Interpretable and Privacy-Preserving Federated Learning via Subspace Representations

Anonymous Authors[1]

Abstract

Federated Learning (FL) has emerged as a vital paradigm for privacy-preserving collaborative learning. While prototype-based interpretability offers explanations in centralized settings, its extension to FL is hindered by the risk of leaking sensitive prototypical patches and the challenge of aggregating heterogeneous client concepts. We propose FedGraSP, a novel manifold-aware, subspace-based framework that aggregates class evidence on the Grassmann manifold for part-level interpretable FL. By encoding representative part-level prototypes as subspaces on the Grassmann manifold and exploiting its rotation invariance, FedGraSP achieves well-generalized collaborative learning without transmitting private raw image patches. Our framework utilizes a projection-retraction-based gradient update to maintain manifold obedience and extends naturally to personalized FL by repurposing the final layer to adapt to each user's dataset distribution. Extensive evaluations on fine-grained classification datasets demonstrate that FedGraSP provides faithful, part-level explanations while maintaining high utility, bridging the gap between transparent reasoning and data minimization in decentralized environments.

1. Introduction

Federated learning (FL) has emerged as a promising framework for training models on distributed and privacy-sensitive datasets without sharing raw data directly (McMahan et al., 2017; Hard et al., 2018; Guan et al., 2024; Pati et al., 2024). When interpretable deep learning is introduced into this setting, a new challenge arises: prototype-based models explain predictions by linking them to representative training instances or image patches, which may encode sensitive information about the underlying training data. This creates a fundamental tension between two desirable goals: interpretability, which requires semantically meaningful evidence for predictions, and privacy, which requires minimizing information leakage about local training data (Zhu et al., 2019; Geiping et al., 2020). Existing approaches to interpretable federated learning largely inherit prototype-based reasoning. However, feature space based prototype which construct decision boundaries directly from backbone outputs suffers from risk of privacy leakage from invertible gradient attack (Tan et al., 2022a;b; Huh et al., 2025). In addition, pixel-level prototypes which construct representative image patch for criteria of output label expose identifiable evidence from private data (Naumova et al., 2024).

Deploying prototype-based interpretable models in FL therefore requires a new design that achieves three goals simultaneously: (a) strong global generalization from small and scattered client datasets, (b) meaningful explanation at the pixel or part level, and (c) reduced exposure of client-specific visual content to an honest-but-curious server. We propose a representation framework in which class evidence is encoded as subspaces on the Grassmann manifold rather than as identifiable prototype patches, as illustrated in Figure 1. This design preserves interpretability while reducing the risk of exposing client-specific training samples. Our approach extends the transparent basis-concept formulation of TesNet (Wang et al., 2021) to the federated setting: TesNet captures unique and discriminative concepts through an orthogonality-constrained embedding space, which leads to better generalization than Euclidean prototype-based plug-in networks such as ProtoPNet (Chen et al., 2019).

Our contributions are summarized as follows.

- **Privacy-preserving interpretable FL** We present a federated framework that learns pixel-grounded part prototypes locally and aggregates them geometrically to construct a globally transparent model. This extends prototype-based reasoning from centralized learning to privacy-preserving FL by blocking the exposure of private image parts.
- **Flexible conversion between global and local mod-**

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

els. To address data heterogeneity across clients, we propose a personalized classifier-deployed FL framework that supports flexible conversion between a single global model and personalized local models. Our framework can support either a universal global model under scattered data or personalized models which are experts on their own dataset .

2. Related Works

2.1. Pixel or part-level evidence in explainable CNN architectures

A substantial body of explainable AI has pursued evidence-based interpretability, where localized visual regions and human-inspectable exemplars justify a model's prediction. ProtoPNet performs classification by matching an input to learned prototypical parts and aggregating their evidence (Chen et al., 2019), and Deformable ProtoPNet improves robustness of part evidence under pose changes via spatially flexible prototypes (Donnelly et al., 2022). PIP-Net emphasizes intuitive prototypes and sparse scoring sheets, making final decisions traceable to a small set of part-level matches (Nauta et al., 2023). TesNet frames interpretability through a transparent embedding layer in which class-wise subspaces are spanned by basis concepts and local feature-map patches are projected onto them, enabling a structured decomposition of evidence (Wang et al., 2021). Complementing these architectures, structured-concept pruning methods employ Grassmann-manifold distance metrics to reconstruct and separate subspaces of different dimensions after slimming operations (Wang et al., 2023).

2.2. Prototypes and interpretability in federated learning

In federated learning, "prototype" has been adopted primarily to indicate an abstract representative of a class in latent feature space, used to cope with statistical/model heterogeneity and non-IID data. FedProto defines a class prototype at each client as the mean of embedding vectors and aggregates prototypes to regularize local training (Tan et al., 2022a), while FedPCL incorporates prototypes into a supervised contrastive learning objective with a lightweight learnable projection network (Tan et al., 2022b). More recent variants extend the same alignment idea to non-Euclidean geometries and classifier constraints. HyperFed constructs fixed, uniformly distributed class prototypes in hyperbolic space (Liao et al., 2023). SphereFed fixes and shares a classifier whose weight vectors span a unit hypersphere (Dong et al., 2022).

