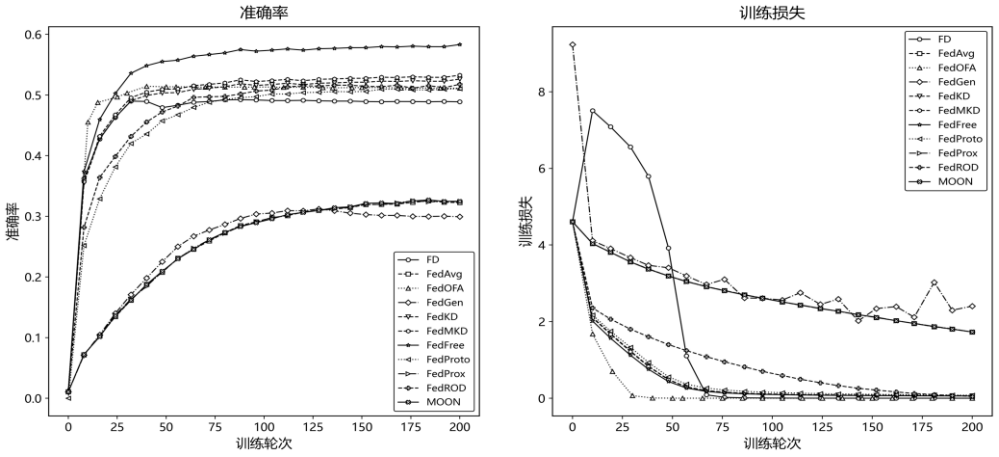
| Datasets | Cifar100 | | | | Tiny-ImageNet | | | |
|---|---|---|---|---|---|---|---|---|
| | $\alpha$=0.01 | $\alpha$=0.1 | $\alpha$=0.5 | $\alpha$=5 | $\alpha$=0.01 | $\alpha$=0.1 | $\alpha$=0.5 | $\alpha$=5 |
| FedAvg | 23.58 | 31.89 | 27.99 | 35.51 | 15.70 | 19.46 | 21.14 | 21.71 |
| FedProx | 23.74 | 31.99 | 35.05 | 35.31 | 15.66 | 19.37 | 21.22 | 27.69 |
| FedGen | 20.89 | 30.96 | 33.88 | 35.64 | 15.87 | 19.19 | 21.06 | 21.44 |
| Per-FedAvg | 49.25 | 44.28 | 35.32 | 24.94 | 39.39 | 25.07 | 16.36 | 12.08 |
| Ditto | 73.29 | 52.87 | 26.28 | 35.72 | 50.62 | 32.15 | 18.98 | 21.79 |
| FedRoD | 67.78 | 50.94 | 36.29 | 25.63 | 49.17 | 36.43 | 23.23 | 16.71 |
| FD | 74.63 | 49.93 | 29.32 | 15.45 | 50.49 | 30.02 | 14.34 | 6.56 |
| FedMKD | 74.66 | 53.28 | 42.17 | 34.90 | 49.82 | 40.15 | 27.82 | 20.17 |
| FedFree | 78.12 | 58.33 | 48.52 | 42.18 | 52.37 | 44.69 | 32.41 | 26.33 |
| FedOFA | 73.79 | 51.13 | 33.77 | 22.21 | 49.29 | 34.88 | 16.22 | 8.55 |

### 3.2.2　异构模型蒸馏

（1）实验设计

异构架构下的知识迁移核心障碍在于不同模型特征空间的天然不兼容。如图 4 所示：前 3 列的同架构模型特征呈现“对角高相似”分布，说明同架构特征空间具有一致性；而后 3 列的跨架构模型特征相似度分散且整体偏低，不同层特征几乎无对齐性，验证了异构架构特征空间的离散性——这种“特征错位”会直接导致知识迁移失效，是本实验需解决的核心问题。

