

Beyond Federated Prototype Learning: Learnable Semantic Anchors with Hyperspherical Contrast for Domain-Skewed Data

Lele Fu^{1,2}, Sheng Huang^{1,2}, Yanyi Lai¹, Tianchi Liao¹, Chuanfu Zhang², Chuan Chen^{1*}

¹School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

²School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China

{fulle, huangsh253, laiyy28, liaotch}@mail2.sysu.edu.cn, {zhangchf9, chenchuan}@mail.sysu.edu.cn

Abstract

Federated prototype learning is in the spotlight as global prototypes are effective in enhancing the learning of local representation spaces, facilitating the ability to generalize the global model. However, when encountering domain-skewed data, conventional federated prototype learning is susceptible to two dilemmas: 1) Local prototypes obtained by averaging intra-class embedding carry domain-specific markers, the margins among aggregated global prototypes could be attenuated and detrimental to inter-class separation. 2) Local domain-skewed embedding may not exhibit a uniform distribution in Euclidean space, which is not conducive to the prototype-induced intra-class compactness. To address the two drawbacks, we go beyond conventional paradigm of federated prototype learning, and propose learnable semantic anchors with hyperspherical contrast (FedLSA) for domain-skewed data. Specifically, we eschew the pattern of yielding prototypes via averaging intra-class embedding and directly learn a set of semantic anchors aided by the global semantic-aware classifier. Meanwhile, the margins between anchors are augmented via pulling apart them, ensuring decent inter-class separation. To guarantee that local domain-skewed representations can be uniformly distributed, local data is projected into the hyperspherical space, and the intra-class compactness is achieved by optimizing the contrastive loss derived from the von Mises-Fisher distribution. Finally, extensive experimental results on three multi-domain datasets show the superiority of the proposed FedLSA compared to existing typical and state-of-the-state methods.

Introduction

Artificial intelligence models rely on massive data for training to yield superior performance, which is often collected from many private devices. However, data privacy security increasingly renders data centralization impossible, severely hindering model evolution. Federated learning (Ye et al. 2023) is a promising way to break through data barrier dilemma, which collaboratively trains a global model among multiple participants with not revealing local data. It is worth emphasizing that data distributions among different clients are likely non-independent and identically distributed (Non-IID), harming the generalizability of global model.

*Corresponding author

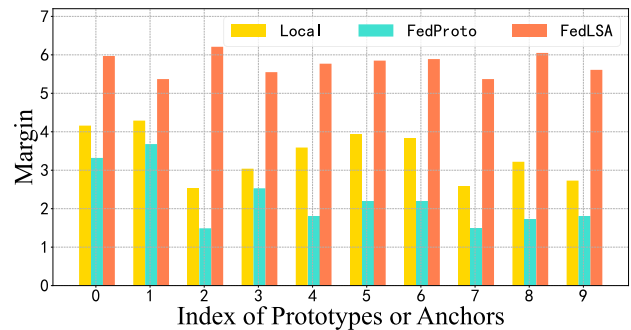


Figure 1: Margin comparison of prototypes/anchors in local clients, FedProto, and FedLSA on Office Caltech dataset. Margin is defined as the minimum L_2 -norm distance between a prototype/anchor and other prototypes/anchors. The maximum margin in all local clients is the baseline, which is compared with the margins of global prototypes in FedProto and anchors in FedLSA. It can be seen that the margins between global prototypes in FedProto are diminished while the margins between anchors in FedLSA are enhanced.

How to conquer the negative influence caused by heterogeneous distributions has become a widely discussed topic.

Most of existing works (Li et al. 2020; Li, He, and Song 2021; Zhang et al. 2024c) adopt contrastive learning, knowledge distillation, meta learning, etc. to strengthen the generalization of global model. Essentially, they expect to converge on the optimization directions of different local models via certain regularization. Notably, a fundamental assumption exists in above approaches that local data in diverse clients suffers from the label skew yet is sampled from the same domain. Conversely, in real-world scenarios, local data could be drawn from varying domains (Zhang et al. 2023a; Gong et al. 2022). To be specific, the statistics information of local data features is quite distinctive, which is more likely to result in large divergences in local model parameters, seriously jeopardizing the generalization of global model. Adversarial training is first introduced into federated learning with domain skew (Peng et al. 2020; Zhang et al. 2021, 2024b) due to its versatility in aligning various data distributions, but its difficulty in training appears as well. (Yan et al. 2024; Bai et al. 2024) aim to explore the represen-

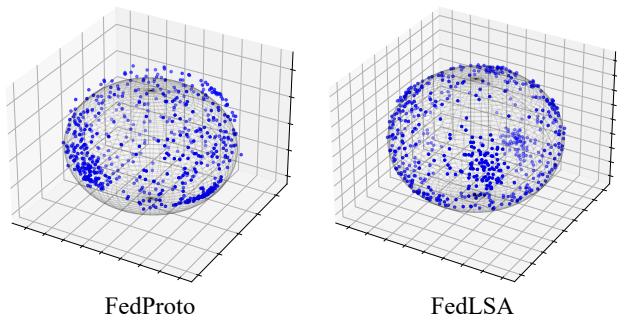


Figure 2: Hyperspherical embedding comparison of FedProto and FedLSA on USPS dataset. The embedding learned from USPS by FedProto and FedLSA is mapped into the hyperspherical space, then it can be observed that the distribution of points in FedProto is extremely nonuniform, even certain points are scattered outside the hypersphere, which is detrimental to implementing prototype-based contrastive learning. On the contrast, FedLSA learns a relatively uniform embedding space.

tations with minimal domain-specific information, ensuring unbiased training of local models. However, the representation disengagement process induces the high computational burden and precludes the practical applications.

