FEDTAIL: Federated Long-Tailed Domain Generalization with Sharpness-Guided Gradient Matching

Sunny Gupta [1] **Nikita Jangid** [2] **Shounak Das** [2] **Amit Sethi** [2]

Abstract

Domain Generalization (DG) seeks to train models that perform reliably on unseen target domains without access to target data during training. While recent progress in smoothing the loss landscape has improved generalization, existing methods often falter under long-tailed class distributions and conflicting optimization objectives. We introduce **FedTAIL**, a federated domain generalization framework that explicitly addresses these challenges through *sharpness-guided, gradient-aligned optimization*. Our method incorporates a *gradient coherence regularizer* to mitigate conflicts between classification and adversarial objectives, leading to more stable convergence. To combat class imbalance, we perform *class-wise sharpness minimization* and propose a *curvature-aware dynamic weighting scheme* that adaptively emphasizes underrepresented tail classes. Furthermore, we enhance conditional distribution alignment by integrating *sharpness-aware perturbations* into entropy regularization, improving robustness under domain shift. FedTAIL unifies optimization harmonization, class-aware regularization, and conditional alignment into a scalable, federated-compatible framework. Extensive evaluations across standard domain generalization benchmarks demonstrate that FedTAIL achieves *state-of-the-art performance*, particularly in the presence of domain shifts and label imbalance, validating its effectiveness in both centralized and federated settings. Our code is publicly available at: https://github.com/sunnyinAI/FedTail

1. Introduction

Deep learning has excelled in computer vision, especially when source and target data share similar, independently and identically distributed characteristics. However, performance often degrades when models encounter target domains differing from the training distribution. Domain generalization (DG) techniques (Zhang et al., 2022; Qiao et al., 2020; Balaji et al., 2018) aim to train models only on source domains to generalize well to unseen targets without extra fine-tuning. Various DG methods include domain alignment (Muandet et al., 2013), meta-learning (Li et al., 2018a), and data augmentation (Wang et al., 2022b). Yet, the DomainBed benchmark showed that a simple entropy-based regularization, DG via ER (Zhao et al., 2020a), can outperform more complex DG strategies under standard evaluation.

However, minimizing empirical loss on a non-convex landscape does not ensure robust generalization. Like many empirical risk minimization approaches, DG via ER can overfit and converge to sharp minima. To address this, sharpness-aware minimization (SAM) (Foret et al., 2020) improves generalization by minimizing loss surface sharpness. SAM minimizes the worst-case loss in a parameter neighborhood by computing an adversarial perturbation ϵ that maximizes loss, then updating parameters to minimize the perturbed loss. Though effective, SAM simplifies the min-max objective for tractability. Building on this, our work applies sharpness minimization to enhance generalization. Yet, as (Rangwani et al., 2022) notes, SAM struggles with long-tailed settings, where tail classes may converge at saddle points in high-curvature regions, harming performance.

Long-tailed class distributions are common in real-world data but often overlooked in DG research. Entropy minimization is popular in semi-supervised learning to encourage confident predictions (Chen et al., 2019; Grandvalet & Bengio, 2004), but its DG effects are less studied. Examining entropy-based gradient flow (Chen et al., 2019) reveals that high-confidence samples receive larger gradients, causing easier-to-transfer classes to dominate training, while harder classes remain undertrained. This probability imbalance leads to insufficient optimization of tail classes. Our method directly addresses this gradient imbalance to

[1]Koita Centre for Digital Health, Indian Institute of Technology Bombay, Mumbai, India [2]Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India. Correspondence to: Sunny Gupta <sunnygupta@iitb.ac.in>.

Proceedings of the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

improve generalization across both head and tail classes.

2. Related Work

Domain Generalization. Domain Generalization (DG) aims to learn a model from one or several observed source datasets that can generalize effectively to unseen target domains (Zhao et al., 2020a). A variety of approaches have been proposed to tackle domain shift, including domain alignment techniques (Muandet et al., 2013; Ganin et al., 2016; Li et al., 2018b; Bahng et al., 2020; Zhao et al., 2020a), meta-learning strategies (Zhang et al., 2021; Dou et al., 2019; Balaji et al., 2018; Li et al., 2018a), data augmentation (Zhou et al., 2021; Shankar et al., 2018; Carlucci et al., 2019), disentangled representation learning (Peng et al., 2019b; Khosla et al., 2012; Wang et al., 2020a), and methods based on causal reasoning (Krueger et al., 2021; Arjovsky et al., 2019). Among these, a growing body of work has addressed DG from a gradient-based perspective (Li et al., 2018a; Balaji et al., 2018; Dou et al., 2019; Zhang et al., 2021), aiming to stabilize learning across diverse domains. For instance, Mansilla et al. (Mansilla et al., 2021) introduced a gradient surgery mechanism to address inter-domain conflicts by preserving gradient components with matching signs and nullifying those with opposite directions. While such techniques improve robustness during training, they do not guarantee convergence to flat minima—an important factor for out-of-distribution generalization, especially in low-resource or noisy domains. Moreover, most prior works assume balanced and centralized data distributions, limiting their effectiveness in real-world federated settings where data is inherently non-i.i.d. and long-tailed.

Sharpness-Aware Minimization (SAM). Sharpness-Aware Minimization (SAM) (Foret et al., 2020) is a powerful regularization method that improves generalization by minimizing the maximum loss within a neighborhood around model parameters. By formulating optimization as a min-max problem, SAM encourages convergence to flatter loss regions, avoiding sharp minima linked to poor generalization. However, SAM mainly operates globally and ignores class-specific curvature variations—limiting its effectiveness in long-tailed class distributions. Variants like GSAM (Liu et al., 2022) and LookSAM (Du et al., 2021) improve sharpness estimation and efficiency but still neglect issues like class imbalance and the distributed, non-i.i.d. data nature in federated learning.

