Federated Incremental Semantic Segmentation

Jiahua Dong¹,²,³*, Duzhen Zhang⁴*, Yang Cong¹,²†, Wei Cong¹,²,³, Henghui Ding⁴, Dengxin Dai⁴
¹State Key Laboratory of Robotics, Shenyang Institute of Automation,
Chinese Academy of Sciences, Shenyang, 110016, China.‡
²Institutes for Robotics and Intelligent Manufacturing,
Chinese Academy of Sciences, Shenyang, 110169, China.
³University of Chinese Academy of Sciences, Beijing, 100049, China.
⁴ETH Zürich, Zürich, 8092, Switzerland.
{dongjiahua1995, congyang81, congwei45, henghui.ding}@gmail.com, dai@vision.ee.ethz.ch

Abstract

Federated learning-based semantic segmentation (FSS) has drawn widespread attention via decentralized training on local clients. However, most FSS models assume categories are fixed in advance, thus heavily undergoing forgetting on old categories in practical applications where local clients receive new categories incrementally while have no memory storage to access old classes. Moreover, new clients collecting novel classes may join in the global training of FSS, which further exacerbates catastrophic forgetting. To surmount the above challenges, we propose a Forgetting-Balanced Learning (FBL) model to address heterogeneous forgetting on old classes from both intra-client and inter-client aspects. Specifically, under the guidance of pseudo labels generated via adaptive class-balanced pseudo labeling, we develop a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss to rectify intra-client heterogeneous forgetting of old categories with background shift. It performs balanced gradient propagation and relation consistency distillation within local clients. Moreover, to tackle heterogeneous forgetting from inter-client aspect, we propose a task transition monitor. It can identify new classes under privacy protection and store the latest old global model for relation distillation. Qualitative experiments reveal large improvement of our model against comparison methods. The code is available at https://github.com/JiahuaDong/FISS.

1. Introduction

Federated learning (FL) [13,20,22,44] is a remarkable decentralized training paradigm to learn a global model across

Figure 1. Exemplary FISS setting for medical diagnosis. Hundreds of hospitals including newly-joined ones receive new classes incrementally according to their own preference. FISS aims to segment new diseases consecutively via collaboratively learning a global segmentation model on private medical data of different hospitals.

distributed local clients without accessing their private data. Under privacy preservation, it has achieved rapid development in semantic segmentation [4,8,30] by training on multiple decentralized local clients to alleviate the constraint of data island that requires enormous finely-labeled pixel annotations [25]. As a result, federated learning-based semantic segmentation (FSS) [28,29] significantly economizes annotation costs in data-scarce scenarios via training a global segmentation model on private data of different clients [29].

However, existing FSS methods [16,25,28,29] unrealistically assume that the learned foreground classes are static and fixed over time, which is impractical in real-world dynamic applications where local clients receive streaming data of new categories consecutively. To tackle this issue, existing
FSS methods [28, 29, 35] typically enforce local clients to store all samples of previously-learned old classes, and then learn a global model to segment new categories continually via FL. Nevertheless, it requires large computation and memory overhead as new classes arrive continuously, limiting the application ability of FSS methods [16, 28]. If local clients have no memory to store old classes, existing FSS methods [16, 29] significantly degrade segmentation behavior on old categories \((i.e., \text{catastrophic forgetting})[40,47,48])\) when learning new classes incrementally. In addition, the pixels labeled as background in the current learning task may belong to old classes from old tasks or new foreground classes from future tasks. This phenomenon is also known as background shift [11, 36] that heavily aggravates heterogeneous forgetting speeds on old categories. More importantly, in practical scenarios, new local clients receiving new categories incrementally may join in global FL training irregularly, thus further exacerbating catastrophic forgetting to some extent.

To surmount the above real-world scenarios, we propose a novel practical problem called Federated Incremental Semantic Segmentation (FISS), where local clients collect new categories consecutively according to their preferences, and new local clients collecting unseen novel classes participate in global FL training irregularly. In the FISS settings, the class distributions are non-independent and identically distributed (Non-IID) across different clients, and training data of old classes is unavailable for all local clients. FISS aims to train a global incremental segmentation model via collaborative FL training on local clients while addressing catastrophic forgetting. In this paper, we use medical lesions segmentation [25, 29] as an example to better illustrate FISS, as shown in Figure 1. Hundreds of hospitals, as well as newly joined ones, collect unseen/new medical lesions continuously in clinical diagnosis. Considering privacy preservation, it is desired for these hospitals to learn a global segmentation model via FL without accessing each other’s data [44, 56].