Considering communication friendliness and decent generalization, federated prototype learning (Tan et al. 2022a; Huang et al. 2023; Dai et al. 2023) for domain-skewed data is developed. They average the representations in same classes to obtain local prototypes and upload them to the server for aggregating the global prototypes. In turn, the global prototypes are used to regularize local training. Although they effectively boost the generalizability of global model and achieve encouraging performance, two limitations remain inevitable. ❶ Local prototypes are yielded by averaging the embeddings in same categories, still stuck with domain-specific markers. Thus, the margins among aggregated global prototypes could be weakened, which is not favorable for achieving inter-class separation. Fig. 1 illustrates the margin comparison of prototypes in local clients and FedProto. It can be seen that the margins between global prototypes of FedProto are weakened compared to that of local prototypes. ❷ The data representations extracted by local feature extractors are scattered in Euclidean space and may not exhibit a uniform distribution, then prototype-based contrastive strategy cannot work in promoting intra-class compactness (Wang and Isola 2020; Liao et al. 2024). Fig. 2 visualizes the embedding in hyperspherical space of FedProto, it can be observed that the scatter learned by FedProto exhibits a remarkably nonuniform distribution, even some points jump out of the hypersphere, which poses a huge difficulty for contrastive learning. As discussed above, the two deficiencies gravely hamper training of local models and further decrease the generalization of global model.

In view of above issues, we propose a **F**ederated **L**earnable **S**emantic **A**nchors (FedLSA) method with Hyperspherical Contrast for Domain-Skewed Data. The pro-

posed FedLSA is beyond conventional federated prototype learning, which abandons the manner of averaging intra-class representations for producing local and global prototypes. Specifically, we train the semantic anchors with the help of global semantic-aware classifier on the server-side, eliminating the impact of local domain-specific information. The margins among anchors are further enhanced via pulling apart diverse anchors for sound inter-class separability. To guarantee uniform distribution of local representations, the local data is projected into the hyperspherical space. Under the guidance of semantic anchors, the intra-class representations are enforced to be compact. From Figs. 1 and 2, we can observe that the proposed FedLSA strengthens the margins of semantic anchors and promises the uniform distribution of local representations, which are more beneficial for inter-class separability and intra-class compactness. Fig. 3 shows the framework of the proposed FedLSA. The main contributions are concluded as follows.

- We learn the semantic anchors via global semantic-aware classifier beyond conventional federated prototype learning, averting the impact of local domain-specific information. Various anchors are pulled away for larger margins, thus better promoting inter-class separability.
- We project the local data into hyperspherical space, where the latent embedding can be uniformly distributed. Under the penalty of contrastive loss induced by von Mises-Fisher distribution, the intra-class representations are more likely approach their corresponding anchors, thus enabling intra-class compactness.
- To verify the effectiveness of the proposed FedLSA, a multitude of experiments on three multi-domain datasets are conducted. Compared with the existing federated learning methods, FedLSA effectively improves the generalizability and performance of global model.

Related Works

Federated Learning for Heterogeneous Data

To protect the local data privacy, federated learning has been proposed as an encouraging distributed model training paradigm. Heterogeneous data across multiple clients arises along with federated learning, and becomes one of the realistic challenges for federated learning applications. When regarding heterogeneous data, there are two kinds of categories: label skew and domain skew. For addressing label skew, massive researches have been evolved. (Li et al. 2020; Li, He, and Song 2021; Huang et al. 2024) introduced local regularity term to force the optimization directions of local models closer to the global direction. To enhance knowledge integration under data heterogeneity, (Yang et al. 2023; Chen et al. 2024; Yao et al. 2024) adopted the knowledge distillation technologies to transfer knowledge from other clients to local models. Learning only one global model may not effectively handle heterogeneous data in multiple clients, some works (Yang, Huang, and Ye 2024; Zhi et al. 2024; Tu et al. 2024) proposed personalized strategies, using global information to foster local personalized modules. Domain-skewed data readily weakens the generalization of

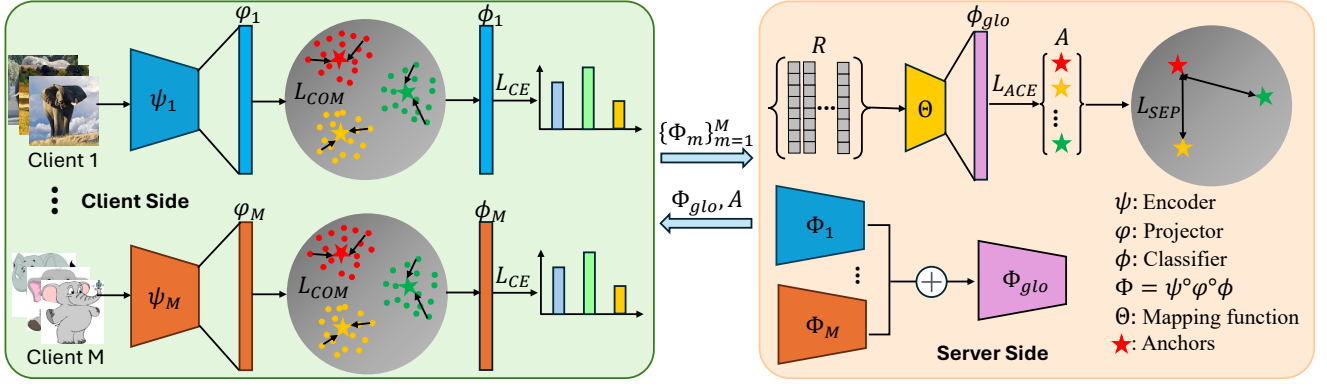


Figure 3: The framework of the proposed FedLSA. On the client side, each client projects the local data into hyperspherical space and performs the contrastive learning under the guidance of semantic anchors, thus achieving the intra-class compactness. On the server side, the semantic anchors are learned with the help of global semantic-aware classifier, whose margins are enhanced via separation loss, then the inter-class separability can be guaranteed.

the global model. (Hong et al. 2021; Huang et al. 2022; Jiang et al. 2024) leveraged the adversarial training to bridge the gap between multiple domains. (Peng et al. 2020; Yan et al. 2024; Bai et al. 2024) disentangled the feature embedding into domain-shared and domain-specific representations. In general, the label skew oriented methods cannot effectively handle domain-skewed data, while the domain skew oriented methods frequently require additional discriminators or domain-invariant feature extractors, increasing the training difficulty and communication burden.