为此，构建联邦学习场景，设置 10 个客户端并采用异构模型配置，模型集合涵盖 8 类主流深度学习架构：包括专为小尺寸图像设计的基础卷积神经网络 FedAvgCNN（适配 32×32 输入与 1600 维特征映射）、轻量级网络 MobileNetV2、经典深度卷积网络 GoogLeNet，以及不同深度层级的 ResNet 系列（ResNet-18/34/50/101/152，通过堆叠残差块实现从浅到深的性能梯度），同时纳入适配小图像输入的 Vision Transformer 模型（ViT-B/16、ViT-B/32），形成“基础 CNN-轻量 CNN-深度 CNN-Transformer”的多元异构模型组合，可系统探究不同模型复杂度、特征提取范式在联邦学习异构场景下的性能表现与适配性。




(a)　准确率　　(b)　训练损失

图3　准确率和损失随轮次变化图

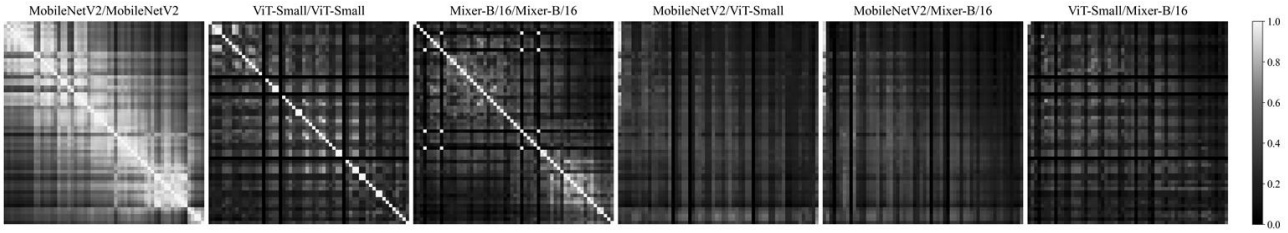



图4　基于中心核对准(Centered Kernel Alignment, CKA)度量的中间特征相似度热力图

(2)特征投影对齐效果验证

为量化验证特征投影模块对不同复杂度模型的对齐效果，本文采用CKA度量各异构模型在投影前后的特征相似度，结果如表2所示。实验表明：投影前，轻量级模型（MobileNetV2）与超大规模模型（ResNet-152、ViT-B/16）的特征相似度仅为0.32~0.38，存在严重的特征空间偏移；经本文多阶段特征投影后，不同复杂度模型的特征相似度提升至0.71~0.78，特征空间对齐效果显著，验证了特征投影模块在跨复杂度异构场景下的有效性。上述结果定量表明，本文方法可有效弥合轻量级模型与大规模模型间的特征空间鸿沟，为跨复杂度异构模型的协同推理提供了可靠的表示对齐基础。

(3)结果分析

实验分别在CIFAR-10与CIFAR-100数据集上展开验证：由表3和图5可知，在CIFAR-10数据集的多算法对比实验中，本文方法实现了最高的准确率，最终准确率达86.14%、峰值准确率达86.61%，显著优于FedKD、FedMRL、FedProto等基线方法，同时也优于FedFree、FedMKD等近年异构联邦蒸馏方法。从训练过程来看，FedOFA在初始阶段（0-25轮）即呈现更陡峭的准确率上升趋势，收敛速率快于多数对比算法，且训练后期（50轮后）的准确率波动极小，体现出更稳定的训练特性。在CIFAR-100数据集的单独对比实验中，随着任务复杂度的提升，本文方法与其他基线的差距更大，最终准确率大幅领先于其他算法，其性能优势从训练

**表2　特征投影模块的模型对齐效果(CKA)**

| 模型组合 | 投影前 | 投影后 | 提升幅度 |
|---|---|---|---|
| MobileNetV2/ResNet-152 | 0.32 | 0.71 | 0.39 |
| MobileNetV2/ViT-B/16 | 0.35 | 0.75 | 0.40 |
| ResNet-152/ViT-B/16 | 0.38 | 0.78 | 0.40 |
| 平均 | 0.35 | 0.75 | 0.40 |

初期持续维持至收敛阶段；从定量结果来看，本文方法最终准确率41.68%、峰值准确率42.22%，较FedFree、FedMKD等方法分别提升0.77个、3.06个百分点，优势进一步凸显。上述结果表明，本文方法在不同复杂度的图像分类任务（CIFAR-10/CIFAR-100）中均能稳定优于基线算法，验证了其在联邦学习异构场景下的有效性与泛化性。