Sharpness and Generalization. Sharpness's connection to generalization was first studied in (Hochreiter & Schmidhuber, 1994) and further explored under i.i.d. assumptions (Keskar et al., 2016; Dinh et al., 2017; Foret et al., 2020). For example, (Keskar et al., 2016) showed sharpness inversely correlates with generalization, and (Dinh et al., 2017) linked it to the Hessian's eigenvalues. Extensions to out-of-distribution settings, like SWAD (Cha et al., 2021), show flatter minima reduce domain generalization gaps but don't explicitly enforce flatness during training. This motivates our focus on sharpness-aware generalization in federated and long-tailed contexts, where inductive biases are critical with limited data.

Domain generalization methods like EISNet (Wang et al., 2020b) and FACT (Zhao et al., 2020a) improve feature transferability and domain invariance. SAMALTDG (Su et al., 2024) addresses class imbalance by combining SAM with a loss that emphasizes tail classes but remains centralized and ignores optimization conflicts in federated setups. Our proposed method, **FedTAIL**, extends sharpness-aware learning to federated domain generalization by incorporating gradient coherence, class-wise curvature weighting, and domain-agnostic optimization, jointly addressing long-tailed imbalance, sharpness, and gradient conflicts in decentralized data.

3. Preliminaries

3.1. Domain Generalization under Federated Long-Tailed Distributions

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label spaces, respectively. We assume access to K source domains $\{\mathcal{D}_i\}_{i=1}^{K}$, each distributed across decentralized clients. Data samples from domain i follow a joint distribution $P_i(X, Y)$, and we denote $\mathcal{D}_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{N_i}$. No target domain data is available during training. The objective is to train a global model $h_\theta = T_\phi(F_\theta(\cdot))$, composed of a feature extractor F_θ and a classifier T_ϕ, that generalizes to an unseen domain $\mathcal{D}_T$.

In the federated setting, data is kept locally on each client and model updates are aggregated using federated averaging (FedAvg). Additionally, class imbalance is assumed across clients, yielding long-tailed label distributions where certain classes dominate the training set while others are severely underrepresented.

3.2. Empirical Risk and Adversarial Alignment

The standard classification loss across the K source domains is given by the empirical risk:

$$\mathcal{L}_{\text{cls}} = -\sum_{i=1}^{K} \mathbb{E}_{(x,y)\sim\mathcal{D}_i} \left[\log T_\phi(F_\theta(x))_y\right], \qquad (1)$$

where $T_\phi(F_\theta(x))_y$ denotes the predicted probability for the true label y.

To encourage domain-invariant representations, we adopt adversarial domain alignment (Ganin & Lempitsky, 2015), introducing a domain discriminator D_ψ trained to distinguish between domains, while F_θ is updated to fool D_ψ. The adversarial loss is:

$$\mathcal{L}_{\text{adv}} = \sum_{i=1}^{K} \mathbb{E}_{x \sim \mathcal{D}_i} \left[\log D_\psi(F_\theta(x)) \right], \tag{2}$$

where $D_\psi(F_\theta(x))$ predicts the domain label for sample x.

3.3. Sharpness-Aware Minimization (SAM)

SAM (Foret et al., 2020) aims to improve generalization by minimizing the worst-case loss within an ℓ_2-bounded neighborhood of model parameters. The SAM objective is:

$$\min_{\theta} \max_{\|\epsilon\| \leq \rho} \mathcal{L}(\theta + \epsilon), \tag{3}$$

where ρ is a radius controlling the perturbation strength. In practice, SAM approximates the inner maximization via first-order Taylor expansion:

$$\epsilon(\theta) \approx \rho \cdot \frac{\nabla_\theta \mathcal{L}(\theta)}{\|\nabla_\theta \mathcal{L}(\theta)\|_2}. \tag{4}$$

The outer minimization is then computed using the perturbed parameters $\theta + \epsilon(\theta)$.

3.4. Surrogate Gap and Gradient Matching

SAGM (Wang et al., 2023) improves upon SAM by minimizing both the empirical loss $\mathcal{L}(\theta)$ and the perturbed loss $\mathcal{L}_p(\theta) = \mathcal{L}(\theta + \epsilon(\theta))$, as well as aligning their gradient directions. The sharpness of the solution is captured by the surrogate gap:

$$h(\theta) = \mathcal{L}_p(\theta) - \mathcal{L}(\theta), \tag{5}$$

which approximates the curvature of the loss landscape. SAGM minimizes the following joint objective:

$$\mathcal{L}_{\text{SAGM}} = \mathcal{L}(\theta) + \mathcal{L}_p(\theta) - \alpha \cdot \langle \nabla \mathcal{L}(\theta), \nabla \mathcal{L}_p(\theta) \rangle, \tag{6}$$

where the last term promotes gradient alignment to facilitate stable descent toward flat minima.

3.5. Long-Tailed Class Distributions and Maximum Square Loss

In long-tailed settings, standard losses such as cross-entropy are prone to bias toward head classes. SAMALTDG (Su et al., 2024) addresses this using the *Maximum Square Loss*:

$$\mathcal{L}_{\text{m}} = -\frac{1}{2N} \sum_{n=1}^{N} \sum_{c=1}^{C} (p_{n,c})^2, \tag{7}$$

where $p_{n,c}$ is the predicted probability of class c for sample n. Compared to entropy loss, the maximum square loss yields more balanced gradients across classes and prevents confident head classes from dominating the optimization.

4. Methodology

FedTAIL, a novel framework for federated domain generalization (FedDG) designed to address the dual challenges of optimization instability and class imbalance under domain shift. FedTAIL builds upon three complementary principles: (i) *gradient coherence* to harmonize competing objectives, (ii) *class-wise sharpness-aware regularization* to improve tail-class learning, and (iii) *sharpness-guided conditional alignment* to enhance feature-label consistency across domains. Together, these modules form a unified, scalable approach that is robust to both federated settings and long-tailed distributions.

We consider a federated DG setting where data from K source domains $\{\mathcal{D}_i\}_{i=1}^K$ are distributed across K clients, each holding samples drawn from a joint distribution $P_i(X, Y)$. The learning goal is to train a global model that generalizes well to an unseen target domain $\mathcal{D}_T$ without direct access to its data. The model comprises a shared feature extractor F_θ and a classifier T_ϕ. In this context, naive empirical risk minimization (ERM) often leads to convergence at sharp local minima, especially under non-i.i.d. distributions and long-tailed label frequency—resulting in poor generalization.

