**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

# FedSaaS: Class-Consistency Federated Semantic Segmentation via Global Prototype Supervision and Local Adversarial Harmonization

**Xiaoyang Yu**[1,2], **Xiaoming Wu**[1,2*], **Xin Wang**[1,2*], **Dongrun Li**[1,2], **Ming Yang**[1,2], **Peng Cheng**[3]
[1]Key Lab. of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center, Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
[2]Shandong Provincial Key Lab. of Industrial Network and Information System Security, Shandong Fundamental Research Center for Computer Science, Jinan, China
[3]College of Control Science and Engineering, Zhejiang University, Hangzhou, China
b1043124010@stu.qlu.edu.cn, wuxm@sdas.org, xinwang@qlu.edu.cn

## Abstract

Federated semantic segmentation enables pixel-level classification in images through collaborative learning while maintaining data privacy. However, existing research commonly overlooks the fine-grained class relationships within the semantic space when addressing heterogeneous problems, particularly domain shift. This oversight results in ambiguities between class representation. To overcome this challenge, we propose a novel federated segmentation framework that strikes class consistency, termed FedSaaS. Specifically, we introduce class exemplars as a criterion for both local- and global-level class representations. On the server side, the uploaded class exemplars are leveraged to model class prototypes, which supervise global branch of clients, ensuring alignment with global-level representation. On the client side, we incorporate an adversarial mechanism to harmonize contributions of global and local branches, leading to consistent output. Moreover, multilevel contrastive losses are employed on both sides to enforce consistency between two-level representations in the same semantic space. Extensive experiments on several driving scene segmentation datasets demonstrate that our framework outperforms state-of-the-art methods, significantly improving average segmentation accuracy and effectively addressing the class-consistency representation problem.

## 1 Introduction

Semantic segmentation plays a pivotal role in various fields, such as autonomous driving [Feng *et al.*, 2020], medical diagnosis [Qureshi *et al.*, 2023], and remote sensing [Yuan *et al.*, 2021], where precise pixel-level classification is critical. While deep learning has significantly improved segmentation accuracy through the use of large datasets, the need for extensive labeled data is often hindered by concerns surrounding data privacy and security. Federated learning (FL) addresses this challenge by facilitating collaborative, decentralized model training while preserving data privacy [McMahan *et al.*, 2017]. As such, FL represents a promising paradigm for enabling cross-region semantic segmentation.

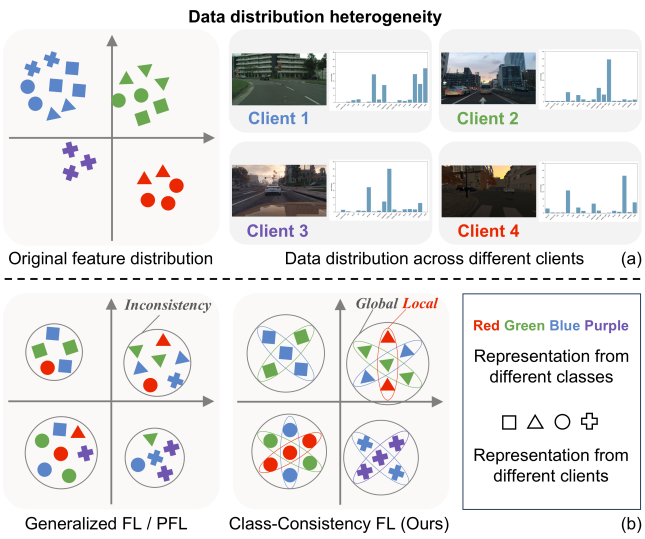


Figure 1: (a) Data from different regions exhibits severe domain shift and label skew. In the semantic space, the same class often shows significantly different distributions across clients. (b) Existing FL methods primarily focus on generalization or personalization (left) but often ignore class-level alignment. Class-consistency FL (right) ensures consistent representation of the same class by aligning and constraining both global- and local-level representations.

However, federated semantic segmentation continues to face significant challenges due to the heterogeneity of client data. In real-world applications, data from different clients often exhibit substantial variation, stemming from factors like sensor biases, environmental changes, and user preferences. This heterogeneity gives rise to discrepancies in the learned representations, which complicates the model's ability to generalize across clients [Huang *et al.*, 2023]. This issue becomes particularly prominent in FL when clients share the same objects but exhibit distinct feature distributions, a

*Corresponding authors.
†Full version of this paper can be found in [Yu *et al.*, 2025].

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

phenomenon known as domain shift. As illustrated in Figure 1(a), driving images from different domains exemplify this challenge: data from clients 1 and 2 are derived from real street scenes in different regions, while clients 3 and 4 consist of synthetically generated data. Domain shift factors can cause the same objects to appear visually distinct, leading to misaligned representation distributions in the semantic space and impeding accurate class identification.

Existing research often overlooks the fine-grained class relationships within the semantic space when addressing domain shift factors. In general FL methods, such as [Wang *et al.*, 2025; Xu *et al.*, 2024; Collins *et al.*, 2021], clients focus on extracting effective local-level representations from their specific domains, while the server aims to capture global-level representations shared across multiple domains. Since these two types of representations correspond to different scales of data understanding, inconsistencies may arise between local and global representations in the semantic space. This issue is particularly pronounced in federated semantic segmentation tasks, where both local- and global-level representations are crucial for capturing fine-grained class-level relationships, both intra-class and inter-class. Existing solutions [Ma *et al.*, 2024; Miao *et al.*, 2023; Kou *et al.*, 2024] often employ style transfer and contrastive learning to enhance generalization or adapt to local characteristics. However, these approaches fail to address the inconsistency problem of class representations between local and global semantic spaces. As shown in Figure 1(b), such inconsistencies can lead to semantic mismatches, where the same class is represented differently across domains. This leads to divergent semantic interpretations, ultimately impairing the model's ability to generalize effectively. To address the challenge of class inconsistency, this paper explores methods for aligning and constraining both local and global class representations within the semantic space.