**表3　各算法最终与峰值准确率（%）**

| 算法 | Cifar10 最终 | Cifar10 峰值 | Cifar100 最终 | Cifar100 峰值 |
|---|---|---|---|---|
| FedOFA | 86.14% | 86.61% | 41.68% | 42.22% |
| FedFree | 85.72% | 86.18% | 40.91% | 41.55% |
| FedMKD | 84.95% | 85.30% | 38.62% | 39.10% |
| FedKD | 79.86% | 80.30% | 29.03% | 29.30% |
| FD | 83.23% | 83.27% | 33.21% | 33.72% |
| FedMRL | 84.53% | 84.67% | 35.49% | 35.85% |
| FedProto | 72.15% | 72.15% | 25.58% | 25.69% |
| FedTGP | 79.47% | 80.17% | 27.91% | 28.25% |
| Local | 84.08% | 84.17% | 34.74% | 34.74% |

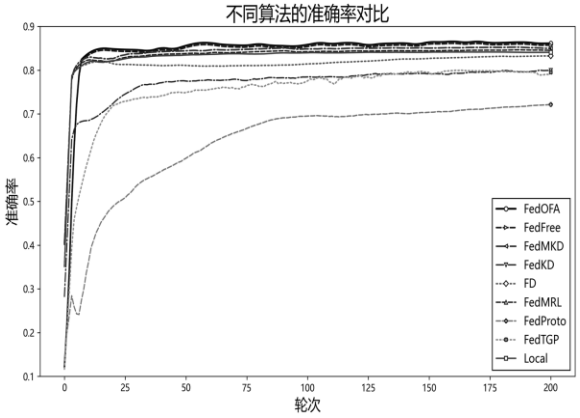
### 3.3　通信效率与鲁棒性分析

#### 3.3.1　通信效率对比

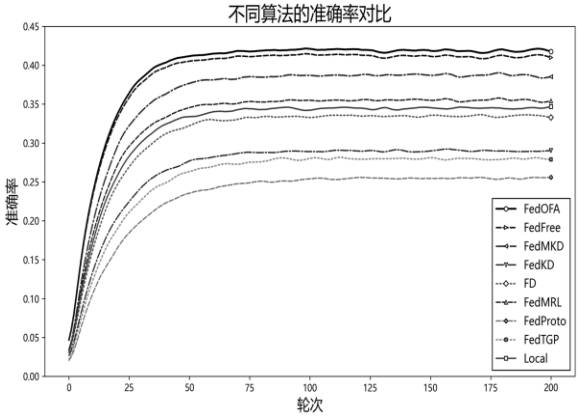
在CIFAR-10、10个客户端的设置下，本文以32位浮点为计量单位，采用“平均每轮通信量”对不同方法的通信开销进行对比，从而避免总轮数差异带来的偏差。图6中横轴为对数刻度，可见各方法通信量跨越多个数量级。

总体上，通信量差异来自于“每轮交换对象”的规模不同。模型交换类方法（FML、FedMRL）需要在每轮上传/下发模型参数，因此单轮通信量达到$10^4$KB量级；FedKD虽引入压缩，但仍需传输压缩参数表示，单轮通信量同样处于MB量级。FedMKD采用轻量级代理模型介导蒸馏，单轮需传输代理模型参数，通信量约为$10^3$KB，虽高于本文方法，但仍远低于MB级的完整模型交换方案。相对地，FD仅交换类别级聚合统计量，通信量最低；FedFree通过分层知识对齐与增益筛选，仅上传少量关键层参数，单轮通信量约为75KB，显著低于模型交换类方法。FedProto/FedTGP交换按类别聚合的原型向量，通信量为KB量级。

本文方法的核心交互对象为公共数据上的蒸馏信号，其通信规模随类别数与公共数据规模线性增长，而不随模型参数量直接增长。因此，单轮通信量约为275KB，显著低于FML/FedMRL/FedKD等模型交换方法的数十MB/轮，同时也优于需传输代理模型的FedMKD。该结果表明，本文方法在保持有效知识迁移的同时具备更优的带宽开销特性，适用于带宽受限与多客户端扩展场景。