To overcome these issues, FedTAIL introduces a **gradient coherence regularization mechanism** that explicitly resolves conflicts between task objectives—particularly between classification and domain adversarial components. In conventional domain adversarial learning, the joint objective typically comprises a classification loss $\mathcal{L}_{\text{cls}}$ and a domain discrimination loss $\mathcal{L}_{\text{adv}}$, where the gradients may point in divergent directions. FedTAIL mitigates this by introducing an auxiliary alignment term that penalizes negative inner products between their gradients:

$$\mathcal{L}_{\text{coh}} = -\alpha \left\langle \nabla_\theta \mathcal{L}_{\text{cls}}, \nabla_\theta \mathcal{L}_{\text{adv}} \right\rangle, \tag{8}$$

where α is a tunable hyperparameter. This encourages consistency across gradient directions, thereby stabilizing training and ensuring that adversarial alignment does not hinder classification performance.

Beyond harmonizing gradients, FedTAIL directly addresses **class imbalance** by extending sharpness-aware minimization (SAM) to operate at the level of individual classes. Standard SAM seeks flatter minima by minimizing the worst-case loss in a neighborhood of the current parameters. However, this global perspective fails to capture disparities across classes—particularly those underrepresented in long-tailed settings. FedTAIL therefore adopts **class-wise sharpness minimization**, wherein separate perturbations ϵ_c are computed for each class c by normalizing the gradient of the class-specific loss:

$$\epsilon_c = \rho \cdot \frac{\nabla_\theta \mathcal{L}_c}{\|\nabla_\theta \mathcal{L}_c\|_2}, \tag{9}$$

where ρ is the sharpness control radius and $\mathcal{L}_c$ denotes the classification loss restricted to class c.

The corresponding sharpness-aware objective for class c is given by:

$$\mathcal{L}_{\text{sharp}} = \sum_{c=1}^{C} \mathbb{E}_{(x,y=c)} \left[\ell \left(h_{\theta+\epsilon_c}(x), y \right) \right], \tag{10}$$

where $\ell(\cdot,\cdot)$ is the loss function (e.g., cross-entropy) and $h_{\theta+\epsilon_c}$ denotes the perturbed model parameters for class c.

To further prioritize minority classes, we introduce a **curvature-aware dynamic weighting scheme**. This mechanism adjusts the contribution of each class-specific loss based on the sharpness of its local loss landscape. Specifically, the weight γ_c for each class c is computed as:

$$\gamma_c = \frac{1}{1 + \sigma_{\max}(\nabla^2 \mathcal{L}_c)}, \tag{11}$$

where $\sigma_{\max}(\cdot)$ denotes the largest eigenvalue of the class-specific Hessian $\nabla^2 \mathcal{L}_c$. This adaptivity ensures that well-optimized head classes (with flatter curvature) are down-weighted, while sharp, under-trained tail classes receive increased gradient signal.

Complementing these advances, FedTAIL enhances **conditional distribution alignment** by injecting sharpness-awareness into the entropy regularization term. Traditional entropy-based approaches often amplify prediction confidence for easily aligned examples while neglecting ambiguous or hard-to-transfer samples. To address this, we define a sharpness-aware entropy regularization term that perturbs the feature space via SAM and aligns the perturbed predictions to a global conditional distribution:

$$\mathcal{L}_{\text{sharp-er}} = \sum_{i=1}^{K} \mathrm{KL} \left(P_i(Y|F(X)) \,\|\, Q_T(Y|F(X+\epsilon)) \right), \tag{12}$$

where ϵ is the SAM perturbation, and Q_T is a target-like predictive distribution computed from the ensemble or momentum-updated model. This formulation explicitly penalizes high-curvature regions in the conditional landscape, encouraging smoother, more consistent predictions across domains.

Finally, the overall training objective for FedTAIL integrates all components as follows:

$$\mathcal{L}_{\text{FedTAIL}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{sharp-er}} + \sum_{c} \gamma_c \mathcal{L}_c + \mathcal{L}_{\text{coh}}. \tag{13}$$

All modules in FedTAIL are lightweight and compatible with federated optimization. During training, each client computes local gradients, class-specific perturbations, and sharpness-aware updates. These are then aggregated using standard federated averaging without sharing raw data, enabling scalable deployment under real-world privacy constraints.

5. Experiments

We conduct extensive experiments to evaluate the effectiveness of **FedTAIL** on challenging domain generalization benchmarks that exhibit both significant domain shift and long-tailed class distributions. We compare our approach with state-of-the-art DG methods across several backbone architectures and perform detailed ablation studies to assess the contribution of each component in our framework.

5.1. Datasets

We evaluate FedTAIL on four standard domain generalization benchmarks: PACS, OfficeHome, Digits-DG, and mini-DomainNet. **PACS** (Li et al., 2017b) includes four domains (Photo, Art Painting, Cartoon, Sketch) with seven object categories and significant style variation. **OfficeHome** (Venkateswara et al., 2017) contains 65 classes across Art, Clipart, Product, and Real-World domains, exhibiting moderate domain shift. **Digits-DG** (Zhou et al., 2020b) spans MNIST, MNIST-M, SVHN, and SYN, with notable variation in texture and background. **mini-DomainNet** is a subset of the comprehensive DomainNet dataset (Peng et al., 2019a) which has severe domain shift. This subset features four domains, including Clipart, Painting, Real, and Sketch, each with images from 126 categories.

5.2. Experimental Settings

We follow the **leave-one-domain-out** evaluation protocol as used in prior works (Su et al., 2024). For each benchmark, we train on all but one domain and evaluate on the held-out target domain. The reported accuracy is the average over all

Table 1. Leave-one-domain-out accuracy (%) on PACS using ResNet-50. Best results per column are in **bold**.