Prototype Learning

The prototype representation is an average of a class of samples, and are generally used to bring the representations in a same categories closer together, thus enabling intra-class compactness. In light of high reliability and easy manipulation, prototype-induced regularization is widely investigated. (Li et al. 2021; Wei et al. 2023; Lu et al. 2024) proposed prototype-based contrastive loss, pulling prototypes closer to their kindred samples and pulling away from non-kindred samples. To eliminate the impact of superfluous information, (Zhou et al. 2022; Li et al. 2023; Fang et al. 2024) learned the adaptive prototypes that are independent of the data distributions. Due to the properties of high expression and low parameters, the idea of prototypes has been introduced into federated learning for local regularization and efficient communication. (Tan et al. 2022a,b; Dai et al. 2023) exchanged the prototypes between the server and clients, thus significantly decreasing the communication cost. Zhang et al. (Zhang et al. 2024a) trained a set of prototypes with larger separation on the server. Huang et al. (Huang et al. 2023) obtained the clustered prototypes to capture the diversity within a class. When faced with the domain-skewed data, none of the above federated prototype learning address the thorny question, namely, how to obtain domain-independent prototype. Fortunately, the proposed FedLSA achieves the exploration of semantic anchors with the help of global semantic-aware classifier, removing the domain-specific information.

Representation Learning on Hyperspherical Space

Hyperspherical space is a specific form of von Mises-Fisher distribution (Banerjee et al. 2005). The feature representation lying in a hyperspherical space has good homogeneity and avoids the common feature collapse problem in neural network training. In light of above merits, learning representations on a hyperspherical space becomes a reasonable choice. Liu et al. (Liu et al. 2017) proposed the operation of hyperspherical convolution to overcome the dilemma caused by large parameters. Mettes et al. (Mettes, van der Pol, and Snoek 2019) proposed hyperspherical prototype network, tackling the representation learning for multiple tasks. Saadabadi et al. (Saadabadi et al. 2024) introduced a dynamic mechanism of label-to-prototype assignment for hyperspherical classification. Zhang et al. (Zhang et al. 2024d) provided a hyperspherical margin weighting manner, constructing the importance for each sample. Wang et al. (Wang et al. 2024) devised a hyperspherical loss to achieve inter-class separation and intra-class cohesion. Current federated prototype methods have not yet noticed the role of hyperspherical space for enhancing representation learning. Fortunately, in the proposed FedLSA, we project local data into the hypersphere space and utilize learnable anchors to enable intra-class compactness and inter-class separation.

Methodology

Preliminaries

Given a federated learning system with a centralized server and M clients, each client trains its local model $\Phi_m = \psi_m \circ \varphi_m \circ \phi_m$ on private data $D_m = \{x_i, y_i\}_{i=1}^{|D_m|}$. ψ_m is the encoder that projects the raw data into embedding space: $z_i = \psi_m(x_i) \in \mathbb{R}^I$, where I is the dimension of embedding. φ_m is a the projector that further projects the high-dimensional embedding into a low-dimensional hyperspherical space: $h_i = \text{nor}(\varphi(z_i)) \in \mathbb{R}^L$, where L is the dimension of hyperspherical embedding and $\text{nor}(\cdot)$ is defined as $\text{nor}(x) = x/\|x\|_2$. ϕ_m is the classifier that maps the embedding into logits output: $q_i = \phi_m(h_i) \in \mathbb{R}^C$, where C is the

number of categories. Domain-skewed data across different clients is mathematically defined as $P_m(x|y) \neq P_n(x|y)$, where $P(\cdot)$ denotes the probability distribution. Domain-skewed data essentially refers to sharing of label space among different clients, while with varying distributions of feature spaces.

Federated Prototype Learning

Federated prototype learning has demonstrated its flourishing vitality due to the efficient communication and client alignment capability. For the c -th prototype p_m^c of the m -th client, it is defined as the average of embedding belonging to a category and expressed as

$$p_m^c = \frac{1}{|D_m^c|} \sum_{i=1}^{|D_m^c|} \psi_m(x_i), \quad (1)$$

$$P_m = [p_m^1; \dots; p_m^C] \in \mathbb{R}^{C \times I},$$

where $|D_m^c|$ denotes the number of samples in the c -th class for the m -th client. After the clients complete the local training, the local prototypes are uploaded to the server for global aggregation, which is formulated as

$$p^c = \frac{1}{M} \sum_{m=1}^M p_m^c, \quad (2)$$

$$\mathcal{P} = [p^1; \dots; p^C] \in \mathbb{R}^{C \times I}.$$

The global prototypes \mathcal{P} are dispatched to regularize local models such as contrastive loss. To obtain a global model with well generalization, some works (Nguyen, Torr, and Lim 2022; Huang et al. 2023) also demand to transmit the local models for federated average. In this paper, we follow the above works and aim to learn a generalizable global model on various domains.

As described above, local prototypes in traditional federated prototypes learning are averaged from intra-class representations, which are necessarily biased by domain-specific information and likely to undermine the margins between aggregated global prototypes. In addition, local data is projected into Euclidean embedding space via local feature extractors, which cannot enable a uniform distribution and is not favorable for the prototype-based contrastive learning.

Learnable Semantic Anchors with Hyperspherical Contrast

In response to above issues, we propose to directly learn the semantic anchors for domain-skewed data beyond traditional federated prototype learning. The proposed FedLSA is composed of two prominent modules: Sematic Anchors Learning with Inter-Class Separability and Hyperspherical Representation Learning with Intra-Class Compactness.

Sematic Anchors Learning with Inter-Class Separability Marginal attenuation of global prototypes caused by domain-skewed data is prone to fuzzy division of various classes of representations, which is not conducive to positive model training. Herein, we propose the Sematic Anchors Learning with Inter-Class Separability to address the

above concern. Concretely, a set of learnable anchors $A = [a_1; \dots; a_C]$, initially parameterized by $R = [r_1; \dots; r_C] \in \mathbb{R}^{C \times I}$, are stored on the server. A mapping function $\Theta(\cdot)$ is used to map the random vectors R to the anchors: $A = \Theta(R)$. Herein, we adopt two-layer Multi-Layer Perceptron (MLP) as the mapping function.