A naive solution for FISS problem is to directly integrate incremental semantic segmentation [1,11,53] and FL [19,50] together. Nevertheless, such a trivial solution requires global server to have strong human prior about which and when local clients can collect new categories, so that global model learned in the latest old task can be stored by local clients to address forgetting on old classes via knowledge distillation [18,43]. Considering privacy preservation in the FISS, this privacy-sensitive prior knowledge cannot be shared between local clients and global server. As a result, this naive solution severely suffers from intra-client heterogeneous forgetting on different old classes caused by background shift [1,11,36,53] and inter-client heterogeneous forgetting across different clients brought by Non-IID class distributions.

To overcome the above-mentioned challenges, we develop a novel Forgetting-Balanced Learning (FBL) model, which alleviates heterogeneous forgetting on old classes from intra-client and inter-client perspectives. Specifically, to tackle intra-client heterogeneous forgetting caused by background shift, we propose an adaptive class-balanced pseudo labeling to adaptively generate confident pseudo labels for old classes. Under the guidance of pseudo labels, we propose a forgetting-balanced semantic compensation loss to rectify different forgetting of old classes with background shift via considering balanced gradient propagation of local clients. In addition, a forgetting-balanced relation consistency loss is designed to distill underlying category-relation consistency between old and new classes for intra-client heterogeneous forgetting compensation. Moreover, considering addressing heterogeneous forgetting from inter-client aspect, we develop a task transition monitor to automatically identify new classes without any human prior, and store the latest old model from global perspective for relation consistency distillation. Experiments on segmentation datasets reveal large improvement of our model over comparison methods. We summarize the main contributions of this work as follows:

- We propose a novel practical problem called Federated Incremental Semantic Segmentation (FISS), where the major challenges are intra-client and inter-client heterogeneous forgetting on old categories caused by intra-client background shift and inter-client Non-IID distributions.
- We propose a Forgetting-Balanced Learning (FBL) model to address the FISS problem via surmounting heterogeneous forgetting from both intra-client and inter-client aspects. As we all know, in the FL field, this is a pioneer attempt to explore a global continual segmentation model.
- We develop a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss to tackle intra-client heterogeneous forgetting across old classes, under the guidance of confident pseudo labels generated via adaptive class-balanced pseudo labeling.
- We design a task transition monitor to surmount inter-client heterogeneous forgetting by accurately recognizing new classes under privacy protection and storing the latest old model from global aspect for relation distillation.

2. Related Work

3. Problem Definition

As claimed in incremental semantic segmentation (ISS) [1, 11, 27, 36], some consecutive segmentation tasks are defined as \( T = \{T_t\}_{t=1}^{T} \), where the \( t \)-th \((t = 1, \cdots, T)\) task \( T_t = \{x^t_i, y^t_i\}_{i=1}^{N_t} \) is composed of \( N_t \) pairs of RGB images \( x^t_i \in \mathbb{R}^{H \times W \times 3} \) and labels \( y^t_i \in \mathbb{R}^{H \times W} \). \( H \) and \( W \) denote height and width of given images. The label space \( \mathcal{Y}_t^{K^t} \) of \( t \)-th incremental task consists of \( K^t \) new categories and background. \( K^t \) new classes have no overlap with \( K^0 = \sum_{i=1}^{t-1} K^i \cup \bigcup_{j=1}^{t-1} 3^{j} \) old classes learned from \( t - 1 \) old tasks. In the \( t \)-th task, we follow ISS methods [11, 36] to annotate \( K^0 \) old classes and other foreground classes from future learning tasks as background (i.e., background shift [1]), due to unavailable training data of \( K^0 \) old classes.

We then extend the settings of incremental semantic segmentation (ISS) [1, 36, 53] to Federated Incremental Semantic Segmentation (FISS). Denote global server as \( S_0 \) and \( L \) local clients as \( \{S_l\}_{l=1}^{L} \). In the FISS, at the \( r \)-th \((r = 1, \cdots, R)\) global round, we randomly select some local clients to aggregate gradients. When we choose the \( l \)-th local client to learn the \( t \)-th segmentation task, the latest global model \( \Theta_x^{t-1} \) is distributed to \( S_l \), and trained on private training data \( T^t_l = \{x^t_{li}, y^t_{li}\}_{i=1}^{N^t_l} \sim \mathcal{P}_l \) of \( S_l \). \( x^t_{li} \) and \( y^t_{li} \in \mathcal{Y}_t^i \) denote the images and labels of the \( l \)-th client. \( \{\mathcal{P}_l\}_{l=1}^{L} \) are non-independent and identically distributed (i.e., Non-IID) across local clients. The label space \( \mathcal{Y}_t^i \subset \mathcal{Y}_t^0 \) of \( S_l \) in the \( t \)-th task is composed of \( K^t_l \) new classes (\( K^t_l \leq K^t \)) that belongs to a subset of \( \mathcal{Y}_t^i = \bigcup_{j=1}^{t} 3^j \mathcal{Y}_t^j \). Following ISS methods [11, 36, 53], we consider background shift in the FISS and also annotate \( K^0 = \sum_{i=1}^{t-1} K^i_l \cup \bigcup_{j=1}^{t-1} 3^j \mathcal{Y}_t^{j} \) old categories from \( t - 1 \) old tasks and other foreground categories from future learning tasks as background. After getting global model \( \Theta^{r,t} \) and performing local training on \( T^t_l \), \( S_l \) obtains a updated local model \( \Theta^{r,t}_l \). Then global server \( S_g \) aggregates local models of selected clients as the global model \( \Theta^{r+1,t} \) for the training of next global round.