(a)　CIFAR-10数据集




(b)　CIFAR-100数据集

图5　算法在CIFAR-10和CIFAR-100数据集上的准确率

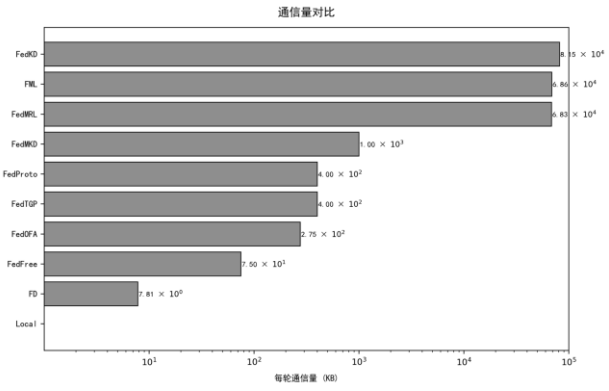
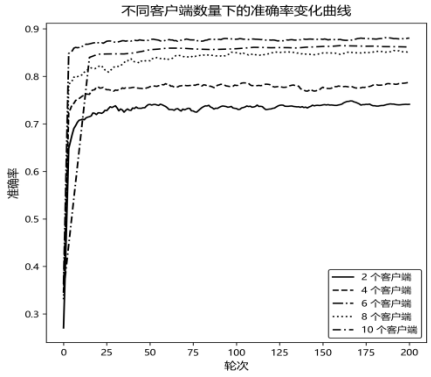


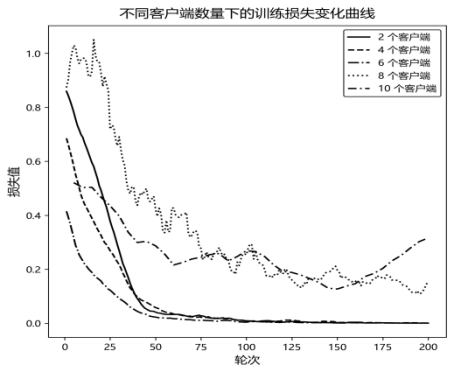

图6　算法通信量

### 3.3.2　鲁棒性分析

为验证所提算法在客户端异构场景下的鲁棒性，实验设置不同数量的异构客户端参与训练。从图 7 中的准确率曲线可见，不同客户端数量配置下，模型准确率均能快速收敛至 0.8 左右的较高水平区间，且收敛后未出现显著的性能衰减；结合损失曲线，各配置下的损失值均逐步收敛至趋近于 0 的区间，且后期波动幅度保持稳定，未因客户端数量增加而出现损失发散的情况，这一现象直接体现了算法在客户端异构性变化时的收敛稳定性。




(a)　准确率




(b)　训练损失

图7　不同客户端数量下的训练曲线

从图 8 来看，当客户端数量从 2 扩展至 10 时，模型准确率整体呈“上升后平稳”的趋势：客户端数量增加至 6 时，准确率提升至 88.27%；继续增加客户端至 8、10 时，准确率虽有小幅波动，但仍维持在 85% 以上的较高水平——这一结果表明，算法在客户端数量动态变化的异构场景中，能够适应数据分布的差异性，未因参与节点的增加而出现性能骤降，体现了其对客户端异构性的鲁棒适应能力。