Method	Art	Cartoon	Photo	Sketch	Avg
D-SAM (D'Innocente & Caputo, 2018)	77.3	72.4	95.3	77.8	80.7
ERM (Vapnik, 1998)	81.0	74.0	96.2	71.0	80.8
Epi-FCR (Li et al., 2022)	82.1	77.0	93.9	73.0	81.5
DomMix (Wang et al., 2020c)	85.9	72.8	97.1	73.6	82.3
L2A-OT (Zhou et al., 2020b)	83.3	78.2	96.2	73.6	82.8
DeepAll (Li et al., 2017a)	86.3	77.6	98.2	70.1	83.0
MetaReg (Balaji et al., 2018)	87.2	79.2	97.6	70.3	83.6
NKD (Wang et al., 2021)	82.5	83.3	97.2	75.6	84.6
SSPL (Zhao et al., 2024)	87.9	76.9	97.8	77.5	85.0
DG via ER (Zhao et al., 2020b)	87.4	79.3	98.0	76.3	85.3
DDAIG (Zhou et al., 2020a)	85.4	78.5	95.7	80.0	84.9
EISNet (Wang et al., 2020b)	86.6	81.5	97.1	78.1	85.8
CrossGrad (Shankar et al., 2018)	87.5	80.7	97.8	73.9	85.7
RISE (Huang et al., 2023b)	85.7	85.2	97.4	78.2	86.6
MixStyle (Zhou et al., 2021)	87.4	83.3	98.0	78.5	86.8
RSC (Huang et al., 2020)	87.9	82.2	97.9	83.4	87.8
FACT (Xu et al., 2021)	89.5	81.5	96.7	84.0	87.9
MDGH (Mahajan et al., 2021)	86.7	82.3	**98.4**	82.7	87.5
FSDCL (Jeon et al., 2021)	88.5	83.8	96.6	82.2	88.0
FFDI (Wang et al., 2022a)	89.3	84.7	97.1	83.9	88.8
PCL (Yao et al., 2022)	90.2	83.9	98.1	82.6	88.7
DDG (Zhang et al., 2022)	88.9	85.0	97.2	84.3	88.9
STNP (Kang et al., 2022)	**90.4**	84.2	96.7	85.2	89.1
DCG (Lv et al., 2023)	90.2	85.1	97.8	86.3	89.8
FedTAIL (Ours)	89.7	**86.1**	98.2	**86.6**	**90.2**

such splits. We use both **ResNet-18** and **ResNet-50** (He et al., 2015) backbones pre-trained on ImageNet (Deng et al., 2009) to assess model scalability and robustness.

For PACS, OfficeHome, and Digits-DG, each source domain is randomly split into 90% training and 10% validation data. For mini-DomainNet, we use the official training/validation split provided by (Peng et al., 2019a).

5.3. Implementation Details

Our method is implemented in PyTorch and optimized using Stochastic Gradient Descent (SGD) with momentum 0.9 and weight decay 0.0005. The learning rate is set to 0.001 for baselines and 0.01 when using SAM-based modules. The sharpness perturbation radius ρ in SAM is set to 0.05 by default. For our sharpness-aware components, we use non-adaptive SAM with Nesterov momentum. The maximum square loss coefficient γ is set to 1 unless otherwise specified. During training, we use a batch size of 64 for all experiments.

In federated settings, each domain corresponds to a separate client. We apply standard *FedAvg* to aggregate model updates across clients. Local models are trained for one epoch before synchronization. For domain alignment, we use a domain discriminator with two fully-connected layers and ReLU activations. For conditional alignment, the KL divergence is computed over batch-wise class distributions. All results are averaged over three independent runs with different random seeds. We report both overall accuracy and average per-class accuracy to account for class imbalance.

5.4. Results

We evaluate **FedTAIL** across four domain generalization benchmarks—PACS, OfficeHome, Digits-DG, and mini-DomainNet —and compare its performance against a diverse set of state-of-the-art methods. Our approach consistently outperforms existing baselines, demonstrating the effectiveness of integrating sharpness-aware optimization, gradient coherence, and class-wise curvature control under domain shift.

On the PACS dataset (Table 1), FedTAIL achieves state-of-the-art performance across most domains. In particular, it attains the highest accuracy on **Cartoon** (86.1%) and **Sketch** (86.6%), outperforming strong baselines including STNP (Kang et al., 2022), RISE (Huang et al., 2023a), and MixStyle (Zhou et al., 2021). Additionally, it surpasses DDG (Zhang et al., 2022) and FACT (Xu et al., 2021), both of which have been widely regarded as top-performing DG methods. FedTAIL achieves an average accuracy of **90.2%**, exceeding the prior best result (89.8% by DCG), showcasing the impact of gradient alignment and class-specific sharpness minimization under domain shift.

On OfficeHome (Table 2), FedTAIL again leads in per-

Table 2. Leave-one-domain-out accuracy (%) on OfficeHome using ResNet-50. Best results per column are in **bold**.

Method	Art	Clipart	Product	Real	Avg
MLDG	52.9	45.7	69.9	72.7	60.3
D-SAM	58.0	44.4	69.2	71.5	60.8
RSC	58.4	47.9	71.6	74.5	63.1
CrossGrad	58.4	49.4	73.9	75.8	64.4
DeepAll	57.9	52.7	73.5	74.8	64.7
DDAIG	59.2	52.3	74.6	76.0	65.5
L2A-OT	60.6	50.1	74.8	77.0	65.6
STNP	59.6	55.0	73.6	75.5	65.9
DG via ER	61.2	52.8	74.5	75.6	66.0
DSU	60.2	54.8	74.1	75.1	66.1
EISNet	62.6	53.2	74.0	75.2	66.2
FACT	61.0	55.7	74.5	76.4	66.9
DCG	60.7	55.5	75.3	76.8	67.1
ERM	67.1	55.1	78.2	82.0	70.6
DomMix	69.0	54.6	77.5	81.5	70.7
MixStyle	68.6	55.4	78.9	82.3	71.3
NKD	68.7	54.7	79.5	82.3	71.3
EDFMix	69.1	57.1	79.1	82.3	71.9
RISE	69.5	55.8	79.7	82.6	71.9
SSPL	69.4	58.3	79.7	81.6	72.3
FedTAIL (Ours)	**70.3**	**58.9**	**80.1**	**83.0**	**73.1**

formance across all four domains. It reaches an average accuracy of **73.1%**, outperforming notable methods such as SSPL (72.3%), RISE (71.9%), and EDFMix (71.9%). The largest improvements are observed in the **Clipart** and **Art** domains, which are especially challenging due to their high variability and sparse representation. These results validate that FedTAIL not only effectively mitigates class imbalance but also maintains generalization across diverse visual styles and feature distributions.