Smaller line of work has begun to explicitly consider the interaction between interpretability and privacy. AGP studies the privacy and interpretability trade-off in gradient-based FL by using Grad-CAM to identify important channels and selectively inject Gaussian noise into shared model parameters. However, its explanations remain post-hoc and privacy protection is still achieved through perturbation because of inherent limitation for gradient trackable network (Li et al., 2024). MyTH moves closer to interpretable FL by adapting ProtoPNet to a biomedical federated setting and comparing local and global patch prototypes (Naumova et al., 2024). However, because MyTH still relies on sharing and averaging explicit patch-level prototypes, collaborative interpretability remains tied to directly client-derived visual evidence. To the best of our knowledge, no prior work has simultaneously preserved the part-level transparency emphasized by MyTH and the collaborative aggregation benefits of prototype-based FL while avoiding destructive noise injection as the primary privacy mechanism.

3. Preliminaries on Riemannian Manifold Learning

A Riemannian manifold is an optimizable smooth manifold equipped with both a coordinate representation and a Riemannian metric that together define all geometric concepts. By working in local coordinates, one gains access to notions such as distances, angles, and gradients in a computationally tractable way.

3.1. Riemannian Manifold

Definition 3.1 (Riemannian Manifold)**.** *A Riemannian manifold $(\mathcal{M}, \rho)$ is a smooth manifold $\mathcal{M}$ equipped with a Riemannian metric ρ, defined as a smoothly varying inner product on the tangent space $T_x\mathcal{M}$ at each point $x \in \mathcal{M}$, $\rho_x(\cdot,\cdot) : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$.*

Definition 3.2 (Tangent Space)**.** *Let $\mathcal{M} \subset \mathbb{R}^n$ be an m-dimensional smooth manifold and $\phi : \mathcal{V} \subset \mathbb{R}^m \to \mathcal{M}$ a smooth local parametrization with $x = \phi(a)$. The tangent space of $\mathcal{M}$ at x, denoted $T_x\mathcal{M}$, is defined as*

$$T_x\mathcal{M} := \mathrm{Range}\big(d\phi_a\big) = \mathrm{span}\left\{\frac{\partial\phi}{\partial q^1}(a), \ldots, \frac{\partial\phi}{\partial q^m}(a)\right\},$$

where $d\phi_a$ is the differential of ϕ at a and $q = (q^1, \ldots, q^m) \in \mathbb{R}^m$ denotes local coordinates. This construction is independent of choice of local parametrization.

3.2. Grassmann Manifold

The Stiefel manifold $\mathrm{St}(k, n)$ is defined as the set of all orthonormal k-frames in $\mathbb{R}^n$, i.e., $\mathrm{St}(k, n) := \{X \in \mathbb{R}^{n\times k} \mid X^\top X = I_k\}$. The Grassmann manifold $\mathrm{Gr}(k, n)$, $n \geq k$, imposes a stricter equivalence by identifying all frames that span the same subspace, and is formally defined as the set of all k-dimensional linear subspaces of $\mathbb{R}^n$. In the projector form, each subspace is identified with its orthogonal pro-

jection matrix, and the manifold is characterized as $n \times n$ matrix which satisfies

$$\mathrm{Gr}(k,n) := \left\{ P \;\middle|\; P^\top = P,\ P^2 = P,\ \mathrm{rank}(P) = k \right\}.$$

Riemannian Metrics. A Riemannian manifold admits multiple valid metrics, each inducing different notions of distance and curvature. Two common choices for Riemannian metrics on $\mathrm{Gr}(k,n)$ are the canonical and projection metrics.

- *Canonical Metric.* The canonical Riemannian metric is a rotation-invariant metric on $\mathrm{Gr}(k,n)$, inherited from the Frobenius inner product on the ambient space of symmetric matrices $\mathrm{Sym}(n)$. At a projector $P \in \mathrm{Gr}(k,n)$, it is defined on the tangent space $\Delta_1, \Delta_2 \in T_P\mathrm{Gr}(k,n)$ as

$$g_P(\Delta_1, \Delta_2) = \frac{1}{2}\,\mathrm{tr}(\Delta_1^\top \Delta_2). \tag{1}$$

  The canonical metric is the foundation of all intrinsic geometric operations on $\mathrm{Gr}(k,n)$.

- *Projection (Extrinsic) Metric.* The projection metric measures distance directly between two projectors in the ambient space $\mathrm{Sym}(n)$, without reference to the tangent space, as

$$d_P(P,Q) = \frac{1}{\sqrt{2}}\,\|P - Q\|_F. \tag{2}$$

  As an extrinsic quantity, d_P is nonetheless rotation-invariant.