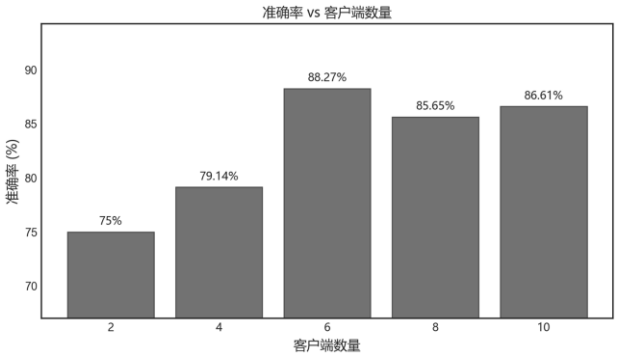



图8　客户端数量对准确率的影响

## 3.4　消融与参数敏感性分析

### 3.4.1　鲁棒性分析

为验证所提出框架中各关键组件的贡献，本文在 CIFAR-10 数据集上开展消融实验，结果见图 9。完整框架准确率为 86.61%，作为对照基线。移除特征投影后，服务器端仅利用最终 logits 与聚合 logits 构造损失，准确率降至 83.11%。该结果表明特征投影模块对性能提升贡献最为显著，提示其在异构表示对齐与中间层知识补充方面发挥了关键作用。移除自适应增强后，以传统 KL 散度替代 OFA 蒸馏损失，准确率降至 84.51%，说明自适应增强机制有助于提升蒸馏过程的有效性，可通过强调可靠监督、抑制噪声而改善泛化表现。移除加权聚合、改用均值聚合后，准确率降至 84.81%，表明基于数据量的加权策略能够更充分地利用高质量客户端信息。综上，三个模块均对最终性能具有稳定增益，三者协同使得完整框架在非 IID 场景下取得更优表现。

### 3.4.2　参数敏感性分析

在 CIFAR-10 数据集上，本文对框架的关键超参数——蒸馏温度 $T$、平衡系数 $\alpha$ 以及自适应调节参数 $\gamma$——进行了参数敏感性分析。实验遵循“单因素变动”原则：除当前考察的超参数外，其余设置均保持为默认配置；在 $T$ 与 $\gamma$ 的敏感性实验中固定 $\alpha$ 为基准值，而在 $\alpha$ 的敏感性实验中仅改变 $\alpha$。如表 4 和图 10 所示，$T = 4$ 取得最高准确率 86.61%；当 $T < 4$ 时，蒸馏分布过于尖锐使得知识迁移过度集中，从而削弱模型的泛化能力；当 $T > 4$ 时，过强的软化会放大噪声信息，导致性能下降。进一步地，当 $\alpha = 0.5$ 表现最优；$\alpha$ 过小会使特征对齐损失权重不足，难以有效约束表征一致性，而 $\alpha$ 过大则相对削弱蒸馏损失的作用，两种情况都

会引起准确率下降。对于$\gamma$，$\gamma = 1.5$在异构程度较高的场景下取得最佳性能；$\gamma \approx 1.0$更适用于同构或低异构设置，而$\gamma > 1.5$时自适应调节强度过高，容易带来过拟合风险。总体而言，上述核心超参数在合理取值范围内表现出较好的稳定性，且最优配置具有一定通用性，说明该框架对复杂调参的依赖较低。

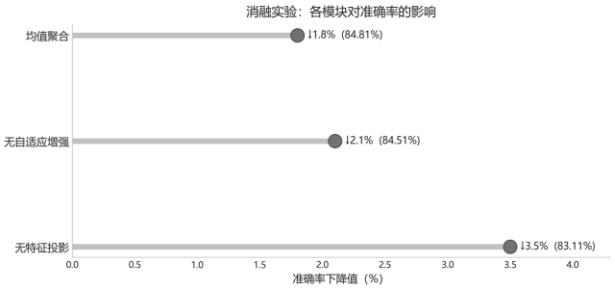



图9　模块消融的精度衰减量化分析

**表4　　参数敏感性分析结果（%）**

| $T$ | Acc(%) | $\alpha$ | Acc(%) | $\gamma$ | Acc(%) |
|---|---|---|---|---|---|
| 2 | 84.41 | 0.3 | 85.21 | 1.0 | 85.61 |
| 4 | 86.61 | 0.5 | 86.61 | 1.2 | 86.11 |
| 6 | 85.61 | 0.7 | 85.41 | 1.5 | 86.61 |
| 8 | 84.01 | — | — | 2.0 | 85.31 |

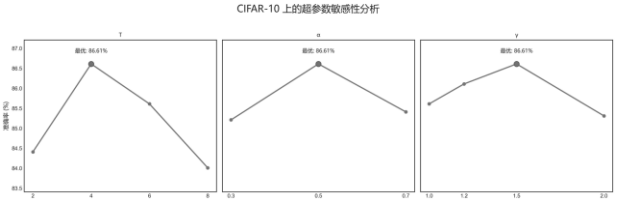



图10　参数敏感度分析

## 4　总结

本文针对传统联邦学习在异构模型环境中存在的参数平均受限与知识迁移效率低下问题，提出了一种结合特征投影机制与自适应增强的异构联邦学习框架——FedOFA。该方法通过在服务器端引入特征空间对齐与自适应目标增强模块，实现了跨架构模型间的高效知识迁移。实验结果表明，该框架在多种数据集上，在准确率、收敛速度方面均表现良好，证明了其在异构环境下的有效性与鲁棒性。