In light of the drawback of global prototypes produced by averaging local prototypes, we argue that good anchors should satisfy two fundamental properties. 1) They are only relevant to the class semantics and independent from the specific domains. 2) Different anchors should be far away from each other for decent separability. To achieve the two goals, we perform the following training strategies. First, when the local models are fused to a global model on the server, the global classifier $\phi_{glo}(\cdot)$ incorporates information from multiple domains and offers the generalized semantic discriminative capabilities. Hence, we adopt $\phi_{glo}(\cdot)$ to project learnable anchors $A = [a_1; \dots; a_C]$ into semantic space and distinguish the learnable anchors of different categories by the typical CrossEntropy (CE) loss:

$$\mathcal{L}_{ACE} = -\mathbf{1}_{y_i} \log(\text{softmax}(\rho_i)), \quad (3)$$

where $\mathbf{1}_{y_i}$ is the one-hot label of the i -th anchor, $\rho_i = \phi_{glo}(a_i)$ is the predicted logits output for the i -th learnable anchor a_i .

By the optimization of \mathcal{L}_{ACE} , different anchors correspond to diverse semantic information, achieving initial differentiation. Nevertheless, good anchors, as mentioned above, need to be further separated from each other. Then, the local representations can be guided to enable better inter-class separability. To fulfill this conception, we conduct the following separability loss, which is written as

$$\mathcal{L}_{SEP} = \log \frac{\sum_{j=1, j \neq i}^C \exp(a_i a_j^T) / \tau}{C - 1}, \quad (4)$$

where τ denotes the temperature parameter, controlling the concentration strength of representations. Thus, the overall loss for learning well-separated anchors is formulated as

$$\mathcal{L}_{LSA} = \mathcal{L}_{ACE} + \alpha \mathcal{L}_{SEP}, \quad (5)$$

where α denotes the trade-off parameter. Overall, we directly learn the semantic anchors by means of network training, which is very different from conventional federated prototype learning. The proposed approach not only ensures that anchors are protected from domain-specific information but also ensures sufficient margins to each other.

Hyperspherical Representation Learning with Intra-Class Compactness When performing the local training, the m -th client encodes the raw data into embedding space: $z_i = \psi_m(x_i)$. However, after the simple projection of a feature extractor, the embedding space may not exhibit a uniform distribution, where the sample points are partially dense and partially sparse. Extensive works (Wang and Isola 2020; Ming et al. 2023) have demonstrated that the embeddings scattered in a hypersphere space exhibit the excellent uniformity, which is important to span a well-separated manifold. Hence, the embedding z_i is further projected into the hyperspherical space: $h_i = \text{nor}(\varphi_m(z_i))$. A hyperspherical space

can be modeled as the von Mises-Fisher distribution (Wang and Isola 2020; Ming et al. 2023), which is formulated as a probability density function p_D :

$$p_D(h; a_c, \kappa) = N_D(\kappa) \exp(\kappa a_c^T h), \quad (6)$$

where h is the hyperspherical embedding, a_c is the mean and can also be viewed as the c -th prototype, κ is the concentration parameter that controls the tightness around the mean, $N_D(\kappa)$ denotes the normalization factor. According to the probability model Eq. (6), the normalized probability that the i -th hyperspherical embedding h_i is allocated to the c -th category is formulated as

$$p(y_i = c | h_i; \{\kappa, a_j\}_{j=1}^C) = \frac{\exp(a_c^T h_i / \tau)}{\sum_{j=1}^C \exp(a_j^T h_i / \tau)}, \quad (7)$$

where $\tau = \frac{1}{\kappa}$ is the temperature parameter. For an embedding space with sound decision boundaries, each data point should fall into corresponding class with the highest probability compared to other classes. To achieve the goal, the maximum likelihood estimation (MLE) is implemented over the Eq. (7): $\max_{\psi_m, \varphi_m} \prod_{i=1}^{|D_m|} p(y_i = c | h_i; \{\kappa, a_j\}_{j=1}^C)$. For obtaining an optimizable loss, we adopt the negative log-likelihood strategy to transfer the above MLE problem into following form:

$$\mathcal{L}_{COM} = -\log \frac{\exp(a_{y_i}^T h_i / \tau)}{\sum_{j=1}^C \exp(a_j^T h_i / \tau)}, \quad (8)$$

where a_{y_i} denotes the anchor corresponding to the i -th hyperspherical embedding h_i . The loss \mathcal{L}_{COM} brings each hyperspherical embedding closer towards its respective anchor, facilitating the intra-class compactness. Moreover, the CE loss is used to endow the local model with basic classification ability:

$$\mathcal{L}_{CE} = -\mathbf{1}_{y_i} \log(\text{softmax}(q_i)), \quad (9)$$

where $q_i = \phi_m(h_i)$ denotes the predicted logits output for the i -th hyperspherical embedding h_i . Thus, the overall loss for learning well-compact hyperspherical embedding can be written as

$$\mathcal{L}_{HC} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{COM}, \quad (10)$$

where λ is the trade-off parameter. In each communication round, the semantic anchors are trained on the server and distributed to clients. Then the participants start the local training with the guidance of semantic anchors in parallel. Algorithm 1 reports the main flow.

Experiments

Datasets

We implement the comparative experiments on three multi-domain datasets: **Digits**, **Office Caltech**, and **PACS**. **Digits** contains 203,587 digital images with four domains, such as MNIST, USPS, SVHN, and SYN. Number categories from 0 to 9 are included in each domain. **Office Caltech** is consisted of 2,533 object images with 10 classes, four kinds domains are included: CALTECH, AMAZON, WEBCAM, and DSLR. **PACS** contains one type of real-world images

Algorithm 1: The flow of FedLSA

Input: Number of clients M , training epochs on clients and server E_c, E_s , communication rounds T , learning rate η , temperature parameter τ , local data $D_m = \{x_i, y_i\}_{i=1}^{|D_m|}$, anchors' initialization parameter R .