In this work, we propose a class-consistency **Fed**erated semantic **S**egmen**t**ation approach via glob**a**l prototype **S**upervision and local adversarial harmonization, termed **FedSaaS**. To measure class representations at both global and local levels in the semantic space, we introduce class exemplars as a criterion, inspired by mask average pooling [Siam *et al.*, 2019]. FedSaaS leverages class exemplars on the server side to train the global model and generate class prototypes that supervise the global branch of the client model, ensuring alignment with global-level class representations. On the client side, we integrate an adversarial mechanism to harmonize the contributions of the local and global branches, thereby achieving consistent outputs. Furthermore, multilevel contrastive losses based on class exemplars are employed on both the client and server sides to enforce consistency within the same semantic space. We evaluate FedSaaS on five autonomous driving scene datasets, constructing datasets with varying levels of heterogeneity (slight and severe). Experimental results demonstrate that FedSaaS outperforms state-of-the-art methods across different degrees of domain shift.

The contributions of this paper are highlighted as follows:

- We propose FedSaaS, a federated semantic segmentation framework that introduces class exemplars to achieve consistency in class representations at both global and local levels within the semantic space.
- We supervise the alignment of local class representations by modeling class prototypes and integrate an adversarial mechanism within the client to harmonize the contributions of the global and local branches, thereby ensuring consistent outputs. Multilevel contrastive losses are introduced to further enhance the consistency between the two-level representations.
- Experiments on driving scene datasets demonstrate the superior performance of FedSaaS. Ablation studies and empirical analyses further validate its effectiveness in achieving consistency, improving segmentation precision, and enhancing communication efficiency.

## 2 Related Work

**Semantic Segmentation**. It is a task that assigns each pixel in an image to a predefined category. State-of-the-art methods predominantly employ encoder-decoder architectures based on various network models, including convolutional neural networks [Long *et al.*, 2015; Ronneberger *et al.*, 2015; Zhao *et al.*, 2017; Chen *et al.*, 2017], vision transformers [Dosovitskiy *et al.*, 2020; Liu *et al.*, 2021], diffusion models [Tian *et al.*, 2024; Amit *et al.*, 2021], and Mamba [Xing *et al.*, 2024]. These methods typically rely on large-scale labeled datasets and centralized training, achieving high segmentation accuracy on public benchmarks. However, the centralized training paradigm raises significant concerns related to data privacy and accessibility. In response, recent research has begun to explore distributed training frameworks as a promising alternative in such contexts.

**Federated Learning.** FL is a suitable paradigm for meeting the distributed training requirements in semantic segmentation tasks. This approach has made significant progress in addressing generalization challenges such as label shift and domain shift [Wang *et al.*, 2025; Xu *et al.*, 2024]. Personalized federated learning (PFL) further extends this paradigm by accounting for the specific needs of each client, adapting the global model to better align with the data distribution of individual clients [Collins *et al.*, 2021]. Nevertheless, unlike traditional FL tasks, semantic segmentation introduces unique challenges due to its requirement for precise pixel-level classification. This necessitates the retention of fine-grained spatial information at the local level, which complicates the direct application of existing general FL methods. These challenges are particularly pronounced in scenarios with severe data heterogeneity.

**Federated Semantic Segmentation.** The problem of FL-based semantic segmentation was first studied by [Michieli and Ozay, 2021]. Existing methods can be broadly categorized into two types: local personalization and global generalization. The first type primarily focuses on PFL frameworks, aiming to adapt local models to the specific characteristics of each client [Tan *et al.*, 2022]. This approach has been widely applied in fields such as medical image analysis and autonomous driving [Xie *et al.*, 2024;

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

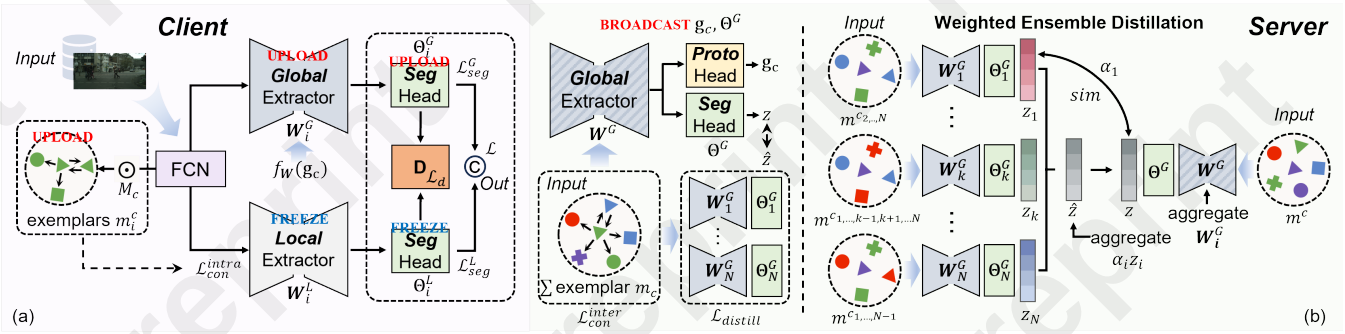


Figure 2: Overview of the FedSaaS Framework. (a) Client-side details. (b) Server-side and weighted integrated distillation details.