Tangent Space. $T_P\mathrm{Gr}(k,n) \subset \mathbb{R}^{n\times n}$ is given by

$$T_P\mathrm{Gr}(k,n) = \{\Delta \mid \Delta^\top = \Delta,\ \Delta = P\Delta P^\perp + P^\perp \Delta P\}, \tag{3}$$

where $P^\perp = I_n - P$ denotes the orthogonal complement projector.

Gradient. Given the Euclidean derivative $\nabla f(X)$ of an objective $f(X)$ at X, the Riemannian gradient is obtained by projection via π_{T_X} as $\nabla_{\mathcal{M}} f(X) = \pi_{T_X}\big(\nabla f(X)\big)$. In case of Grassmann manifold adoption, the projection $\Pi_{T_P\mathrm{Gr}} : \mathrm{Sym}(n) \to T_P\mathrm{Gr}(p,n)$ can be defined as

$$S \mapsto (I_n - P)SP + PS(I_n - P), \tag{4}$$

which is induced from (3) by only leaving the components that change the range of P.

4. Proposed Method

Figure 2 illustrates the local architecture used in Federated Learning with Grassmannian Subspace Prototypes (FedGraSP). Each client employs a standard CNN backbone to map an input image x into a latent feature map. To reduce channel dimensionality and alleviate the computational cost of subsequent operations, we append a 1×1 convolutional bottleneck, which produces $\mathbf{Z} \in \mathbb{R}^{B\times n\times H\times W}$.

We propose a simple yet effective federated framework that learns (i) a strong global model from decentralized and limited client data, and (ii) a personalized model for each client by lightweight adaptation. We decompose the network into a shared body and a client-adaptive head. The body consists of the backbone and bottleneck convolutional layer w_f together with the Grassmannian prototype layer $\mathbf{P}$, whereas the head is the class-combination matrix $\mathbf{G}$. The overall procedure consists of three stages per communication round:

1. **Local update.** Each client updates the shared body parameters $(w_f, \mathbf{P})$ using its private dataset $\mathcal{D}_i$, while keeping the head $\mathbf{G}$ fixed.
2. **Server aggregation.** The server aggregates w_f in Euclidean space via parameter averaging and aggregates $\mathbf{P}$ on the Grassmann manifold using a geometry-aware rule.
3. **Personalization.** After completing the global training rounds, each client fine-tunes the head $\mathbf{G}$ on $\mathcal{D}_i$ while freezing $(w_f, \mathbf{P})$ received from the server. Since $\mathbf{G}$ is a square matrix under our score design, personalization reduces to a simple plug-and-play update of $\mathbf{G}$.

Although our local architecture is conceptually related to TesNet, we parameterize class concepts as subspaces rather than explicit orthonormal frames and accordingly adapt the learning procedure (Wang et al., 2021). The representation becomes independent of the ordering and orientation of basis vectors by operating on subspaces. Furthermore, it mitigates the risk of exposing client-specific patch-level information that could otherwise be inferred from directly shared prototype vectors.

Given an input x_i, the local model produces logits $\phi_c(x_i; f, \mathbf{P}, h)$ by passing class scores through the head layer (Fig. 2). Let $\mathbf{Z}_i = f(x_i)$ and let p denote a patch embedding extracted from $\mathbf{Z}_i$. The score of class c for the i-th sample is computed by summing the top-k projection energies as $s_{ci} = \sum_{p\in \text{top-}k(\mathbf{Z}_i)} \left\|\mathbf{P}_\mathbf{c}^\top p\right\|_2^2$. The identification loss is the standard cross-entropy

$$\mathcal{L}_{id} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{ic} \log \phi_c(x_i; f, \mathbf{P}, h), \tag{5}$$

where y_{ic} is a one-hot label indicator and N is the number of training samples.

Class-aware compactness-separation. We adopt a

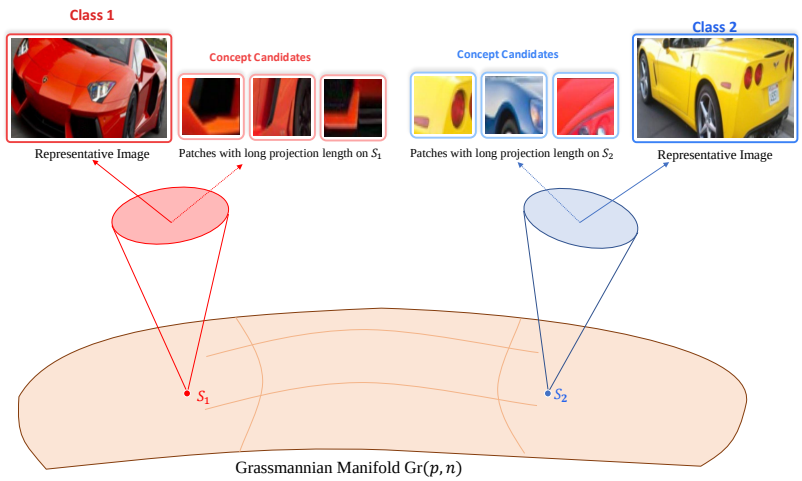


Figure 1. Illustration of the encoded subspaces on the Grassmannian manifold. Unlike fixed prototypes, FedGraSP learns a subspace (represented by colored circles) for each class that captures unique characteristics while preserving privacy.