Output: Global model Φ_{glo} .

- 1: **Client Side:**
- 2: **for** $m = 1 : M$ **in parallel do**
- 3: **for** epoch $e = 1 : E_c$ **do**
- 4: $h_i = \text{nor}(\varphi_m(\psi_m(x_i)))$;
- 5: // Calculate the CE loss \mathcal{L}_{CE} //
- 6: $\mathcal{L}_{CE} \leftarrow (\text{softmax}(\phi_m(h_i)), y_i)$ using Eq. (9);
- 7: // Calculate the compactness loss \mathcal{L}_{COM} //
- 8: $\mathcal{L}_{COM} \leftarrow (\{a_j\}_{j=1}^C, h_i)$ using Eq. (8);
- 9: $\mathcal{L}_{HC} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{COM}$;
- 10: $\Phi_m^e \leftarrow \Phi_m^{e-1} - \eta \nabla \mathcal{L}_{HC}$;
- 11: **end for**
- 12: **end for**
- 13: Upload the local model Φ_m to the server;
- 14: **Server Side:**
- 15: **for** $t = 1 : T$ **do**
- 16: **for** $e = 1 : E_s$ **do**
- 17: $A = \Theta(R)$;
- 18: // Calculate the CE loss \mathcal{L}_{ACE} //
- 19: $\mathcal{L}_{ACE} \leftarrow (\text{softmax}(\phi_{glo}(a^i)), y_i)$ using Eq. (3);
- 20: // Calculate the separation loss \mathcal{L}_{SEP} //
- 21: $\mathcal{L}_{SEP} \leftarrow (\{a_j\}_{j=1}^C)$ using Eq. (4);
- 22: $\mathcal{L}_{LSA} = \mathcal{L}_{ACE} + \alpha \mathcal{L}_{SEP}$;
- 23: $R^e \leftarrow R^{e-1} - \eta \nabla \mathcal{L}_{LSA}$;
- 24: $\Theta^e \leftarrow \Theta^{e-1} - \eta \nabla \mathcal{L}_{LSA}$;
- 25: **end for**
- 26: Broadcast the semantic anchors A and global model Φ_{glo} to clients;
- 27: **end for**

and three types of art images, a total of 9,991 images with 7 categories are covered. The federated learning system is equipped with 10 clients and 1 server. To simulate that different clients hold data from different domains, we split the data of a domain to certain clients. Taking Digits as an example, the specific division is MNIST: 1, USPS: 4, SVHN: 2, SYN: 3. Table 1 summarizes the prominent characteristic of the datasets. More details about the datasets and data division can be referred in supplementary material.

Dataset	Instances	Domains	Classes	Feature Dimensions
Digits	203,587	4	10	$32 \times 32 \times 3$
Office Caltech	2,533	4	10	$32 \times 32 \times 3$
PACS	9,991	4	7	$225 \times 225 \times 3$

Table 1: Statistics of three multi-domain datasets.

Compared Methods

The proposed FedLSA is compared with eight federated learning methods, including both typical and state-of-the-

Methods	Digits						Office Caltech					
	MNIST	USPS	SVHN	SYN	AVERAGE	Δ	CALTECH	AMAZON	WEBCAM	DSLRL	AVERAGE	Δ
FedAvg	79.11	79.07	63.20	22.19	60.89	-	56.70	66.32	43.10	46.67	53.20	-
FedProx	80.91	78.97	61.32	25.37	61.64	+0.75	55.80	65.79	50.00	46.67	54.57	+1.37
MOON	83.18	79.42	61.25	23.78	61.91	+1.02	60.27	68.95	51.72	50.00	57.74	+4.54
FedSR	75.41	79.42	62.49	24.08	60.35	-0.54	58.93	64.74	44.83	53.33	55.46	+2.26
FedProto	79.87	78.13	60.83	25.97	61.20	+0.31	54.02	62.63	41.38	50.00	52.01	-1.19
FedGA	80.17	81.22	49.27	31.69	60.59	-0.30	53.12	57.89	62.07	66.67	<u>59.94</u>	<u>+6.74</u>
FPL	79.98	79.47	62.68	26.22	<u>62.09</u>	<u>+1.20</u>	58.93	68.42	53.45	53.33	58.53	+5.33
FedTGP	80.20	78.67	61.22	26.97	61.77	+0.88	56.25	65.79	50.00	53.33	56.34	+3.14
FedLSA	81.50	76.63	65.17	30.95	63.56	+2.67	55.80	64.74	60.34	60.00	60.22	+7.02

Table 2: Performance comparison (%) of all compared methods on Digits and Office Caltech datasets, where *AVERAGE* and Δ denote the average performance of all domains and the improvements over FedAvg, respectively.

Methods	PACS					
	PHOTO	ART	CARTOON	SKETCH	AVERAGE	Δ
FedAvg	47.50	31.22	47.87	34.86	40.36	-
FedProx	50.30	31.87	48.30	37.32	41.95	+1.59
MOON	49.10	32.52	48.01	36.81	41.61	+1.25
FedSR	49.10	30.89	51.42	36.30	41.93	+1.57
FedProto	50.70	32.85	50.28	34.44	42.07	+1.71
FedGA	51.50	31.87	52.13	34.52	<u>42.51</u>	<u>+2.15</u>
FPL	48.50	33.33	47.16	32.91	40.48	+0.12
FedTGP	49.90	34.80	50.85	34.27	42.46	+2.10
FedLSA	52.30	35.77	53.41	33.42	43.73	+3.37

Table 3: Performance comparison (%) of all compared methods on PACS dataset, where *AVERAGE* and Δ denote the average performance on all domains and the improvements over FedAvg, respectively.

art. **FedAvg** (McMahan et al. 2017) and **FedProx** (Li et al. 2020) are two widely used federated learning algorithms. **MOON** (Li, He, and Song 2021), **FedSR** (Nguyen, Torr, and Lim 2022), **FedGA** (Zhang et al. 2023b), **FedProto** (Tan et al. 2022a), **FPL** (Huang et al. 2023), and **FedTGP** (Zhang et al. 2024a) are six state-of-the-art federated learning approaches. Concretely, **MOON** (Li, He, and Song 2021) employs the global presentations as the contrast signal to combat catastrophic forgetting. **FedSR** (Nguyen, Torr, and Lim 2022) uses the mutual information mechanism to mine domain-independent information of each client. **FedGA** (Zhang et al. 2023b) seeks fair treatment of various clients when encountering heterogenous data. **FedProto** (Tan et al. 2022a) proposes a prototype based federated learning schema. **FPL** (Huang et al. 2023) utilizes a group of prototypes rather than individual prototype to promote the contrastive learning. **FedTGP** (Zhang et al. 2024a) trains the global prototypes based on local prototypes on the server.