Wang *et al.*, 2023b; Kou *et al.*, 2024]. These works typically fuse various forms of local features learned by clients with shared features on the server to align with the distribution of local data. However, the fusion process often lacks appropriate constraints, which can result in an overemphasis on either global or local features, leading to an imbalance between local and global outputs. To address this limitation, we introduce an adversarial mechanism within each client, which promotes mutual learning between local and global branches, thereby facilitating a balanced contribution from both. The second category, global generalization, aims to enhance generalization by learning shared knowledge across domains, with a particular focus on addressing style heterogeneity within each domain [Ma *et al.*, 2024; Fantauzzo *et al.*, 2022]. Commonly adopted techniques include knowledge distillation [Wang *et al.*, 2023a], ensemble learning [Gong *et al.*, 2022], and prompt learning [Su *et al.*, 2024]. To mitigate fine-grained domain heterogeneity, an effective strategy is to combine contrastive loss to map local pixel embeddings into a global semantic space [Miao *et al.*, 2023; Tan *et al.*, 2024] and learn global class embeddings. Inaccurate mapping often occurs, leading to inconsistencies between local and global representations. To tackle this challenge, we introduce class exemplars to explore both intra-class and inter-class relationships, which are then used to supervise the client-side global branches, ensuring alignment with global-level representations in the semantic space.

# 3 Methodology

## 3.1 Overview

Figure 2(a) illustrates the overall framework of the proposed FedSaaS approach. The training objective is to ensure that the local and global representations of each class, derived from all client datasets, are mapped into a shared semantic space. For a total of $N$ clients, each client holds its own dataset $(\mathcal{D}_1, \ldots, \mathcal{D}_N)$. To maintain consistency between local and global representations for each class, we train a model $\mathcal{F}$ on each client $(\mathcal{F}_1, \ldots, \mathcal{F}_N)$, with collaborative training occurring across them. Specifically, each model $\mathcal{F}$ consists of two branches: a global branch $\mathcal{F}^G$ and a local branch $\mathcal{F}^L$. Although the data used by the two branches partially overlap, their representations remain inconsistent. Our goal is to align the representations at both levels by mapping them into a unified semantic space, thereby enabling harmonized outputs.

We follow the model decoupling idea proposed by [Collins *et al.*, 2021], dividing the backbone into two components: 1) a feature extractor $\mathbf{W} : \mathbb{R}^D \to \mathbb{R}^K$, which maps input samples to the feature semantic space, and 2) a segmentation head $\mathbf{\Theta} : \mathbb{R}^K \to \mathbb{R}^C$, which maps feature semantic space to label space. The final fully connected layer in a given backbone network is treated as the segmentation head. Here, the parameters $D$, $K$, and $C$ represent the dimensions of the input, feature, and label spaces, respectively.

The FedSaaS training process involves operations on both the client and server sides. On the client side, the training dataset consists of raw image data. On the server side, the training dataset comprises class exemplars uploaded from clients. These exemplars are obtained by multiplying the output of pretrained fully convolutional network (FCN) with the mask image corresponding to each class in the original image. To ensure compatibility for this multiplication, the output size of the FCN must match the size of the original image. This process is formally defined as: $m_i^c = \text{FCN}(x) \odot M_i^c$, where $x \in \mathbb{R}^{H \times W}$ represents the original image, and $M_i^c \in \mathbb{R}^{H \times W}$ denotes the mask of a specific class, and $\odot$ denotes the Hadamard product. Class exemplars capture the spatial distribution and correlations of the corresponding class, making them a valuable criterion for achieving alignment between local- and global-level class representations.

## 3.2 Weighted Ensemble Distillation

On the server side, a global branch $\mathcal{F}^G$ is trained using weighted ensemble distillation based on class exemplars to map and constrain the global class representation. For the global branch $\mathcal{F}_k^G$ of client $k$, when provided with unseen exemplars $m_i^c, i \neq k$, the predicted logits are computed as $z_k = \mathcal{F}_k^G(\mathbf{W}_k^G, \mathbf{\Theta}_k^G; m_i^c)$. Simultaneously, the server-side global branch $\mathcal{F}^G$ generates predictions for the same exemplars, producing logits $z = \mathcal{F}^G(\mathbf{W}^G, \mathbf{\Theta}^G; m_i^c)$. The similarity between the outputs of the client-side and server-side global branches is measured by $sim(z_k, z)$. A higher value of $sim(z_k, z)$ indicates that the client's global branch exhibits better generalization to unseen domains, and thus, it is assigned a higher weight. To quantify this, we define a weight function $\alpha_i$ that reflects the generalization capability of each

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

client's global branch: $\alpha_i = \frac{sim(z_k,z)}{\sum_{i=1,i\neq k}^{N} sim(z_k,z)}$, where the similarity measure $sim(\cdot)$ can be implemented using metrics such as Kullback-Leibler divergence or cosine similarity.

Subsequently, the logits from the client-side global branches are aggregated in a weighted manner to produce new predicted logits: $\hat{z} = \sum_{i=1}^{N} \alpha_i \cdot z_i$. The server-side global branch is then trained by minimizing the following loss function designed to learn from the aggregated logits:

$$\mathcal{L}_{distill} = \mathbb{E}_{m^c}[sim(z, \hat{z})]. \tag{1}$$

This process enables the server-side global branch to effectively integrate knowledge from multiple clients, thereby achieving enhanced generalization across diverse domains.

To enhance the constraint of global semantic consistency, we introduce inter-client contrastive learning on the server side, leveraging class exemplars. Class exemplars encapsulate semantic information specific to client categories, providing a direct basis for constructing positive and negative samples in two-level contrastive learning. We define the inter-client contrastive loss as:

$$\mathcal{L}_{con}^{inter} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(v_i^c v^+/\tau\right)}{\exp\left(v_i^c v^+/\tau\right) + \sum_{v^-} \exp\left(v_i^c v^-/\tau\right)}, \tag{2}$$

where $v_i^c$ is the normalized vector of the class exemplar $m_i^c$, $v^+$ represents positive samples from the same class, $v^-$ represents negative samples from other classes, and $\tau$ is the temperature coefficient. The overall loss function for training the server-side global branch is

$$\mathcal{L}_g = \mathcal{L}_{distill} + \mathcal{L}_{con}^{inter}. \tag{3}$$

By training the global branch from the client to the server, the consistency of similar representations is improved, ensuring aligned semantic representations across clients.