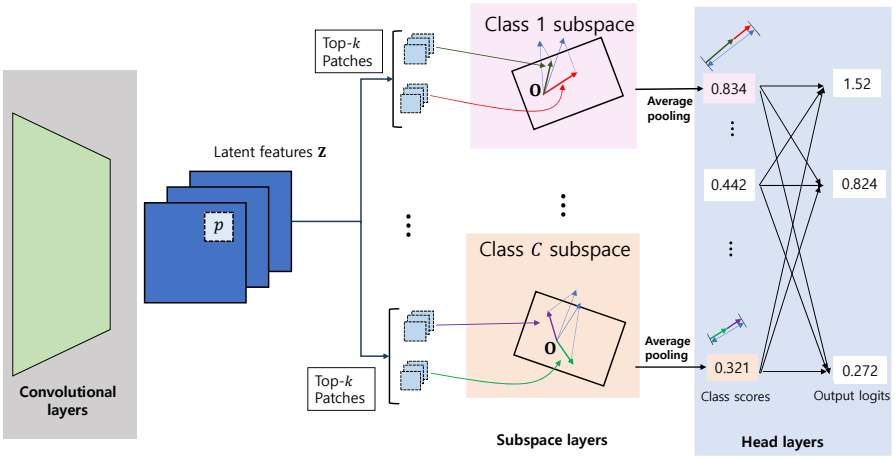


Figure 2. Overall architecture of local model for FedGraSP. The main difference with TesNet architecture is that the score for each class is derived based on the projection length to subspace represented by Grassmann manifold. This makes it difficult for the server to reconstruct the exact image part of the local data.

compactness-separation objective as

$$\mathcal{L}_{cs} = \mathcal{L}_{compactness} + \mu \mathcal{L}_{separation},$$

$$\mathcal{L}_{compactness} = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{C} \mathbb{1}_{\{j=c_i\}} \tilde{s}_{ji},$$

$$\mathcal{L}_{separation} = \frac{1}{N} \sum_{i=1}^{N} \min_{j:\, j \neq c_i} \tilde{s}_{ji}.$$

$\tilde{s}_{ji}$ denotes the normalized score of class j for the i-th image, which is the normalized projection length onto the ground-truth subspace c_i as

$$\tilde{s}_{ji} = \sum_{p \in \text{top-k}(\mathbf{Z}_i)} \frac{\|\mathbf{P_j}^\top p\|^2}{\|p\|^2}.$$

Since the subspace layer consists of orthogonal bases with unit norm, we employ normalized scores to ensure that symbolic patches align meaningfully with the target subspace.

The overall optimization problem, which balances all previously introduced components, is formulated as $\mathcal{L}_{total} = \mathcal{L}_{id} + \lambda_2 \mathcal{L}_{cs}$, where λ_1 and λ_2 are hyperparameters that control the relative contributions of each term.

While the Euclidean convolutional layers can be optimized

Algorithm 1 Consensus update on $\mathrm{Gr}(p, n)$

Input: Projection matrices $\{\Pi_{kc}\}_{k=1}^{N}$, step size α, adjacency weights $\{a_j\}_{j=1}^{|\mathcal{S}|}$
Output: Consensus projectors $\{\Pi_c^\star\}_{c=1}^{C}$
  for $c = 1$ **to** C **do**
    Initialize $\Pi_c^\star \leftarrow \Pi_{1c}$
    for $t = 1$ **to** T **do**
      $\Xi \leftarrow \sum_{j=1}^{N} a_j \Pi_{jc}$
      $\Theta \leftarrow \Pi_c^\star \Xi (I_n - \Pi_c^\star) + (I_n - \Pi_c^\star)\Xi \Pi_c^\star$ {Tangent direction}
      $S \leftarrow \Pi_c^\star + \alpha\Theta$
      Compute EVD: $S = Q\Lambda Q^\top$
      Retain the p leading eigenvectors $Q_{[p]}$
      $\Pi_c^\star \leftarrow Q_{[p]} Q_{[p]}^\top$ {Retraction onto $\mathrm{Gr}(p, n)$}
    end for
  end for

using standard optimizers such as Adam, the subspace layer requires geometry-aware optimization. Specifically, we employ projection–retraction-based gradient updates to ensure that the parameters remain on the underlying manifold throughout training. A first-order accurate, efficient update in discrete time then follows the gradient update and retraction step in (4) as $\mathbf{P}_c^{(t+1)} = R_{\mathbf{P}_c^{(t)}}(\mathbf{P}_c^{(t)} \nabla_{\mathbf{P}_c^{(t)}} \mathcal{L}_{total} \mathbf{P}_c^{(t)\perp} + \mathbf{P}_c^{(t)\perp} \nabla_{\mathbf{P}_c^{(t)}} \mathcal{L}_{total} \mathbf{P}_c^{(t)})$. After full iterative training with $\mathcal{L}_{total}$, we train a single FC layer at the stage of personalization. The corresponding loss enforces a sparse constraint on the weight matrix $\mathbf{G}$ to follow the derived score while capturing the relationship among class scores. The loss for last layer is $\mathcal{L}_h = \mathcal{L}_{id} + \lambda_4 \sum_{c=1}^{C} \sum_{j \neq c} |\mathbf{G}_{(c,j)}|$. FedGraSP treats $\mathbf{G}$ as a means to align global subspace representations with client-specific label distributions rather than as a classifier.