Implementation Details

We perform the experiments with a simple convolutional neural network, which is equipped with 2-layer convolution layers and 3-layer linear layers. The specific parameters of

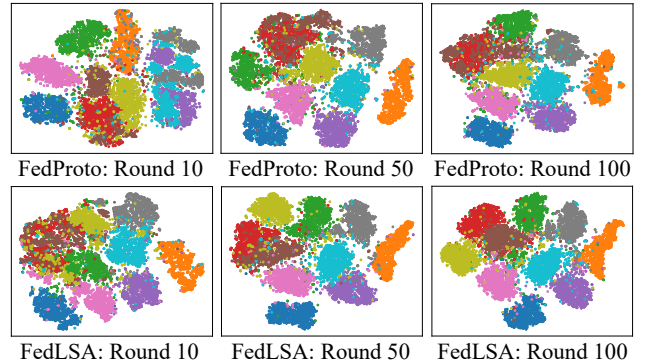


Figure 4: Scatter visualization of FedProto and FedLSA on MNIST, where the t-SNE is used for dimension reduction.

model architecture is introduced in the supplementary material. Stochastic Gradient Descent (SGD) is employed as optimizer. The communication round is set as 100, the epoch numbers on the clients and server are set as 5 and 500, respectively. The batch size is fixed as 64. In the proposed FedLSA, three hyperparameters α , λ , τ are tuned in $\{0.02, 0.4\}$, $\{0.1, 0.7\}$, $\{0.1, 0.2\}$ respectively. All comparative experiments are implemented on a server with Intel(R) Xeon Gold 6230R CPU, RTX 4090GPU, and 128G RAM.

Performance Comparison

Tables 2 and 3 record the results on three multi-domain datasets. On the one hand, in terms of average performance on all domains, the proposed FedLSA achieves the best results on three test datasets, demonstrating the effectiveness of FedLSA. Particularly, the performance of FedLSA is further upgraded compared to prototype-based approaches, such as FedProto, FPL, and FedTGP. It is worth emphasizing that on Office Caltech dataset, FedProto results in performance degradation compared to FedAvg, because the domain-skewed data induces the margin decay of global prototypes, which severely affects the classification effect. While FedLSA achieves an increase of 7.02%, so learning the anchors with enhanced margins that are only relevant to the semantic is necessary. On the other hand, it is normal

\mathcal{L}_{SEP}	\mathcal{L}_{COM}	Digits				
		MNIST	USPS	SVHN	SYN	AVERAGE
✗	✗	79.11	79.07	63.20	22.19	60.89
✗	✓	78.71	74.54	62.46	28.51	61.06
✓	✗	81.57	75.64	61.01	26.62	61.21
✓	✓	81.50	76.63	65.17	30.95	63.56

Table 4: Ablation results (%) with respect to two principal losses on Digits dataset.

\mathcal{L}_{SEP}	\mathcal{L}_{COM}	Office Caltech				
		CALTECH	AMAZON	WEBCAM	DSLR	AVERAGE
✗	✗	56.70	66.32	43.10	46.67	53.20
✗	✓	54.91	61.58	56.90	63.33	59.18
✓	✗	55.80	64.21	53.45	60.00	58.37
✓	✓	55.80	64.74	60.34	60.00	60.22

Table 5: Ablation results (%) with respect to two principal losses on Office Caltech dataset.

that the proposed FedLSA does not work best all domains. For well-generalized anchors, it should be adapt to various domains as much as possible and achieve balanced results among them. Therefore, realizing optimal average performance is the advantage of the proposed FedLSA. Moreover, the scatter of FedProto and FedLSA on MNIST is visualized with the communication round increasing in Fig. 4. It can be seen that FedProto is consistently inferior than the proposed FedLSA in terms of intra-class compactness, which is attributed to the fact that the semantic anchors learned by FedLSA are more applicable to cross-domain data, and the hyperspherical space projection is more conducive to prototype contrastive learning.

Ablation Study

Separability loss \mathcal{L}_{SEP} and compactness loss \mathcal{L}_{COM} support FedLSA for excellent performance, their significances are validated by a series of ablation experiments in Tables 4 and 5. The CE loss \mathcal{L}_{ACE} is a basic loss like \mathcal{L}_{CE} that enables semantic differentiation of anchors, so it's not necessary to be ablated. When neither \mathcal{L}_{SEP} nor \mathcal{L}_{COM} exists, FedLSA degenerates to FedAvg, realizing the baseline performance. Further, the performance is improved as either \mathcal{L}_{SEP} or \mathcal{L}_{COM} is included. Notably, when \mathcal{L}_{SEP} is included and \mathcal{L}_{COM} is dropped, we use L_2 -norm to constrain intra-class representations to close to their anchors, otherwise only using \mathcal{L}_{SEP} is ineffective. Overall, the complete model achieves the best performance thanks to the inter-class separability and intra-class compactness induced by \mathcal{L}_{SEP} and \mathcal{L}_{COM} .