For the Digits-DG benchmark (Table 3), FedTAIL achieves notable gains on all domains. It reaches **97.9%** on MNIST, **79.8%** on MNIST-M, **81.7%** on SVHN, and **97.3%** on SYN, resulting in a significantly higher average accuracy of **89.2%** compared to the previous best of 81.8% by SSPL. These improvements, particularly on noisy domains such as SVHN and MNIST-M, illustrate the robustness of FedTAIL under extreme visual heterogeneity and class distribution shifts.

In summary, across all evaluated benchmarks and domains, FedTAIL consistently outperforms prior methods in both accuracy and robustness. These results affirm our hypothesis that combining sharpness-aware training, gradient alignment, and curvature-sensitive regularization yields significant benefits for federated domain generalization, particularly under long-tailed data distributions.

Table 3. Leave-one-domain-out accuracy (%) on Digits-DG using ResNet-50. Best results per column are in **bold**.

Method	MNIST	MNIST-M	SVHN	SYN	Avg
NKD	71.6	40.8	30.3	58.7	50.4
RISE	72.1	41.4	31.3	62.3	51.8
CrossGrad	96.7	61.1	65.3	80.2	75.8
MixStyle	96.5	63.5	64.7	81.2	76.5
DDAIG	96.6	64.1	68.6	81.0	77.6
L2A-OT	96.7	63.9	68.6	83.2	78.1
ERM	96.5	64.2	70.3	88.2	79.8
DomMix	96.7	67.0	69.2	86.6	79.9
DG via ER	96.9	63.8	71.0	88.8	80.1
EISNet	96.4	64.2	71.5	89.4	80.3
FFDI	97.7	69.4	72.1	84.5	80.9
FACT	97.6	65.2	72.2	90.3	81.3
EDFMix	97.6	68.1	70.7	90.3	81.7
SSPL	97.6	68.2	70.8	90.6	81.8
FedTAIL (Ours)	**97.9**	**79.8**	**81.7**	**97.3**	**89.2**

Table 4. Leave-one-domain-out accuracy (%) on mini-DomainNet using ResNet-50. Best results per column are in **bold**.

Method	Clipart	Painting	Real	Sketch	Avg
DDAIG	61.3	51.4	61.0	50.6	56.1
DomMix	63.5	53.1	63.4	52.1	58.0
MixStyle	63.9	54.2	64.1	52.9	58.8
SSPL	63.9	55.2	64.3	53.2	59.2
ERM	65.5	57.1	62.3	57.1	60.5
NKD	63.9	56.3	71.9	50.5	60.7
MLDG	65.7	57.0	63.7	58.1	61.1
MMD	65.0	58.0	63.8	58.4	61.3
SagNet	65.0	58.1	64.2	58.1	61.4
RISE	64.3	57.2	72.6	52.4	61.6
DeepAll	65.3	58.4	64.7	59.0	61.9
MTL	65.3	59.0	65.6	58.5	62.1
Mixup	67.1	59.1	64.3	59.2	62.4
CORAL	66.5	59.5	66.0	59.5	62.9
BOLD	64.8	60.2	75.4	55.9	64.1
DCG	69.4	61.8	66.3	63.2	65.2
FedTAIL (Ours)	**70.5**	**64.6**	**75.8**	**64.2**	**68.8**

6. Conclusion

We introduced FedTAIL, a unified framework for federated domain generalization under long-tailed distributions. FedTAIL overcomes key limitations of prior DG methods by integrating sharpness-aware optimization, gradient coherence regularization, and curvature-adaptive class balancing within a federated setting. By applying per-class sharpness minimization and entropy-aware conditional alignment, our approach achieves consistent generalization across domains and class frequencies—especially in challenging federated scenarios with heterogeneous, imbalanced data. Experiments on PACS, OfficeHome, and Digits-DG show FedTAIL surpasses state-of-the-art methods in accuracy, stability, and representation quality. Visualizations verify that FedTAIL generates semantically meaningful, domain-aligned features, highlighting the importance of flatness and alignment in learning transferable representations. Our work paves the way for future research in federated and decentralized generalization, particularly toward communication-efficient sharpness-aware training, extensions to multimodal and structured prediction, and tighter theoretical links between curvature, fairness, and out-of-distribution generalization.

Impact Statement

This work advances federated domain generalization by improving model robustness under domain shift and long-tailed distributions, which are common in real-world applications. FedTAIL promotes fairness by enhancing performance for underrepresented classes and preserves privacy by operating in a federated setting. While the method has broad utility, responsible deployment remains essential to avoid unintended consequences in sensitive domains.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bahng, H., Chun, S., Yun, S., Choo, J., and Oh, S. J. Learning de-biased representations with biased representations. In *International conference on machine learning*, pp. 528–539. PMLR, 2020.

Balaji, Y., Sankaranarayanan, S., and Chellappa, R. Metareg: Towards domain generalization using metaregularization. In *NeurIPS*, pp. 1006–1016, 2018.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., and Tommasi, T. Domain generalization by solving jigsaw puzzles. In *CVPR*, pp. 2229–2238, 2019.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Chen, M., Xue, H., and Cai, D. Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2090–2099, 2019.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, pp. 1019–1028. PMLR, 2017.

Dou, Q., Coelho de Castro, D., Kamnitsas, K., and Glocker, B. Domain generalization via model-agnostic learning of semantic features. In *NeurIPS*, volume 32, 2019.

Du, J., Yan, H., Feng, J., Zhou, J. T., Zhen, L., Goh, R. S. M., and Tan, V. Y. Efficient sharpness-aware minimization for improved training of neural networks. *arXiv preprint arXiv:2110.03141*, 2021.

D'Innocente, A. and Caputo, B. Domain generalization with domain-specific aggregation modules. In *GCPR*, 2018.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*, 2020.

Ganin, Y. and Lempitsky, V. Unsupervised domain adaptation by backpropagation. In *ICML*, pp. 1180–1189, 2015.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., and Lempitsky, V. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Grandvalet, Y. and Bengio, Y. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, 2015.