尽管FedOFA在模型异构方面取得了良好平衡，但在数据极端异构情况下表现不佳。未来工作可以聚焦数据-模型异构耦合场景的协同优化，针对当前框架未充分消解的双异构耦合偏差，设计动态感知的分层知识迁移机制，为数据分布极端偏离的客户端优先传递通用基础特征，同时引入联邦域自适应正则项实现特征分布的跨客户端校准。

## 参考文献：

[1] Verbraeken J, Wolting M, Katzy J, et al. A survey on distributed machine learning[J]. ACM Computing Surveys, 2020, 53(2): 1-33.

[2] Kairouz P, Mcmahan H B, Avent B, et al. Advances and open problems in federated learning[J]. Foundations and Trends in Machine Learning, 2021, 14(1-2): 1-210.

[3] Mcmahan H B, Moore E, Ramage D, et al. Communication efficient learning of deep networks from decentralized data[C]//Proc of the 20th Int Conf on Artificial Intelligence and Statistics. Fort Lauderdale, 2017: 1273-1282.

[4] Jeong E, Oh S, Kim H, et al. Communication-efficient on device machine learning: Federated distillation and augmentation under non-IID private data[J/OL]. arXiv preprint, 2018: arXiv:1811.11479.

[5] Li L, Gou J P, Yu B S, et al. Federated Distillation: A Survey[J/OL]. arXiv preprint, 2024: arXiv:2404.08564.

[6] Ge H H, Pokhrel S R, Liu Z Y, et al. PFL-DKD: Modeling decoupled knowledge fusion with distillation for improving personalized federated learning[J]. Computer Networks, 2024, 254: 110758.

[7] Liu W, Chen J Y, Wang B, et al. FedDM: A discrepancy-aware federated learning method based on multibranch feature fusion for non-IID data environments[J]. IEEE Internet of Things Journal, 2025, 12(21): 45825-45835.

[8] Gou J P, Yu B, Maybank S J, et al. Knowledge distillation: a survey[J]. International Journal of Computer Vision, 2021, 129(6): 1789-1819.

[9] Huang Y L, Hu K, Zhang Y, et al. Distilling knowledge from heterogeneous architectures for semantic segmentation[C]//Proc of the AAAI Conference on Artificial Intelligence. Vancouver, 2024, 38: 3824-3832.

[10] Li Y C, Wang X Y, Xu W C, et al. Feature distillation is the better choice for model-heterogeneous federated learning[C]//Proc of the 39th Advances in Neural Information Processing Systems. Vancouver, 2025: 104726-104744.

[11] Lin T, Kong L J, Stich S U, et al. Ensemble distillation for robust model fusion in federated learning[C]//Proc of NeurIPS 2020. Vancouver, 2020: 2351-2363.

[12] 郑晶晶, 赖金山, 张凤荔, 等. FedASHP：高性能通信的自适应稀疏高效联邦学习方法[J/OL]. 小型微型计算机系统, 1-9[2026-04-15]. https://doi.org/10.20009/j.cnki.21-1106/TP.2025-0342.
Zheng J J, Lai J S, Zhang F L, et al. FedASHP: An adaptive sparse efficient federated learning method for high-performance communication [J/OL]. Journal of Chinese Computer Systems, 1-9[2026-04-15]. https://doi.org/10.20009/j.cnki.21-1106/TP.2025-0342.

[13] 刘佳瑜, 王勇, 杨静, 等. FedDW：基于一致性优化的权重蒸馏异构联邦学习方法[J]. 计算机研究与发展, 2026, 63(4): 998-1009.
Liu J Y, Wang Y, Yang J, et al. FedDW: Weight distillation heterogeneous federated learning method based on consensus optimization[J]. Journal of Computer Research and Development, 2026, 63(4): 998-1009.