### 3.3 Global Prototype Supervision

The server further generates class prototypes to supervise the client-side global branch, ensuring alignment between local- and global-level class representations. As illustrated in Figure 3, these class prototypes are designed based on the deep representations of class exemplars, denoted as $h^c(x, y)$.

To construct the class prototypes, we generate category distribution vectors and category co-occurrence relationships separately. First, we perform a weighted average pooling across all class exemplars to obtain the category distribution vector for each class. The weight is calculated as: $\beta_c = 1 - K_c^{-1}/max(K_{c'}^{-1}) - min(K_{c'}^{-1})$, where $K_c$ represents the number of class exemplars for class $c$ and $K_c^{-1}$ denotes its inverse value. This weighting scheme addresses the issue of underrepresentation for classes with fewer exemplars during training by assigning them larger weights. Based on the deep representations $h^c(x, y)$ and the computed weights, the class distribution vector $v_c$ for each class is calculated as:

$$\mathbf{v}_c = \Phi\left[\frac{1}{K_c}\sum_{k=1}^{K_c} \beta_c \cdot h^c(x, y)\right], \tag{4}$$

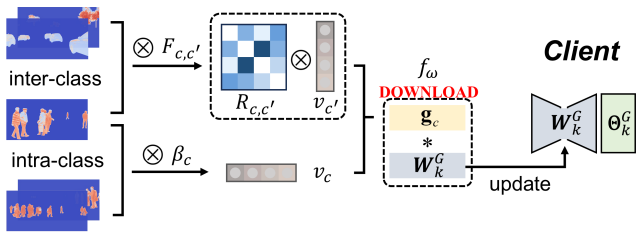


Figure 3: Diagram of the prototype head and process of global prototype supervision.

where $\Phi[\cdot]$ represents a dimensionality reduction operation.

Next, we generate the co-occurrence relationship $\phi_{c,c'}$ between class $c$ and class $c'$, which reflects their relative distribution across all spatial positions. It is computed as

$$\phi_{c,c'} = \frac{\sum_{\Omega_c}\sum_{\Omega_{(c,c')}} \mathbb{I}[h^c(x,y) \neq 0 \wedge h^{c'}(x',y') \neq 0]}{\sum_{\Omega} \mathbb{I}[h^c(x,y) \neq 0]}, \tag{5}$$

where $\Omega_{(c,c')}$ represents the neighborhood region of classes $c$ and $c'$, and $\mathbb{I}[\cdot]$ denotes the indicator function. Based on the co-occurrence relationship, we further incorporate positional information and regional relationships of class exemplars to calculate the correlation $R_{c,c'}$ between classes $c$ and $c'$:

$$R_{c,c'} = \phi_{c,c'} \cdot \sum_{\Omega_c} h^c(x,y) \cdot \cdot \left(\sum_{\Omega_{(c,c')}} (h^c(x',y')) \cdot K_d(x,y,x',y')\right), \tag{6}$$

where $K_d(\cdot)$ represents a Gaussian kernel function that weights the influence of distance between adjacent pixels.

By combining the distribution vector $\mathbf{v}_c$ and the correlation $R_{c,c'}$ for each class, we obtain the class prototype $\mathbf{g}_c$:

$$\mathbf{g}_c = \mathbf{v}_c + \frac{1}{|C_l|}\sum_{c' \in C_l, c' \neq c} R_{c,c'} \cdot v_{c'}, \tag{7}$$

where $C_l$ denotes the total number of classes.

Class prototypes contain two key features: intra-class distribution and inter-class co-occurrence relationship. Once the clients receive the class prototypes $\mathbf{g}_c$, these prototypes are converted into dynamic weights for the convolutional kernels in the global branch. To achieve this, we define a weighted generation network $f_{\mathbf{w}}(\mathbf{g}_c)$, which projects $\mathbf{g}_c$ into the weight dimension. This network can be implemented using a multilayer perceptron (MLP). Through the supervision of class prototypes, the local-level class representations generated by the client-side global branch are effectively aligned with the global representations, ensuring consistency across domains.

### 3.4 Local Adversarial Harmonization

On the client side, the local branch and global branch focus on different levels of class representations, necessitating a mechanism to balance their contributions to the final model output.

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

To address this, we propose a local adversarial harmonization mechanism that facilitates mutual learning between the two branches by confusing a newly trained discriminator. This mechanism is divided into two stages: discriminator training and branches training.

**Discriminator training.** Given the output logits of the local and global branches, a domain discriminator is trained to distinguish their origins. The objective is to maximize the discriminator's classification accuracy, ensuring it can correctly differentiate between the outputs of the local and global branches. To this end, the discriminator maximizes a binary cross-entropy loss function:

$$\mathcal{L}_d = -\mathbb{E}_d[p \log \hat{p} + (1-p)\log(1-\hat{p})], \tag{8}$$

where $p \in \{0, 1\}$ represents the ground-truth label, indicating whether the current input originates from the local branch ($p = 0$) or the global branch ($p = 1$), and $\hat{p}$ is the prediction output of the domain discriminator.