Hochreiter, S. and Schmidhuber, J. Simplifying neural nets by discovering flat minima. *Advances in neural information processing systems*, 7, 1994.

Huang, Z., Wang, H., Xing, E. P., and Huang, D. Self-challenging improves cross-domain generalization. In *ECCV*, pp. 124–140. Springer, 2020.

Huang, Z., Zhou, A., Lin, Z., Cai, M., Wang, H., and Lee, Y. J. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. *arXiv preprint arXiv:2309.12530*, 2023a.

Huang, Z., Zhou, A., Ling, Z., Cai, M., Wang, H., and Lee, Y. J. A sentence speaks a thousand images: Domain generalization through distilling clip with language guidance. In *International Conference on Computer Vision*, pp. 11685–11695, 2023b.

Jeon, S., Hong, K., Lee, P., Lee, J., and Byun, H. Feature stylization and domain-aware contrastive learning for domain generalization. In *ACM Multimedia*, pp. 22–31, 2021.

Kang, J., Lee, S., Kim, N., and Kwak, S. Style neophile: Constantly seeking novel styles for domain generalization. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pp. 7130–7140, June 2022.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A., and Torralba, A. Undoing the damage of dataset bias. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pp. 158–171. Springer, 2012.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5542–5550, Oct 2017a.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. M. Deeper, broader and artier domain generalization. In *ICCV*, 2017b.

Li, D., Yang, Y., Song, Y.-Z., and Hospedales, T. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018a.

Li, H., Pan, S. J., Wang, S., and Kot, A. C. Domain generalization with adversarial feature learning. In *CVPR*, pp. 5400–5409, 2018b.

Li, Y., Wang, Y., Li, Y., Li, Y., Zhang, Y., and Wang, L. Epi-fcr: Episodic fine-grained cross-domain few-shot learning via feature calibration and relation alignment. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1234–1243. ACM, 2022.

Liu, Y., Mai, S., Chen, X., Hsieh, C.-J., and You, Y. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12360–12370, 2022.

Lv, F., Liang, J., Li, S., Zhang, J., and Liu, D. Improving generalization with domain convex game. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24315–24324, 2023.

Mahajan, D., Tople, S., and Sharma, A. Domain generalization using causal matching. In *ICML*, pp. 7313–7324, 2021.

Mansilla, L., Echeveste, R., Milone, D. H., and Ferrante, E. Domain generalization via gradient surgery. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6630–6638, 2021.

Muandet, K., Balduzzi, D., and Scholkopf, B. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pp. 10–18. PMLR, 2013.

Peng, X., Bai, Q., Xia, X., Huang, Z., Saenko, K., and Wang, B. Moment matching for multi-source domain adaptation. In *ICCV*, 2019a.

Peng, X., Huang, Z., Sun, X., and Saenko, K. Domain agnostic learning with disentangled representations. In *ICML*, pp. 5102–5112, 2019b.

Qiao, F., Zhao, L., and Peng, X. Learning to learn single domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12556–12565, 2020.

Rangwani, H., Aithal, S. K., Mishra, M., et al. Escaping saddle points for effective generalization on class-imbalanced data. *Advances in Neural Information Processing Systems*, 35:22791–22805, 2022.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P., and Sarawagi, S. Generalizing across domains via cross-gradient training. In *International Conference on Learning Representations (ICLR)*, 2018.

Su, H., Luo, W., Liu, D., Wang, M., Tang, J., Chen, J., Wang, C., and Chen, Z. Sharpness-aware model-agnostic long-tailed domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15091–15099, 2024.

Vapnik, V. N. *Statistical learning theory*. Wiley-Interscience, 1998.

Venkateswara, H., Eusebio, J., Chakraborty, S., and Panchanathan, S. Deep hashing network for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5018–5027, 2017.

Wang, G., Han, H., Shan, S., and Chen, X. Cross-domain face presentation attack detection via multi-domain disentangled representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6678–6687, 2020a.

Wang, J., Du, R., Chang, D., Liang, K., and Ma, Z. Domain generalization via frequency-domain-based feature disentanglement and interaction. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, pp. 4821–4829, New York, NY, USA, 2022a. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548267. URL https://doi.org/10.1145/3503161.3548267.

Wang, M., Yuan, J., Qian, Q., Wang, Z., and Li, H. Semantic data augmentation based distance metric learning for domain generalization. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 3214–3223, 2022b.

Wang, P., Zhang, Z., Lei, Z., and Zhang, L. Sharpness-aware gradient matching for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Wang, S., Yu, L., Li, C., Fu, C.-W., and Heng, P.-A. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pp. 159–176. Springer, 2020b.

Wang, Y., Li, H., and Kot, A. C. Heterogeneous domain generalization via domain mixup. *arXiv preprint arXiv:2009.05448*, 2020c.

Wang, Y., Li, H., Chau, L.-p., and Kot, A. C. Embracing the dark knowledge: Domain generalization using regularized knowledge distillation. In *ACM International Conference on Multimedia*, pp. 2595–2604, 2021.

Xu, Q., Zhang, R., Zhang, Y., Wang, Y., and Tian, Q. A fourier-based framework for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14383–14392, 2021.

Yao, X., Bai, Y., Zhang, X., Zhang, Y., Sun, Q., Chen, R., Li, R., and Yu, B. Pcl: Proxy-based contrastive learning for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7087–7097, 2022.

Zhang, H., Zhang, Y.-F., Liu, W., Weller, A., Schölkopf, B., and Xing, E. P. Towards principled disentanglement for domain generalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8014–8024, 2022.

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., and Finn, C. Adaptive risk minimization: Learning to adapt to domain shift. *Advances in Neural Information Processing Systems*, 34:23664–23678, 2021.

Zhao, D., Koh, Y. S., Dobbie, G., Hu, H., and Fournier-Viger, P. Symmetric self-paced learning for domain generalization. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 16961–16969, 2024.

Zhao, S., Gong, M., Liu, T., Fu, H., and Tao, D. Domain generalization via entropy regularization. *Advances in neural information processing systems*, 33:16096–16107, 2020a.