Hyperparameter Study

Trade-off parameters α and λ play the role of balancing different losses, we perform the proposed FedLSA with varying parameter settings to investigate their specific effects. Fig. 5 shows the results when α and λ are set in $\{0.001, \dots, 1\}$ on Digits and Office Caltech. We can observe

that FedLSA exhibits little sensitivity to changes of α and λ on Digits, which may be attributed by the data simplicity that reduces the training difficulty. For Office Caltech dataset, a smaller λ only allows FedLSA to achieve moderate performance, since a smaller λ cannot favorably promote intra-class compactness. The temperature parameter τ is also an important factor for FedLSA, Fig. 6 presents the performance under various values of τ in $\{0.05, \dots, 1\}$, neither too high nor too low τ is conducive to satisfactory results. Too high τ detracts from the strength of the contrastive learning, while too low τ tends to reduce representation discriminability. Hence, a suitable value of τ is important.

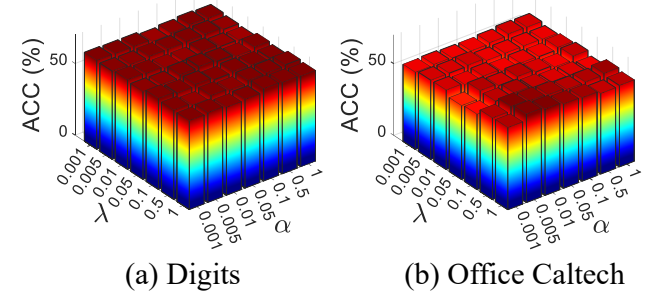


Figure 5: Hyperparameter study with respect to α and λ on Digits and Office Caltech, where α and λ are varied in $\{0.001, \dots, 1\}$, respectively.

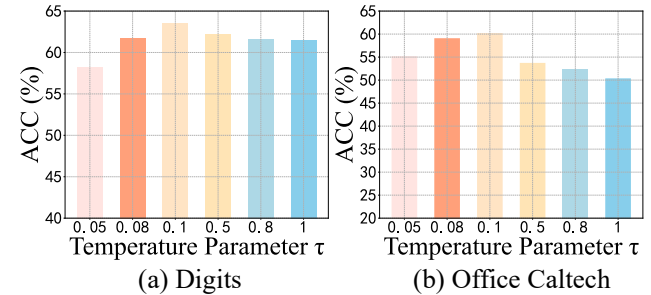


Figure 6: Sensitivity investigation with respect to temperature parameter τ on Digits and Office Caltech, where τ is tuned in $\{0.05, \dots, 1\}$.

Conclusion

In this paper, we propose a novel federated learning method for domain-skewed data. The proposed FedLSA directly learns the semantic anchors with enhanced margins via the global semantic-aware classifier, avoiding the adverse impact of domain-specific information. Furthermore, local data is mapped into a hyperspherical space with uniform distribution, then guided by semantic anchors, better separability in representation space for each client can be realized. To verify the effectiveness of the proposed FedLSA, substantial comparative and ablation experiments are conducted, and the experimental results prove the advance of the proposed FedLSA compared to the typical and state-of-the-art federated learning approaches.

Acknowledgments

The research is supported by the National Key Research and Development Program of China (2023YFB2703700), the National Natural Science Foundation of China (62176269), the Guangzhou Science and Technology Program (2023A04J0314).

References

- Bai, S.; Zhang, J.; Guo, S.; Li, S.; Guo, J.; Hou, J.; Han, T.; and Lu, X. 2024. DiPrompt: Disentangled Prompt Tuning for Multiple Latent Domain Generalization in Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27284–27293.
- Banerjee, A.; Dhillon, I. S.; Ghosh, J.; Sra, S.; and Ridge-way, G. 2005. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(9).
- Chen, H.; Zhang, Y.; Krompass, D.; Gu, J.; and Tresp, V. 2024. Feddat: An approach for foundation model finetuning in multi-modal heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 11285–11293.
- Dai, Y.; Chen, Z.; Li, J.; Heinecke, S.; Sun, L.; and Xu, R. 2023. Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7314–7322.
- Fang, Y.; Chen, C.; Zhang, W.; Wu, J.; Zhang, Z.; and Xie, S. 2024. Prototype learning for adversarial domain adaptation. *Pattern Recognition*, 155: 110653.
- Gong, X.; Sharma, A.; Karanam, S.; Wu, Z.; Chen, T.; Dormann, D.; and Innanje, A. 2022. Preserving privacy in federated learning with ensemble cross-domain knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 11891–11899.
- Hong, J.; Zhu, Z.; Yu, S.; Wang, Z.; Dodge, H. H.; and Zhou, J. 2021. Federated Adversarial Debiasing for Fair and Transferable Representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 617–627.
- Huang, S.; Fu, L.; Li, Y.; Chen, C.; Zheng, Z.; and Dai, H.-N. 2024. A Cross-Client Coordinator in Federated Learning Framework for Conquering Heterogeneity. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. Doi=10.1109/TNNLS.2024.3439878.
- Huang, W.; Ye, M.; Du, B.; and Gao, X. 2022. Few-shot model agnostic federated learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, 7309–7316.
- Huang, W.; Ye, M.; Shi, Z.; Li, H.; and Du, B. 2023. Rethinking federated learning with domain shift: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16312–16322.
- Jiang, L.; Wang, X.; Yang, X.; Shu, J.; Lin, H.; and Yi, X. 2024. FedPA: Generator-Based Heterogeneous Federated Prototype Adversarial Learning. *IEEE Transactions on Dependable and Secure Computing*. Doi=10.1109/TDSC.2024.3419211.
- Li, G.; Jampani, V.; Sevilla-Lara, L.; Sun, D.; Kim, J.; and Kim, J. 2021. Adaptive prototype learning and allocation for few-shot segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8334–8343.
- Li, H.; Song, J.; Gao, L.; Zhu, X.; and Shen, H. 2023. Prototype-based Aleatoric Uncertainty Quantification for Cross-modal Retrieval. In *Proceedings of the Advances in Neural Information Processing Systems*, 24564–24585.
- Li, Q.; He, B.; and Song, D. 2021. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10713–10722.
- Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, 429–450.
- Liao, X.; Liu, W.; Chen, C.; Zhou, P.; Yu, F.; Zhu, H.; Yao, B.; Wang, T.; Zheng, X.; and Tan, Y. 2024. Rethinking the Representation in Federated Unsupervised Learning with Non-IID Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22841–22850.
- Liu, W.; Zhang, Y.; Li, X.; Liu, Z.; Dai, B.; Zhao, T.; and Song, L. 2017. Deep Hyperspherical Learning. In *Proceedings of the Advances in Neural Information Processing Systems*, 3950–3960.
- Lu, H.; Gong, D.; Wang, S.; Xue, J.; Yao, L.; and Moore, K. 2024. Learning with mixture of prototypes for out-of-distribution detection. In *Proceedings of the International Conference on Learning Representations*.
- McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 1273–1282.
- Mettes, P.; van der Pol, E.; and Snoek, C. 2019. Hyperspherical Prototype Networks. In *Proceedings of the Advances in Neural Information Processing Systems*, 1485–1495.
- Ming, Y.; Sun, Y.; Dia, O.; and Li, Y. 2023. How to exploit hyperspherical embeddings for out-of-distribution detection? In *Proceedings of the International Conference on Learning Representations*.
- Nguyen, A. T.; Torr, P.; and Lim, S. N. 2022. Fedsr: A simple and effective domain generalization method for federated learning. *Advances in Neural Information Processing Systems*, 35: 38831–38843.
- Peng, X.; Huang, Z.; Zhu, Y.; and Saenko, K. 2020. Federated Adversarial Domain Adaptation. In *Proceedings of the International Conference on Learning Representations*.
- Saadabadi, M. S. E.; Dabouei, A.; Malakshan, S. R.; and Nasrabadi, N. M. 2024. Hyperspherical Classification with Dynamic Label-to-Prototype Assignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17333–17342.