Since all layers except the subspace prototype layer lie in Euclidean space, the server can aggregate their parameters by linear averaging. In contrast, a naive Euclidean average of subspace parameters generally does not remain on the Grassmann manifold and can lead to degenerate or unstable aggregated subspaces. Therefore, we aggregate subspaces by computing an intrinsic Riemannian mean on the Grassmann manifold to obtain a meaningful global prototype. Let the server receive class-wise subspace prototypes from N clients in the form of rank-p orthogonal projectors $\{\Pi_j\}_{j=1}^{N} \subset \mathbb{R}^{n \times n}$. This can be identified with the situation in which the server is connected to all clients with uniform weights $\{a_{jk}\}_{j=1}^{N}$, as we need to find the central point of the polytope. We can formulate smooth agreement potential to be maximized as

$$\mathcal{P}_L(\{\Pi_k\}) = \frac{1}{2N^2} \sum_{j} a_{jk} \operatorname{tr}\left(\Pi_j^\top \Pi_k\right). \tag{6}$$

Under the canonical Riemannian metric-based gradient update, the consensus update procedure induce not to derive the intersection point for N geodesics which necessitate complicated matrix exponential. Consensus update has a closed-form Riemannian gradient by (4) and EVD-based retraction, as summarized in Algorithm 1.

5. Experiment

Datasets. Stanford Dogs dataset, which contains 120 dog species (Khosla et al., 2011) and Stanford Cars dataset, which includes 196 car models (Krause et al., 2013).We split each user's dataset in 80:20 ratio for train and test.

Local model structure. We employ ResNet50 as backbone network f, followed by two consecutive 1×1 convolutional layers to reduce the output channel dimension to 64. FedGraSP is trained using a projection matrix representation defined as the Grassmann manifold $\mathrm{Gr}(64, 3)$.

Implementation details. We run 3 local epochs for Stanford Cars with 80 global epochs and 5 local epochs for Stanford Dogs with 30 global epochs. There are 8 clients participating in every communication round. We simulate label heterogeneity using a Dirichlet distribution for the non-IID setting. Smaller α corresponds to a higher degree of label imbalance. We adopt the same initial learning rate as TesNet (Wang et al., 2021) and maintain it for whole global epochs.

Baselines. As we propose a flexible FL framework for personal and global model construction, we construct different baselines for each scenario. 'Centralized TesNet' trains TesNet under a centralized setting (Wang et al., 2021). 'MyTH' trains ProtoPNet, a transparent Euclidean-space prototype building method in an FL framework (Chen et al., 2019; Naumova et al., 2024). 'TesNet with Stiefel impose identical local training loss, with strict observance on Stiefel-manifold of subspace layer. It is aggregated with rotation-invariant metric of subspace layer (Sarlette & Sepulchre, 2009).'Personalized TesNet' runs local training as in TesNet while aggregating all layers by FedAvg, with the subspace layer left frozen since it has no domain for consistent mean-point computation.

Across fine-grained classification benchmarks, methods that enforce strict Grassmann-manifold constraints significantly outperform "TesNet with Stiefel". This advantage stems from the approximated subspace representation, which encompasses diverse representative patches rather than collapsing each class onto a single prototype vector. FedGraSP outperforms its variants most clearly on Stanford Cars, where implicit data distributions and severe class imbalance at low α make prototype-based approaches brittle. The subspace representation generalizes well by absorbing such variability. MyTH also performs competitively

Table 1. Global Model Test Accuracy on Stanford Cars .

Method	IID	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
Centralized TesNet		84.20 ± 1.25		
TesNet with Stiefel	32.11 ± 1.57	18.12 ± 1.52	18.25 ± 1.44	19.79 ± 2.09
FedGraSP	$\mathbf{69.55 \pm 1.97}$	$\mathbf{56.12 \pm 1.75}$	$\mathbf{56.47 \pm 1.22}$	$\mathbf{57.35 \pm 2.73}$
MyTH	33.64 ± 1.85	31.66 ± 4.11	29.16 ± 2.34	29.47 ± 3.11

Table 2. Personalized Model Test Accuracy on Stanford Cars.