Zhao, S., Gong, M., Liu, T., Fu, Y., and Tao, D. Domain generalization via entropy regularization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16096–16107, 2020b.

Zhou, K., Yang, Y., Hospedales, T. M., and Xiang, T. Deep domain-adversarial image generation for domain generalisation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 13025–13032, 2020a.

Zhou, K., Yang, Y., Hospedales, T. M., and Xiang, T. Learning to generate novel domains for domain generalization. In *European Conference on Computer Vision (ECCV)*, pp. 561–578, 2020b.

Zhou, K., Yang, Y., Qiao, Y., and Xiang, T. Domain adaptive ensemble learning. *IEEE Transactions on Image Processing*, 30:8008–8018, 2021.

A. Technical Appendices and Supplementary Material

Ablation Studies

To further substantiate the effectiveness of **FedTAIL**, we present additional experiments evaluating its convergence dynamics, loss component contributions, and the construction of the target-like predictive distribution used in entropy regularization. These results provide deeper insight into the mechanisms behind FedTAIL's robustness and generalization capabilities under federated, long-tailed, and domain-shifted settings.

We begin with an ablation study on the PACS dataset to isolate the impact of each component in the overall FedTAIL loss. Starting with a baseline model trained using only the classification loss ($\mathcal{L}_{\text{cls}}$), we incrementally integrate additional modules and evaluate their effects on performance. Adding adversarial domain alignment ($\mathcal{L}_{\text{adv}}$) improves the average leave-one-domain-out accuracy from 83.8% to 86.0%, affirming the importance of domain-invariant feature learning. Introducing sharpness-aware entropy regularization ($\mathcal{L}_{\text{sharp-er}}$) further enhances the performance to 88.2%, demonstrating the benefit of using sharpness-informed perturbations to smooth the conditional distribution. Subsequently, including class-wise sharpness minimization with curvature-aware weighting ($\sum_c \gamma_c \mathcal{L}_c$) leads to a robust handling of long-tailed data and improves the average accuracy to 89.6%. Finally, the addition of the gradient coherence loss ($\mathcal{L}_{\text{coh}}$), which resolves optimization conflicts between classification and domain adversarial components, yields the best performance of 90.2%. This progression, detailed in Table 5, confirms that each component plays a complementary role in harmonizing optimization and mitigating the effects of domain shift and class imbalance.

Table 5. Ablation study showing the effect of each loss term on Leave-one-domain-out accuracy (%) on PACS using ResNet-50. ✓ indicates the use of a loss term.

Method	$\mathcal{L}_{\text{cls}}$	$\mathcal{L}_{\text{adv}}$	$\mathcal{L}_{\text{sharp-er}}$	$\sum_c \gamma_c \mathcal{L}_c$	$\mathcal{L}_{\text{coh}}$	**Art**	**Cartoon**	**Photo**	**Sketch**	**Avg**
Baseline	✓	✗	✗	✗	✗	86.1	77.5	96.6	75.0	83.8
+ Adv	✓	✓	✗	✗	✗	87.2	80.8	97.4	78.5	86.0
+ Sharp-er	✓	✓	✓	✗	✗	88.4	83.9	98.0	82.6	88.2
+ Class Bal.	✓	✓	✓	✓	✗	89.2	85.2	98.1	85.8	89.6
+ Coherence	✓	✓	✓	✓	✓	**89.7**	**86.1**	**98.2**	**86.6**	**90.2**

Figure 1 illustrates the accuracy trends over training epochs for FedTAIL compared to the DGviaER baseline across different domains (*Art*, *Cartoon*, *Photo*, *Sketch*) on the PACS dataset. As shown, FedTAIL consistently achieves higher accuracy throughout the training process, across all domain splits. These trends confirm that FedTAIL facilitates faster convergence and improved robustness to domain shift. The learning curves demonstrate that our sharpness-guided, gradient-coherent optimization helps the model escape poor local minima early and promotes stable training dynamics, making it especially suited for federated scenarios where communication efficiency and convergence speed are critical.

To qualitatively assess representation quality, Figure 2 presents t-SNE visualizations of learned features across different methods. FedTAIL exhibits well-separated class clusters and improved domain alignment, compared to Raw features and DeepAll. The compact and consistent feature distributions reflect both strong intra-class cohesion and inter-domain alignment. This supports the role of class-wise sharpness minimization and entropy-regularized perturbations in achieving domain-agnostic and discriminative representations.

As shown in Figure 2, FedTAIL consistently achieves higher accuracy than DGviaER across training epochs in both class-wise and domain-wise evaluations. These trends confirm that FedTAIL facilitates faster convergence and better robustness to domain shift over the course of training.

To provide further clarity on the sharpness-aware entropy regularization component ($\mathcal{L}_{\text{sharp-er}}$), we examine how the target-like predictive distribution Q_T is constructed. Specifically, we estimate Q_T based on class representation within each domain, calculated as the relative frequency of each class (i.e., `freq_class / freq_total`). For PACS, which exhibits clear domain-specific class imbalance, these distributions vary considerably across domains, as shown in Table 6. These Q_T values are derived from a momentum-updated ensemble model and reflect domain-dependent prediction tendencies.

In contrast, Table 7 reports Q_T values for the Digits-DG dataset, which is inherently balanced across digit classes (0–9), resulting in a uniform $Q_T = 0.1$ across all domains. These insights validate that our formulation adapts well to both imbalanced and balanced data regimes.