- Tan, Y.; Long, G.; Liu, L.; Zhou, T.; Lu, Q.; Jiang, J.; and Zhang, C. 2022a. Fedproto: Federated prototype learning across heterogeneous clients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8432–8440.
- Tan, Y.; Long, G.; Ma, J.; LIU, L.; Zhou, T.; and Jiang, J. 2022b. Federated Learning from Pre-Trained Models: A Contrastive Learning Approach. In *Advances in Neural Information Processing Systems*, volume 35, 19332–19344.
- Tu, J.; Huang, J.; Yang, L.; and Lin, W. 2024. Personalized Federated Learning with Layer-Wise Feature Transformation via Meta-Learning. *ACM Transactions on Knowledge Discovery from Data*, 18(4): 1–21.
- Wang, H.; Cao, J.; Shi, Z.-L.; Leung, C.-S.; Feng, R.; Cao, W.; and He, Y. 2024. Image Classification on Hypersphere Loss. *IEEE Transactions on Industrial Informatics*, 20(4): 6531–6541.
- Wang, T.; and Isola, P. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the International Conference on Machine Learning*, 9929–9939.
- Wei, Y.; Ye, J.; Huang, Z.; Zhang, J.; and Shan, H. 2023. Online Prototype Learning for Online Continual Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 18764–18774.
- Yan, Y.; Wang, H.; Huang, Y.; He, N.; Zhu, L.; Xu, Y.; Li, Y.; and Zheng, Y. 2024. Cross-Modal Vertical Federated Learning for MRI Reconstruction. *IEEE Journal of Biomedical and Health Informatics*, 1–13. Doi=10.1109/JBHI.2024.3360720.
- Yang, X.; Huang, W.; and Ye, M. 2024. FedAS: Bridging Inconsistency in Personalized Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11986–11995.
- Yang, Z.; Zhang, Y.; Zheng, Y.; Tian, X.; Peng, H.; Liu, T.; and Han, B. 2023. FedFed: Feature Distillation against Data Heterogeneity in Federated Learning. In *Advances in Neural Information Processing Systems*, 60397–60428.
- Yao, D.; Pan, W.; Dai, Y.; Wan, Y.; Ding, X.; Yu, C.; Jin, H.; Xu, Z.; and Sun, L. 2024. FedGKD: Toward Heterogeneous Federated Learning via Global Knowledge Distillation. *IEEE Transactions on Computers*, 73(1): 3–17.
- Ye, M.; Fang, X.; Du, B.; Yuen, P. C.; and Tao, D. 2023. Heterogeneous federated learning: State-of-the-art and research challenges. *ACM Computing Surveys*, 56(3): 1–44.
- Zhang, J.; Liu, Y.; Hua, Y.; and Cao, J. 2024a. Fedtgp: Trainable global prototypes with adaptive-margin-enhanced contrastive learning for data and model heterogeneity in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16768–16776.
- Zhang, J.; Zhao, L.; Yu, K.; Min, G.; Al-Dubai, A. Y.; and Zomaya, A. Y. 2024b. A Novel Federated Learning Scheme for Generative Adversarial Networks. *IEEE Transactions on Mobile Computing*, 23(5): 3633–3649.
- Zhang, L.; Fu, L.; Liu, C.; Yang, Z.; Yang, J.; Zheng, Z.; and Chen, C. 2024c. Toward Few-Label Vertical Federated Learning. 18(7).
- Zhang, L.; Fu, L.; Wang, T.; Chen, C.; and Zhang, C. 2023a. Mutual information-driven multi-view clustering. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 3268–3277.
- Zhang, L.; Lei, X.; Shi, Y.; Huang, H.; and Chen, C. 2021. Federated learning with domain generalization. *arXiv preprint arXiv:2111.10487*.
- Zhang, R.; Xu, Q.; Yao, J.; Zhang, Y.; Tian, Q.; and Wang, Y. 2023b. Federated domain generalization with generalization adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3954–3963.
- Zhang, S.; Li, Y.; Wang, Z.; Li, J.; and Liu, C. 2024d. Learning with Noisy Labels Using Hyperspherical Margin Weighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 16848–16856.
- Zhi, M.; Bi, Y.; Xu, W.; Wang, H.; and Xiang, T. 2024. Knowledge-Aware Parameter Coaching for Personalized Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 17069–17077.
- Zhou, T.; Wang, W.; Konukoglu, E.; and Van Gool, L. 2022. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2582–2593.