Method	IID	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
TesNet with Stiefel	32.11 ± 1.91	19.36 ± 1.74	19.13 ± 1.61	21.34 ± 2.09
FedGraSP	$\mathbf{69.63 \pm 1.83}$	$\mathbf{57.11 \pm 1.46}$	$\mathbf{56.81 \pm 1.45}$	$\mathbf{57.88 \pm 2.70}$
Personalized TesNet	13.06 ± 2.95	10.35 ± 4.12	8.85 ± 1.89	9.52 ± 2.52
MyTH	33.80 ± 1.93	31.67 ± 4.09	29.20 ± 2.34	29.44 ± 3.13

Table 3. Global Model Test Accuracy on Stanford Dogs.

Method	IID	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
Centralized TesNet		82.73 ± 0.08		
TesNet with Stiefel	79.85 ± 0.24	37.29 ± 1.25	61.08 ± 1.62	66.79 ± 0.20
FedGraSP	$\mathbf{80.09 \pm 0.10}$	35.00 ± 3.39	61.51 ± 1.39	71.12 ± 0.29
MyTH	73.84 ± 0.62	$\mathbf{47.89 \pm 1.56}$	$\mathbf{71.86 \pm 1.34}$	$\mathbf{72.55 \pm 1.30}$

Table 4. Personalized Model Test Accuracy on Stanford Dogs.

Method	IID	$\alpha = 0.1$	$\alpha = 0.5$	$\alpha = 1$
TesNet with Stiefel	79.62 ± 0.26	51.08 ± 1.23	66.35 ± 1.43	69.43 ± 0.56
FedGraSP	80.08 ± 0.18	$\mathbf{61.91 \pm 3.18}$	68.74 ± 1.22	$\mathbf{72.55 \pm 0.49}$
Personalized TesNet	$\mathbf{82.25 \pm 0.30}$	47.99 ± 5.40	61.43 ± 10.99	65.83 ± 4.85
MyTH	73.84 ± 0.63	47.09 ± 1.39	$\mathbf{71.92 \pm 1.37}$	$\mathbf{72.55 \pm 1.26}$

in several settings, particularly for the global model, by exploiting the expressivity of Euclidean prototypes. Under the personalized setting, the "Personalized TesNet" baseline suffers from severe overfitting, as it omits subspace aggregation and relies entirely on local optimization, underscoring the importance of maintaining manifold consistency to preserve the intended orthogonality constraints in federated training. The performance gain of personalized FedGraSP arises from a redefined role of the final fully connected (FC) layer as a lightweight personalization module: rather than acting as a classifier, as in TesNet, it aligns the global subspace representations with client-specific label distributions by reweighting the class-specific similarity scores. This enables effective personalization without compromising global consistency, and yields stronger performance than MyTH, which relies on unconstrained subspace representations and therefore offers weaker privacy guarantees.

6. Conclusion

We have proposed FedGraSP, a novel part-level prototype-constructing interpretable FL framework that utilizes the permutation invariance of the Grassmann manifold and avoids patch-transparent prototype construction by taking advantage of rotation invariance. We propose a unique method for local training, and the projection-metric-based distance optimization contributes to prompt mean-point computation under strict manifold constraints. FedGraSP also facilitates the conversion between the global model and custom private models by changing the score-extracting technique for each class. Numerical results show that FedGraSP outperforms other baselines that expose private image patch on harsh condition and prove its high generalization capability. Still, it needs more improvement on images with implicit features or especially scant data.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Dong, X., Zhang, S. Q., Li, A., and Kung, H. Spherefed: Hyperspherical federated learning. In *European Conference on Computer Vision*, pp. 165–184. Springer, 2022.

Donnelly, J., Barnett, A. J., and Chen, C. Deformable protopnet: An interpretable image classifier using deformable prototypes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10265–10275, June 2022.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting gradients – how easy is it to break privacy in federated learning? *CoRR*, abs/2003.14053, 2020.

Guan, H., Yap, P.-T., Bozoki, A., and Liu, M. Federated learning for medical image analysis: A survey. *Pattern recognition*, 151:110424, 2024.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Huh, Y., Kim, B., and Choi, W. Feature reconstruction aided federated learning for image semantic communication. *arXiv preprint arXiv:2508.02048*, 2025.

Khosla, A., Jayadevaprakash, N., Yao, B., and Fei-Fei, L. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 554–561, 2013.

Li, Z., Chen, H., Ni, Z., Gao, Y., and Lou, W. Towards adaptive privacy protection for interpretable federated learning. *IEEE Transactions on Mobile Computing*, 23 (12):14471–14483, 2024.

Liao, X., Liu, W., Chen, C., Zhou, P., Zhu, H., Tan, Y., Wang, J., and Qi, Y. Hyperfed: Hyperbolic prototypes exploration with consistent aggregation for non-iid data in federated learning. In Elkind, E. (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pp. 3957–3965. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/440. URL https://doi.org/10.24963/ijcai.2023/440. Main Track.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Naumova, K., Devos, A., Karimireddy, S. P., Jaggi, M., and Hartley, M.-A. Mythisyourthat for interpretable identification of systematic bias in federated learning for biomedical images. *npj Digital Medicine*, 7(1):238, 2024. doi: 10.1038/s41746-024-01226-1. URL https://pubmed.ncbi.nlm.nih.gov/39242810/.