In the \( t \)-th task, motivated by [10], all local clients \( \{S_l\}_{l=1}^{L} \) are divided into three categories: \( \{S_l\}_{l=1}^{L} = S_o \cup S_c \cup S_n \). Specifically, \( S_o \) is composed of \( L_o \) local clients that have accumulated past experience for previous tasks but cannot collect new training data of the \( t \)-th task; \( S_c \) consisting of \( L_c \) local clients can receive new training data of current task and has learning experience for old classes; \( S_n \) includes \( L_n \) new local clients with unseen novel classes but without past learning experience of old classes. These local clients are randomly determined in each incremental task. New clients \( S_n \) are added randomly at any global round in the FISS, increasing \( L = L_o + L_c + L_n \) gradually as continuous tasks. More importantly, we don’t have prior knowledge about the class distributions \( \{\mathcal{P}_j\}_{j=1}^{K} \) quantity and order of segmentation tasks, when and which local clients receive new classes. In this paper, FISS aims to learn a global model \( \Theta^{R,T} \) to segment new categories continuously while surmounting heterogeneous forgetting on old categories brought by background shift, under the privacy preservation of local clients.

4. The Proposed Model

Figure 2 presents the overview of our model to address the FISS problem. Our FBL model overcomes intra-client heterogeneous forgetting via a forgetting-balanced semantic compensation loss (Section 4.2) and a forgetting-balanced relation consistency loss (Section 4.3), under the guidance of adaptive class-balanced pseudo labeling (Section 4.1) to mine pseudo labels for old classes with background shift. Meanwhile, it addresses inter-client heterogeneous forgetting via a task transition monitor (Section 4.4) to recognize new classes and store old model for relation distillation.

4.1. Adaptive Class-Balanced Pseudo Labeling

For the \( l \)-th local client \( S_l \in \{S_c \cup S_n\} \), the semantic segmentation loss \( L_{SE} \) for a mini-batch \( \{x^t_{li}, y^t_{li}\}_{i=1}^{B} \subset T^t_l \) sampled from the \( t \)-th incremental task is formulated as:

\[
L_{SE} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{HW} D_{CE}(\mathcal{P}_l(x^t_{li}, \Theta^{r,t}), (y^t_{li})) \tag{1}
\]

where \( D_{CE}(. , .) \) denotes the cross-entropy loss. At the \( r \)-th global round, global model \( \Theta^{r,t} \) is transmitted from global server \( S_g \) to \( S_l \). \( \mathcal{P}_l(x^t_{li}, \Theta^{r,t}) \in \mathbb{R}^{1 + K^t + K^t} \) is the probability at the \( j \)-th \((j = 1, \cdots, HW)\) pixel predicted by \( \Theta^{r,t} \), and it predicts background, \( K^0 \) old classes, and \( K^t \) new classes for the \( j \)-th pixel. \( y^t_{li,j} \in \mathcal{Y}_t^j \) is corresponding label at the \( j \)-th pixel. \( B \) represents the batch size.

As aforementioned, in the FISS settings, local client \( S_l \) has no memory to store \( K^0 \) old classes, while back-
ground pixels may belong to $K^o$ old classes, other foreground classes from future tasks or real background (i.e., background shift [1,11]). As a result, it enforces the updating of local model $\Theta^{t-1}_i$ (i.e., Eq. (1)) to suffer from intra-client heterogeneous forgetting among different old classes brought by background shift, after $S_i$ receives the global model $\Theta^{t-1}$ from $S_g$ for local training. To this end, as shown in Figure 2, we develop an adaptive class-balanced pseudo labeling to adaptively mine confident pseudo labels for old classes labeled as background pixels in the $t$-th segmentation task. Different from existing ISS methods [2,11,51] that only utilize a constant probability threshold to select pseudo labels for all classes, our FBL model considers class balance to mine pseudo labels for old classes via introducing class-specific entropy threshold for each old class, which are adaptively determined as continual learning process. These class-balanced pseudo labels of $K^o$ old classes are essential to alleviate heterogeneous forgetting across different old classes within local clients.