**Branches training.** To counter the discriminator, the training objective of the local and global branches is to make the domain discriminator incapable of distinguishing their outputs. This is achieved by incorporating adversarial constraints into their loss functions, which aim to minimize the discriminator loss and thereby deceive the discriminator. For the global branch, we adjust the parameters of the feature extractor under the supervision of global prototypes. The parameters of the segmentation head are first replaced with the server-side parameters $\mathbf{\Theta}^G$, and subsequently optimized using a segmentation loss $\mathcal{L}_{seg}^G$. The training losses for the two branches are defined as:

$$\begin{aligned} \mathcal{L}_{global} &= \mathcal{L}_{seg}^G + \lambda \mathcal{L}_d, \\ \mathcal{L}_{local} &= \mathcal{L}_{seg}^L + \lambda \mathcal{L}_d + \mathcal{L}_{con}^{intra}, \end{aligned} \tag{9}$$

where $\mathcal{L}_{seg}^L$ is the local segmentation loss and $\lambda$ is a hyperparameter that adjusts the weight of discriminator loss $\mathcal{L}_d$. A contrastive loss, $\mathcal{L}_{con}^{intra}$, derived from the client's own class exemplars, is added to the local branch loss to constrain the local-level class representations:

$$\mathcal{L}_{con}^{intra} = -\log \frac{\exp(v^c v^+/\tau)}{\exp(v^c v^+/\tau) + \sum_{v^-} \exp(v^c v^-/\tau)}. \tag{10}$$

The logits from the two branches are summed and averaged to produce the final output. Through the local adversarial mechanism, the segmentation model dynamically harmonizes the contributions of local and global representations, ensuring consistent outputs.

# 4 Experiments

## 4.1 Experimental Settings

To evaluate the performance of the proposed method, we select the driving scene segmentation task for both training and validation. This task involves a wide range of complex categorical objects and presents real-world class imbalance issues, posing significant challenges in achieving semantic consistency across categories. We construct datasets for two heterogeneity scenarios: slight and severe, based on training difficulty. These datasets differ notably in terms of domain shift and label shift. For the slight heterogeneity scenario, we use the widely adopted baseline dataset, Cityscapes [Cordts *et al.*, 2016], which includes street views from multiple cities across Europe. Due to small geographical and device-related differences, this dataset exhibits minimal domain shift. For the severe heterogeneity case, we utilize five driving scene datasets: Cityscapes, Mapillary Vistas [Neuhold *et al.*, 2017], BDD100K [Yu *et al.*, 2020], GTA5 [Richter *et al.*, 2016], and Synthia [Ros *et al.*, 2016]. This collection not only contains real or virtual street views from cities worldwide but also includes simulated datasets captured from various angles and devices, resulting in both severe domain and label shifts. In this scenario, the model's ability to generalize across diverse scenes is more rigorously tested.

We select representative FL methods and state-of-the-art (SOTA) federated segmentation approaches for comparison, including FedAvg [McMahan *et al.*, 2017], FedProx [Li *et al.*, 2020], FedDrive [Fantauzzo *et al.*, 2022], FedSeg [Miao *et al.*, 2023], and FedST [Ma *et al.*, 2024]. Due to differences in dataset size and annotation types, we standardize the data preprocessing by cropping all images and resizing them to $512 \times 1024$. For model training, we adopt the BiSeNet V2 architecture [Yu *et al.*, 2021], a lightweight network designed to capture both spatial features and high-level semantic context. The temperature coefficient $\tau$ for the multilevel contrastive loss and the weight $\lambda$ for the adversarial loss are set to $0.05$ and $0.1$, respectively. We set batch size to 16, with 10 local iterations and 50 communication rounds. To evaluate performance, we use two common semantic segmentation metrics: mean Intersection over Union (mIoU), which measures the intersection over union between predicted and ground truth pixels averaged across all categories, and Pixel Accuracy, which describes the ratio of correctly classified pixels.

## 4.2 Main Results

**Quantitative Analysis.** We evaluate the performance on validation datasets, reporting the average and fluctuation deviation over three independent tests. Table 1 presents a comparison between our FedSaaS method and other SOTA methods. As shown, our method achieves the best performance across all data environments. In the slight heterogeneity scenario, we observe a modest improvement over the current best methods, with increases of 2.75% in Accuracy and 1.66% in mIoU. In the severe heterogeneity case, FedSaaS performs significantly better, with improvements of 6.68% in Accuracy and 4.48% in mIoU. This is because FedSaaS leverages class consistency to jointly represent the characteristics of classes at both the local and global levels. Additionally, to assess generalization, we separate Cityscapes from the severely heterogeneous dataset and use it as an unseen domain for testing. We average the results of the remaining four clients, and the scores demonstrate that FedSaaS outperforms the SOTA methods in this scenario as well.

**Qualitative Analysis.** Figure 4 presents the visualization results of testing different clients under the severe heterogeneity scenario. It is observed that Feddrive and FedST, both designed to address the domain shift problem, improve segmentation accuracy only for certain classes compared to