Nauta, M., Schlötterer, J., van Keulen, M., and Seifert, C. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2744–2753, June 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Nauta_PIP-Net_Patch-Based_Intuitive_Prototypes_for_Interpretable_Image_Classification_CVPR_2023_paper.html.

Pati, S., Kumar, S., Varma, A., Edwards, B., Lu, C., Qu, L., Wang, J. J., Lakshminarayanan, A., Wang, S.-h., Sheller, M. J., et al. Privacy preservation for federated learning in health care. *Patterns*, 5(7), 2024.

Sarlette, A. and Sepulchre, R. Consensus on homogeneous manifolds. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*. IEEE, 2009.

Tan, Y., Long, G., Liu, L., Zhou, T., Lu, Q., Jiang, J., and Zhang, C. Fedproto: Federated prototype learning across heterogeneous clients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):8432–8440, 2022a. doi: 10.1609/aaai.v36i8.20819.

Tan, Y., Long, G., Ma, J., Liu, L., Zhou, T., and Jiang, J. Federated learning from pre-trained models: A contrastive learning approach. *Advances in neural information processing systems*, 35:19332–19344, 2022b.

Wang, J., Liu, H., Wang, X., and Jing, L. Interpretable image recognition by constructing

transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 895–904, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Wang_Interpretable_Image_Recognition_by_Constructing_Transparent_Embedding_Space_ICCV_2021_paper.html.

Wang, J., Liu, H., and Jing, L. Transparent embedding space for interpretable image recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(5): 3204–3219, 2023.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. *CoRR*, abs/1906.08935, 2019.

A. Geometric Interpretation of the Grassmann Representation

Proposition A.1 (Fiber-dimension interpretation of Grassmann)**.** *Consider the quotient map* $\rho : \mathrm{St}(k,n) \to \mathrm{Gr}(k,n)$, $\rho(Y) = YY^\top$. *Then* ρ *is surjective and the fiber over* $P = YY^\top \in \mathrm{Gr}(k,n)$ *is*

$$\rho^{-1}(P) := \{YQ : Q \in O(k)\}, \tag{7}$$

$$\dim \rho^{-1}(P) = \dim O(k) = \frac{k(k-1)}{2}. \tag{8}$$

Therefore, a Grassmann point uniquely determines the underlying subspace, but not the particular orthonormal basis representing it.

Proof. Let $P = YY^\top$ with $Y \in \mathrm{St}(k,n)$. First, if $Q \in O(k)$, the mapping becomes

$$\rho(YQ) = (YQ)(YQ)^\top = YQQ^\top Y^\top = YY^\top = P,$$

and leads to

$$\{YQ : Q \in O(k)\} \subseteq \rho^{-1}(P).$$

Conversely, assuming $Z \in \rho^{-1}(P)$, we have

$$ZZ^\top = P = YY^\top.$$

Since $YY^\top$ and $ZZ^\top$ are the orthogonal projectors onto $\mathrm{range}(Y)$ and $\mathrm{range}(Z)$, respectively, the equality $ZZ^\top = YY^\top$ implies

$$\mathrm{range}(Z) = \mathrm{range}(Y).$$

Therefore, Y and Z are two orthonormal bases of the same k-dimensional subspace. Hence there exists a matrix $Q \in \mathbb{R}^{k\times k}$ such that $Z = YQ$. Since $Z^\top Z = I_k$ and $Y^\top Y = I_k$, we obtain

$$I_k = Z^\top Z = Q^\top Y^\top YQ = Q^\top Q,$$

so $Q \in O(k)$ and

$$\rho^{-1}(P) = \{YQ : Q \in O(k)\}.$$

Since $\dim O(k) = \frac{k(k-1)}{2}$, the dimension formula is valid. □

B. Privacy Interpretation via Non-Identifiability

Proposition B.1 (Dimension of the ambiguity set under a linear release map)**.** *Let* $x \in \mathbb{R}^m$ *denote a private patch vector, and suppose the released information is*

$$y = Ax, \qquad A \in \mathbb{R}^{r\times m}.$$

Assume that $\mathrm{rank}(A) = \rho < m$. *Then the feasible set*

$$\mathcal{F}(y) := \{z \in \mathbb{R}^m : Az = y\}$$

is an affine subspace of dimension $m - \rho$. *In particular,* x *is not uniquely determined by* y, *and there are infinitely many private patches consistent with the same released information.*