[14] 王瑞锦, 王金波, 张凤荔, 等. 联邦原型学习的特征图中毒攻击和双重防御机制[J]. 软件学报, 2025, 36(3): 1355-1374.
Wang R J, Wang J B, Zhang F L, et al. Feature graph poisoning attack and double defense mechanism for federated prototype learning[J]. Journal of Software, 2025, 36(3): 1355-1374.

[15] Siddika F, Hossen M A, Zhang W S, et al. Dual-Distilled Heterogeneous Federated Learning with Adaptive Margins for Trainable Global Prototypes[C]//Proc of the 42nd Int Conf on Machine Learning. Vienna, 2025.

[16] Song W S, Yan M W, Li X Z, et al. Bidirectional decoupled distillation for heterogeneous federated learning[J]. Entropy, 2024, 26(9): 762.

[17] Li K, Ding Y, Zhu Z Q, et al. Transforming gaps into gains: bridging model and data heterogeneity in federated learning via knowledge weak-aware zones[C]//Proc of the 39th Advances in Neural Information Processing Systems. Vancouver, 2025.

[18] Yu Y, Kim N. Heterogeneous knowledge distillation using conceptual learning[J]. IEEE Access, 2024, 12: 52803-52814.

[19] Li T, Sahu A K, Talwalkar A, et al. Federated learning: challenges, methods, and future directions[J]. IEEE Signal Processing Magazine, 2020, 37(3): 50-60.

[20] Bottou L, Curtis F E, Nocedal J. Optimization methods for large-scale machine learning[J]. SIAM Review, 2018, 60(2): 223-311.

[21] Hsu T M H, Qi H, Brown M. Measuring the effects of non-identical data distribution for federated visual classification[J/OL]. arXiv preprint arXiv:1909.06335, 2019.

[22] Li T, Sahu A K, Zaheer M, et al. FedProx: Federated optimization in heterogeneous networks[C]//Proc of MLSys 2020. Austin, 2020: 429-450.

[23] Li Q B, He B S, Song D. Model-contrastive federated learning[C]//Proc of the IEEE/CVF Conf on Computer Vision and Pattern Recognition. Nashville, 2021: 10713-10722.

[24] Wu C, Wu F, Lyu L, et al. Communication-efficient federated learning via knowledge distillation[J]. Nature Communications, 2022, 13(1): 2032.

[25] Tan Y, Long G D, Liu L, et al. FedProto: Federated prototype learning across heterogeneous clients[C]//Proc of AAAI 2022. Vancouver, 2022: 8042-8050.

[26] Chen H Y, Chao W L. On bridging generic and personalized federated learning for image classification[C]//Proc of the 10th Int Conf on Learning Representations. Virtual, 2022.

[27] Yi L P, Yu H, Ren C, et al. Federated Model Heterogeneous Matryoshka Representation Learning[C]//Proc of the 38th Advances in Neural Information Processing Systems. Vancouver, 2024: 66431-66454.

[28] Zhang J Q, Liu Y, Hua Y, et al. FedTGP: Trainable Global Prototypes with Adaptive-Margin-Enhanced Contrastive Learning for Data and Model Heterogeneity in Federated Learning[C]//Proc of the 38th AAAI Conference on Artificial Intelligence. AAAI Press, 2024: 16768-16776.

[29] Du H Z, Xiang Y R, Cai Y W, et al. FedFree: Breaking Knowledge-sharing Barriers through Layer-wise Alignment in Heterogeneous Federated Learning[C]//The 39th Annual Conference on Neural Information Processing Systems. Vancouver, Canada: Neural Information Processing Systems Foundation, 2025.

[30] Han P, Xiao H, Zheng S, et al. FedMKD: Hybrid Feature Guided Multilayer Fusion Knowledge Distillation in Heterogeneous Federated Learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2026, 37(3): 1463-1476.

陈宁江（1975- ），男，广西南宁人，博士，广西大学教授、博士生导师，CCF 杰出会员，主要研究方向为云计算和大数据、智能软件工程等。

章德华（2001- ），男，壮族，江西抚州人，主要研究方向为边缘计算、联邦学习等。