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

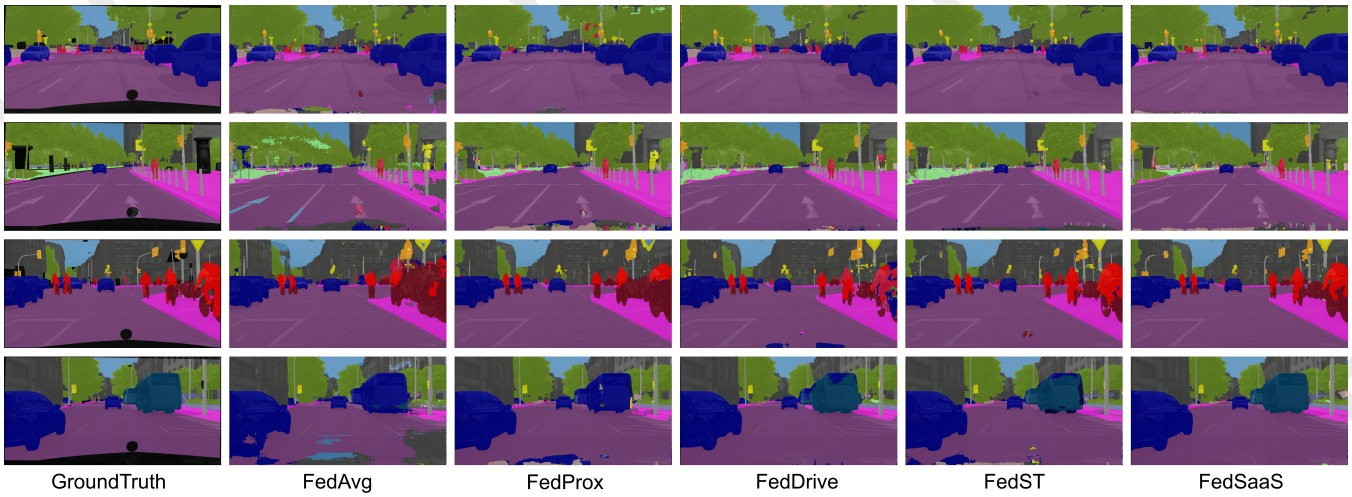


Figure 4: Comparison of visual results from different methods on datasets with severe heterogeneity.

| Method | Slight Heterogeneity | | Severe Heterogeneity | | Unseen-Domain | |
|---|---|---|---|---|---|---|
| | Acc ± std | mIoU ± std | Acc ± std | mIoU ± std | Acc ± std | mIoU ± std |
| FedAvg (PMLR 2017) | 79.20 ± 1.96 | 47.92 ± 1.53 | 66.58 ± 2.58 | 37.19 ± 2.12 | 61.61 ± 1.68 | 34.28 ± 2.04 |
| FedProx (MLSys 2020) | 78.93 ± 0.67 | 47.62 ± 0.90 | 67.06 ± 1.26 | 38.08 ± 1.62 | 63.09 ± 1.91 | 35.37 ± 1.54 |
| FedDrive (IROS 2022) | 84.74 ± 1.31 | 51.14 ± 0.77 | 73.24 ± 1.31 | 42.83 ± 0.56 | 70.40 ± 1.85 | 43.06 ± 0.78 |
| FedSeg (CVPR 2023) | 84.36 ± 1.81 | 51.36 ± 1.29 | 72.86 ± 2.89 | 40.14 ± 2.04 | 67.04 ± 2.25 | 37.48 ± 2.30 |
| FedST (AAAI 2024) | 85.32 ± 0.68 | 52.60 ± 0.36 | 75.58 ± 1.99 | 44.19 ± 1.37 | 73.43 ± 1.89 | 44.77 ± 0.96 |
| FedSaaS (Ours) | **88.57 ± 0.95** | **54.34 ± 0.62** | **82.26 ± 2.30** | **48.67 ± 1.18** | **74.15 ± 1.47** | **45.96 ± 1.14** |
| Backbone | 81.09 ± 0.73 | 49.61 ± 0.58 | 68.14 ± 1.77 | 37.55 ± 0.93 | 65.24 ± 1.66 | 35.63 ± 1.47 |
| + *Proto.* | 85.53 ± 1.04 | 53.57 ± 0.73 | 76.06 ± 2.49 | 43.08 ± 1.72 | 70.08 ± 1.36 | 39.08 ± 1.25 |
| + *Proto.* + $\mathcal{L}_{con}$ | 88.21 ± 1.27 | 54.21 ± 0.66 | 80.38 ± 2.61 | 46.21 ± 1.46 | 73.96 ± 1.86 | 45.83 ± 1.30 |
| + *Proto.* + $\mathcal{L}_{con}$ + $\mathcal{L}_d$ | **88.57 ± 0.95** | **54.34 ± 0.62** | **82.26 ± 2.30** | **48.67 ± 1.18** | **74.15 ± 1.47** | **46.06 ± 1.14** |

Table 1: Performance comparison (%) under slight and severe heterogeneity scenarios across various methods (top), and effectiveness validation of FedSaaS modules (bottom).

traditional FL methods. In contrast, our FedSaaS method improves segmentation performance across all categories. By aligning and constraining both local and global class representations, our approach ensures high segmentation performance for all clients. More detailed comparisons are shown in the Appendix of full version of this paper.

**Ablation Studies.** We evaluate the effectiveness of each module, as shown in the lower half of Table 1. The initial backbone structure consists of local and global branches on client side and follows the FedAvg configuration. The outputs from both branches are combined through summation and averaging to produce the final output. Subsequently, we incorporate class exemplars into the backbone, enabling the segmentation head $\mathbf{\Theta}^G$ to be trained on server side while simultaneously generating class prototypes (abbreviated as *Proto.* in the subsequent charts) to supervise parameter updates in client-side global branch. This enhancement significantly improves the performance of the overall framework, resulting in an accuracy increase ranging from 4.42% to 7.92%.

We further incorporate the two-level contrastive loss (abbreviated as $\mathcal{L}_{con}$ in the table) based on class exemplars to enforce class representation constraints at both the local and global levels. Compared to the results from the slight heterogeneity scenario, the combination of prototype supervision and multilevel contrastive loss yields a particularly notable performance improvement in the severe heterogeneity case, with an increase of 12.24% in Accuracy and 8.66% in mIoU. This highlights the importance of achieving class consistency in scenarios with significant domain shifts. We also introduce an adversarial harmonization module before the outputs of the two client branches. By comparing the results before and after the inclusion of this module, we observe an enhancement in the model's performance. Notably, the alignment between global and local representations is achieved through the above mechanisms, which enables the adversarial mechanism to operate effectively on both branches.

### 4.3 Empirical Analysis

**Visualization.** We use t-SNE [Van der Maaten and Hinton, 2008] to visualize the pixel embeddings of semantic classes under the severe heterogeneity scenario, as shown in Figure 5.