Proof. Since $y = Ax$, the set $\mathcal{F}(y)$ is nonempty. Let $x_0 \in \mathcal{F}(y)$ be any particular solution. We claim that

$$\mathcal{F}(y) = x_0 + \ker(A).$$

First, let $z \in \mathcal{F}(y)$. Then

$$A(z - x_0) = Az - Ax_0 = y - y = 0,$$

so $z - x_0 \in \ker(A)$, and hence $z \in x_0 + \ker(A)$. Conversely, assuming $z = x_0 + v$ with $v \in \ker(A)$, we obtain

$$Az = A(x_0 + v) = Ax_0 + Av = y + 0 = y,$$

so $z \in \mathcal{F}(y)$ and

$$\mathcal{F}(y) = x_0 + \ker(A).$$

Thus $\mathcal{F}(y)$ is an affine translation of $\ker(A)$, and its dimension is

$$\dim \mathcal{F}(y) = \dim \ker(A) = m - \operatorname{rank}(A) = m - \rho,$$

by the rank–nullity theorem. Because $\rho < m$, we have $m - \rho > 0$, and $\mathcal{F}(y)$ contains infinitely many points. The private patch x cannot be uniquely recovered from the released information y. □

Remark B.2 (Local linearization for nonlinear release maps)**.** *Suppose the released representation is given by a smooth map $g : \mathbb{R}^m \to \mathbb{R}^r$. Around a point $x_0 \in \mathbb{R}^m$, we have the first-order expansion*

$$g(x_0 + \delta) = g(x_0) + J_g(x_0)\,\delta + O(\|\delta\|^2),$$

where $J_g(x_0) \in \mathbb{R}^{r \times m}$ is the Jacobian matrix of g at x_0. Therefore, to first order, indistinguishable perturbations δ satisfy

$$J_g(x_0)\,\delta = 0.$$

If $\operatorname{rank}(J_g(x_0)) = \rho < m$, then the first-order ambiguity space has dimension

$$\dim \ker(J_g(x_0)) = m - \rho.$$

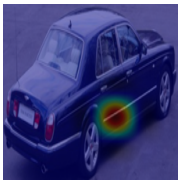
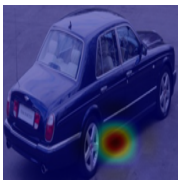
Hence the private variable is locally non-identifiable up to first order.

Remark B.3 (Privacy interpretation)**.** *Proposition B.1 shows that privacy can be interpreted as* non-identifiability*: the larger the quantity*

$$m - \rho = \dim \ker(A),$$

the more degrees of freedom remain hidden from an adversary. In this sense, the dimension of the feasible set $\mathcal{F}(y)$ quantifies how hard it is to infer the exact private patch from the released representation.

Evidence for this car being a Bentley Continental Flying Spur Sedan 2007

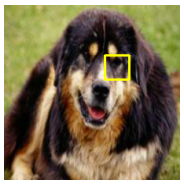
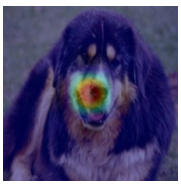
Original image (with top patch)	Activation map	Class Score	Diagonal weight	Final logits
		3.404	$\times$ 1	$=$ 3.404
		2.756	$\times$ 1	$=$ 2.756
		2.054	$\times$ 1	$=$ 2.054
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Evidence for this car being an Acura TSX Sedan 2012

Original image (with top patch)	Activation map	Class Score	Diagonal weight	Final logits
		3.526	$\times$ 1.097	$=$ 3.870
		3.524	$\times$ 1.097	$=$ 3.868
		3.385	$\times$ 1.097	$=$ 3.715
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Figure 3. Example of class scores derived from trained models in non iid-setting ($\alpha = 1$) with Stanford Dogs for a prediction. (Top) Visualized patch from global model (Bottom) Visualized patch from client model

Evidence for this dog being a Tibetan mastiff

Original image (with top patch)	Activation map	Class Score	Diagonal weight	Final logits
		3.444	$\times$ 1	$=$ 3.444
		3.443	$\times$ 1	$=$ 3.443
		3.443	$\times$ 1	$=$ 3.443
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Evidence for this dog being a pug

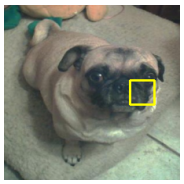
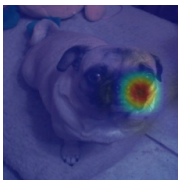
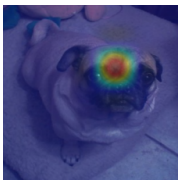
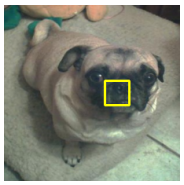
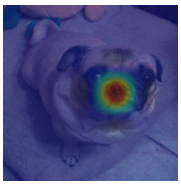
Original image (with top patch)	Activation map	Class Score	Diagonal weight	Final logits
		3.473	$\times$ 2.390	$=$ 8.300
		3.472	$\times$ 2.390	$=$ 8.297
		3.470	$\times$ 2.390	$=$ 8.295
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Figure 4. Example of class scores derived from trained models in non iid-setting ($\alpha = 1$) with Stanford Cars for a prediction. (Top) Visualized patch from global model (Bottom) Visualized patch from client model