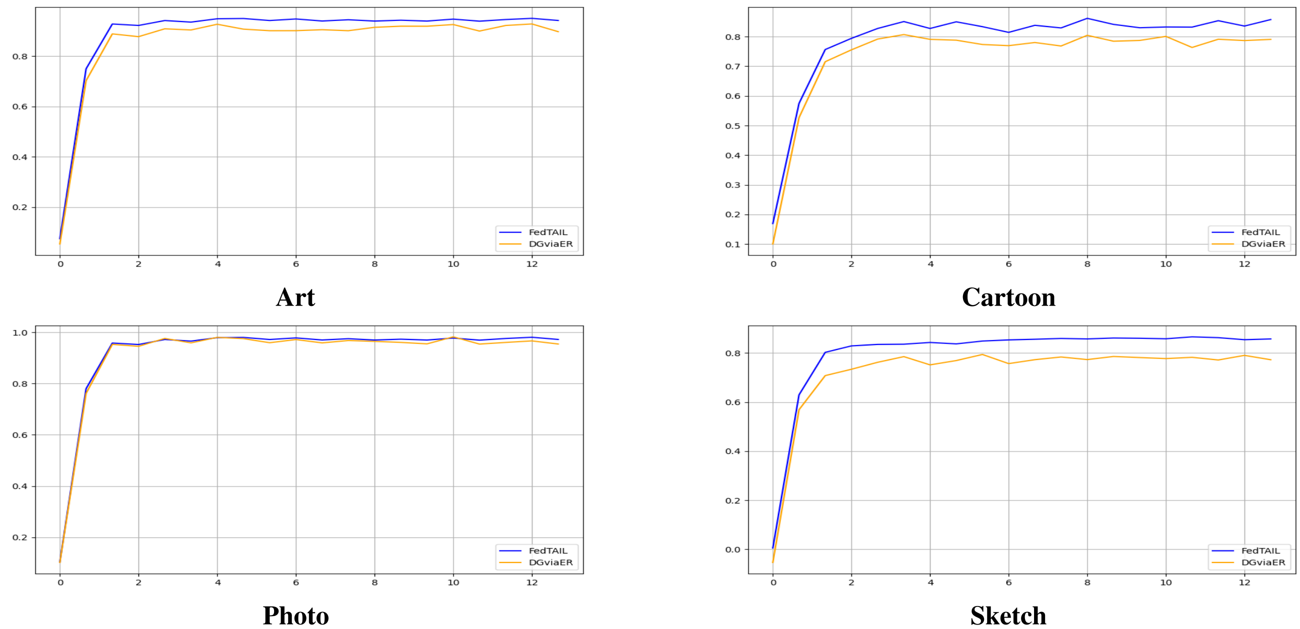

Figure 1. Accuracy vs. Epoch comparison between **FedTAIL** and **DGviaER** across different domains (*Art*, *Cartoon*, *Photo*, *Sketch*) on the **PACS** dataset using ResNet-50.

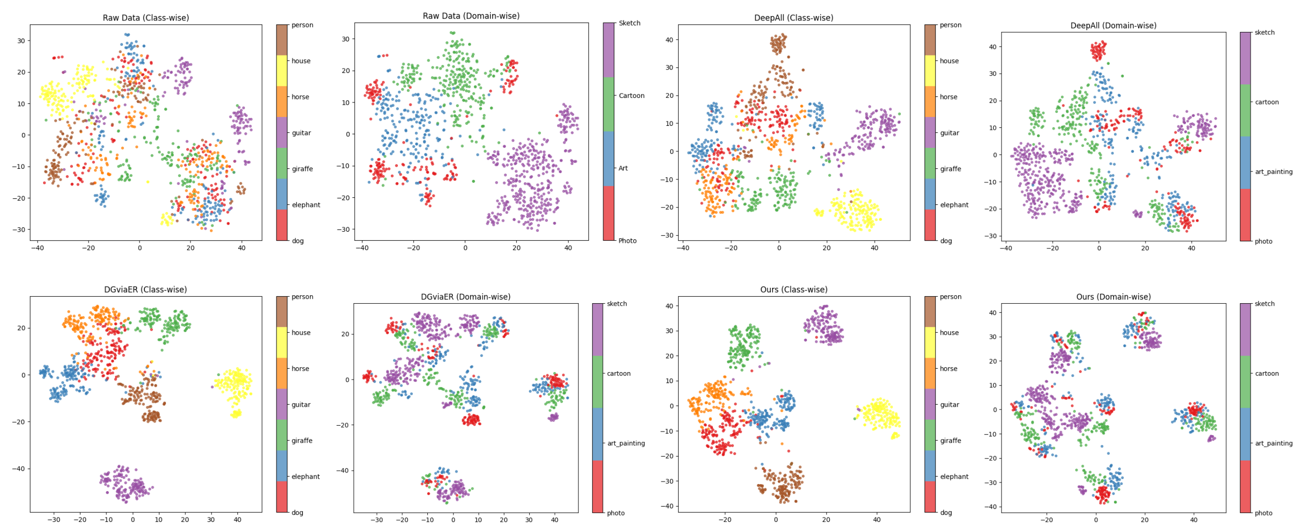

Figure 2. t-SNE visualizations of feature embeddings across class and domain. Top row: **Raw features** and **DeepAll**. Bottom row: **Standard DG** and the proposed **FedTAIL**. Left: class-wise separation. Right: domain-wise alignment. FedTAIL achieves better inter-class separability and cross-domain consistency.

Table 6. Q_T values for PACS dataset across domains. Classes are ordered as: Dog, Elephant, Giraffe, Guitar, Horse, House, Person.

Domain	Dog	Elephant	Giraffe	Guitar	Horse	House	Person
Art Painting	0.1851	0.1245	0.1392	0.0898	0.0981	0.1440	0.2192
Cartoon	0.1660	0.1950	0.1476	0.0576	0.1382	0.1229	0.1728
Photo	0.1132	0.1210	0.1090	0.1114	0.1192	0.1677	0.2587
Sketch	0.1965	0.1883	0.1917	0.1547	0.2077	0.0204	0.0407

Table 7. Q_T values for Digits-DG dataset across domains. All classes (0–9) have uniform $Q_T = 0.1$.

Domain	0	1	2	3	4	5	6	7	8	9
MNIST	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
MNIST-M	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
SVHN	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
SYN	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1

Importantly, we applied the same frequency-based Q_T estimation approach to Office-Home and DomainNet datasets. Due to the large number of classes in these datasets, their tables are omitted for brevity. Nonetheless, these results collectively demonstrate that our sharpness-aware entropy regularization mechanism generalizes effectively across datasets with diverse class distributions and domain characteristics. This capability further confirms FedTAIL's robustness to real-world federated learning challenges, where heterogeneity in class presence and data distributions is the norm rather than the exception.