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

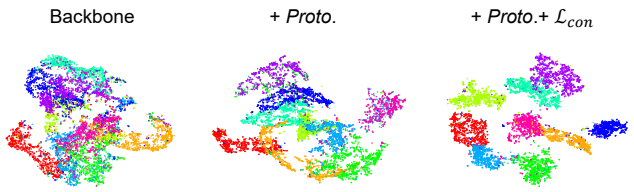


Figure 5: Visualization of the pixel embeddings for different semantic classes.

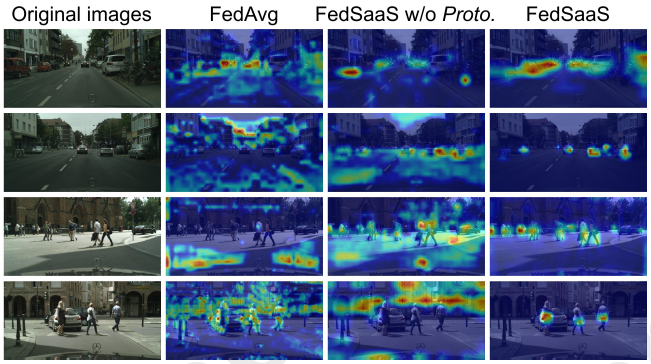


Figure 6: Grad-CAM visualization of the feature attention regions in the last convolutional layer of the global branch.

The results demonstrate that, without additional modules, the model produces poor embeddings, with most semantic class pixels intermingled. Incorporating the *Proto.* module results in a noticeable separation between pixels of different semantic classes. The introduction of the two-level contrastive loss $\mathcal{L}_{con}$ further enhances the divergence between categories in the embedding space, highlighting its critical role in constraining semantic class representations.

We visualize the global branch's attention to categories using Grad-CAM [Selvaraju *et al.*, 2017]. As shown in Figure 6, we compare the attention given by the model to the most common categories in driving scenes—namely, pedestrians and vehicles—by analyzing the output of the last convolutional layer in both the FedAvg and FedSaaS global branches. It is evident that both FedAvg and FedSaaS w/o *Proto.* exhibit insufficient focus and attention to the image. After incorporating the *Proto.* module, the global branch is able to accurately locate and identify the corresponding categories in the respective channels. This supervision of local class alignment through global class prototypes enables the model to gain a richer understanding of various categories, which is crucial for achieving high segmentation accuracy.

**Communication Efficiency.** Figure 7 illustrates the performance curves of different methods during training. It is observed that, compared to other methods within the same communication round, FedSaaS achieves superior performance with enhanced communication efficiency. Additionally, the need to upload class exemplars from clients introduces extra communication overhead. We examine the impact of randomly uploading a subset of class exemplars on performance. As shown in Table 2, as the number of uploaded exemplars increases, segmentation accuracy improves. Notably, test accuracy when uploading approximately 50%-75% of the exemplars is significantly higher than when uploading only 25%. This suggests that, in scenarios where the dataset is large or communication constraints are tight, uploading around half of the class exemplars can still achieve relatively high accuracy, thereby reducing transmission costs.

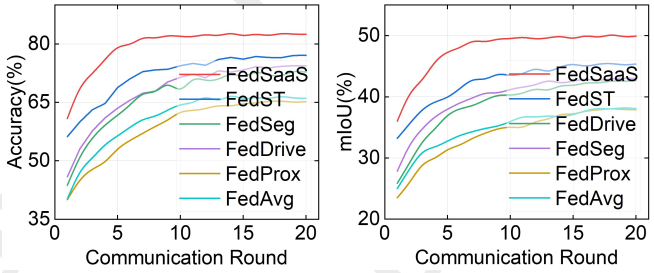


Figure 7: Comparison of communication efficiency.

| Upload ratio | Slight Heterogeneity Acc | Slight Heterogeneity mIoU | Severe Heterogeneity Acc | Severe Heterogeneity mIoU |
|---|---|---|---|---|
| 25% | 84.26 | 51.27 | 69.12 | 35.97 |
| 50% | 86.53 | 53.84 | 75.50 | 43.28 |
| 75% | 87.96 | 54.08 | 77.74 | 44.36 |

Table 2: Performance comparison (%) of different upload ratios of class exemplars under slight and severe heterogeneity scenarios.

**Stability Analysis.** The branch and discriminator are optimized toward opposing objectives: the branch aims to minimize its loss, while the adversarial dynamics force the discriminator's loss to increase. This competition may introduce initial instability in the training process, particularly under severe data heterogeneity, manifesting as transient loss fluctuations or occasional branch misalignment. To mitigate these effects, we adopt two strategies: 1) We initialize the adversarial weight $\lambda$ at a reduced value to limit early-stage perturbation. As optimization stabilizes, $\lambda$ is progressively increased to its target value. 2) During training, sharp performance declines trigger a rollback to the best checkpoint, accompanied by gradient clipping on $\mathcal{L}_d$ to suppress instability. As shown in Figure 7, these measures maintain stable performance growth without compromising convergence speed.

## 5 Conclusion

In this paper, we have proposed a class-consistency FL approach tailored for semantic segmentation tasks. To address the ambiguity in class representations caused by domain shifts, we have introduced a novel framework, FedSaaS, which leverages class exemplars as a criterion to ensure consistency between local and global representations. Specifically, we have sequentially incorporated class prototypes and adversarial mechanism to achieve two-level representation alignment, thereby ensuring consistent outputs at both branches. Extensive experiments conducted on several datasets demonstrate that FedSaaS outperforms state-of-the-art methods in addressing the class-consistency problem.

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

## Acknowledgments

This work was supported in part by the NSFC under Grant 62402256, in part by the Taishan Scholars Program under Grants tsqn202211203 and tsqn202408239, in part by the Shandong Provincial Nature Science Foundation of China under Grant ZR2024MF100, and in part by the QLU/SDAS Pilot Project for Integrated Innovation of Science, Education, and Industry under Grant 2024ZDZX08.

## References

[Amit *et al.*, 2021] Tomer Amit, Tal Shaharbany, Eliya Nachmani, and Lior Wolf. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*, 2021.

[Chen *et al.*, 2017] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2017.

[Collins *et al.*, 2021] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pages 2089–2099. PMLR, 2021.

[Cordts *et al.*, 2016] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, et al. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[Fantauzzo *et al.*, 2022] Lidia Fantauzzo, Eros Fanì, Debora Caldarola, Antonio Tavera, Fabio Cermelli, Marco Ciccone, and Barbara Caputo. Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 11504–11511. IEEE, 2022.

[Feng *et al.*, 2020] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2020.

[Gong *et al.*, 2022] Xuan Gong, Abhishek Sharma, Srikrishna Karanam, Ziyan Wu, Terrence Chen, David Doermann, and Arun Innanje. Preserving privacy in federated learning with ensemble cross-domain knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11891–11899, 2022.

[Huang *et al.*, 2023] Wenke Huang, Mang Ye, Zekun Shi, He Li, and Bo Du. Rethinking federated learning with domain shift: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16312–16322. IEEE, 2023.

[Kou *et al.*, 2024] Wei-Bin Kou, Qingfeng Lin, Ming Tang, Sheng Xu, Rongguang Ye, Yang Leng, Shuai Wang, Guofa Li, Zhenyu Chen, et al. pfedlvm: A large vision model (lvm)-driven and latent feature-based personalized federated learning framework in autonomous driving. *arXiv preprint arXiv:2405.04146*, 2024.

[Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE/CVF CConference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015.

[Ma *et al.*, 2024] Boyuan Ma, Xiang Yin, Jing Tan, Yongfeng Chen, Haiyou Huang, Hao Wang, Weihua Xue, and Xiaojuan Ban. Fedst: Federated style transfer learning for non-iid image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4053–4061, 2024.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

[Miao *et al.*, 2023] Jiaxu Miao, Zongxin Yang, Leilei Fan, and Yi Yang. Fedseg: Class-heterogeneous federated learning for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8042–8052, 2023.

[Michieli and Ozay, 2021] Umberto Michieli and Mete Ozay. Prototype guided federated learning of visual feature representations. *arXiv preprint arXiv:2105.08982*, 2021.

[Neuhold *et al.*, 2017] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4990–4999, 2017.

[Qureshi *et al.*, 2023] Imran Qureshi, Junhua Yan, Qaisar Abbas, Kashif Shaheed, Awais Bin Riaz, Abdul Wahid,

**Preprint – IJCAI 2025**: This is the accepted version made available for conference attendees.
**Do not cite**. The final version will appear in the IJCAI 2025 proceedings.

Muhammad Waseem Jan Khan, and Piotr Szczuko. Medical image segmentation using deep semantic-based methods: A review of techniques, applications and emerging trends. *Information Fusion*, 90:316–352, 2023.

[Richter *et al.*, 2016] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118. Springer, 2016.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Ros *et al.*, 2016] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3234–3243, 2016.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 618–626, 2017.

[Siam *et al.*, 2019] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5249–5258, 2019.

[Su *et al.*, 2024] Shangchao Su, Mingzhao Yang, Bin Li, and Xiangyang Xue. Federated adaptive prompt tuning for multi-domain collaborative learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 15117–15125, 2024.

[Tan *et al.*, 2022] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):9587–9603, 2022.

[Tan *et al.*, 2024] Jieyi Tan, Yansheng Li, Sergey A Bartalev, Bo Dang, Wei Chen, Yongjun Zhang, and Liangqi Yuan. Bridging data islands: Geographic heterogeneity-aware federated learning for collaborative remote sensing semantic segmentation. *arXiv preprint arXiv:2404.09292*, 2024.

[Tian *et al.*, 2024] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse attend and segment: Unsupervised zero-shot segmentation using stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3554–3563, 2024.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008.

[Wang *et al.*, 2023a] Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, and Zhigang Zeng. Dafkd: Domain-aware federated knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20412–20421, 2023.

[Wang *et al.*, 2023b] Jiacheng Wang, Yueming Jin, Danail Stoyanov, and Liansheng Wang. Feddp: Dual personalization in federated medical image segmentation. *IEEE Transactions on Medical Imaging*, 2023.

[Wang *et al.*, 2025] Xin Wang, Yanhan Wang, Ming Yang, Feng Li, Xiaoming Wu, Lisheng Fan, et al. FedSiam-DA: Dual-aggregated federated learning via siamese network for non-IID data. *IEEE Transactions on Mobile Computing*, 24(2):985–998, 2025.

[Xie *et al.*, 2024] Luyuan Xie, Manqing Lin, Siyuan Liu, ChenMing Xu, Tianyu Luan, Cong Li, Yuejian Fang, Qingni Shen, and Zhonghai Wu. pflfe: Cross-silo personalized federated learning via feature enhancement on medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 599–610. Springer, 2024.

[Xing *et al.*, 2024] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[Xu *et al.*, 2024] Yi Xu, Ying Li, Haoyu Luo, Xiaoliang Fan, and Xiao Liu. Fblg: A local graph based approach for handling dual skewed non-iid data in federated learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 8, pages 5289–5297, 2024.

[Yu *et al.*, 2020] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2636–2645, 2020.

[Yu *et al.*, 2021] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, 129:3051–3068, 2021.

[Yu *et al.*, 2025] Xiaoyang Yu, Xiaoming Wu, xin Wang, Dongrun Li, Ming Yang, and Peng Cheng. Fedsaas: Class-consistency federated semantic segmentation via global prototype supervision and local adversarial harmonization. *arXiv preprint arXiv:2505.09385*, 2025.

[Yuan *et al.*, 2021] Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021.

[Zhao *et al.*, 2017] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.