In the $t$-th task, as shown in Figure 2, given a sample $\{(x_{i,t}, y_{i,t})\} \subset T_i^t$, we feed it into old global model $\Theta^{t-1}_{\omega}$ of the last task and current local model $\Theta^{t}_i$ to obtain the probabilities $P^{t-1}_i(x_{i,t}, \Theta^{t-1}_{\omega}) \in \mathbb{R}^{H \times W \times (1 + K^o)}$ and $P^{t}_i(x_{i,t}, \Theta^{t}_{i}) \in \mathbb{R}^{H \times W \times (1 + K^f + K^o)}$ respectively. Then pseudo label $\hat{y}_{i,t}^t \in \mathbb{R}^{H \times W}$ of given image $x_{i,t}$ is defined as:

$$
(\hat{y}_{i,t})_j = \begin{cases} 
  k, & \text{if } (y_{i,t})_j \notin \mathcal{Y}^b \text{ and } k = (y_{i,t})_j; \\
  k, & \text{if } (y_{i,t})_j \in \mathcal{Y}^b \text{ and } h_j \leq \gamma_k^t \\
  \text{and } k = \arg \max \limits_{\mathcal{Y}_{i,t}} P_{i}^{t-1}(x_{i,t}, \Theta^{t-1}_i); \\
  0, & \text{otherwise},
\end{cases}
$$

Algorithm 1: Determination of $\{\gamma_k^t\}_{k=1}^{K^o}$ in Eq. (2).

Input: $T_i^t = \{(x_{i,t}, y_{i,t})\}_{i=1}^{N_i}$ and the selection proportion $\rho_i$.

for $i = 1, \ldots, N_i^t$ do

$H_{i}^t = \mathcal{H}(P_i^t(x_{i,t}, \Theta^{t-1}_{i})) \in \mathbb{R}^{H \times W}$;

$L_i^t = \arg \max \limits_{\mathcal{Y}_{i,t}} P_{i}^{t-1}(x_{i,t}, \Theta^{t-1}_i) \in \mathbb{R}^{H \times W}$;

for $k = 1, \ldots, K^o$ do

$H_k^t = H_i^t; L_i^t \in k$;

$M_k^t = [M_k^t; \text{matrix_to_vector}(H_k^t)]$;

for $k = 1, \ldots, K^o$ do

$E_k^t = \text{sort}(M_k^t, \text{order} = \text{ascending})$;

$\gamma_k^t = E_k^t[\text{length}(E_k^t) \cdot \rho_i]$;

end

end

where $(\hat{y}_{i,t})_j$ is pseudo label of the $j$-th pixel from $\hat{y}_{i,t}$. $P_i^{t-1}(x_{i,t}, \Theta^{t-1}_{i})$ is softmax probability of the $j$-th pixel from $P_i^{t-1}(x_{i,t}, \Theta^{t-1}_{i})$. $h_j = \mathcal{H}(P_i^t(x_{i,t}, \Theta^{t-1}_{i}))$ represents entropy of the $j$-th pixel, and $\mathcal{H}(p) = \sum_i p_i \log p_i$ is entropy measure function. $\{\gamma_k^t\}_{k=1}^{K^o}$ denote class-specific entropy threshold to adaptively select class-balanced pseudo labels with high confidence. As shown in Eq. (2), in the $t$-th task $T_i^t$, when the $j$-th pixel belongs to background label space $\mathcal{Y}^b$ (i.e., $(y_{i,t})_j \notin \mathcal{Y}^b$) and the entropy $h_j$ is less than $\gamma_k^t$, pseudo label is determined by $(\hat{y}_{i,t})_j = \arg \max \limits_{\mathcal{Y}_{i,t}} P_{i}^{t-1}(x_{i,t}, \Theta^{t-1}_i)$. If the $j$-th pixel is not labeled as background (i.e., $(y_{i,t})_j \notin \mathcal{Y}^b$), we consider pseudo label as new foreground classes: $(\hat{y}_{i,t})_j = (y_{i,t})_j$. Otherwise, $(\hat{y}_{i,t})_j = 0$ denotes real background for the $j$-th pixel of $\hat{y}_{i,t}$.

The determination of $\{\gamma_k^t\}_{k=1}^{K^o}$ is summarized in Algorithm 1. After computing entropy $\mathcal{H}(\hat{y}_{i,t})_{i=1}^{N_i}$ for all samples in the $t$-th task $T_i^t$, we sort the entropy of all pixels predicted as the $k$-th class. $\gamma_k^t$ is determined via the entropy ranked
at \(|\text{length}(\mathbf{E}_b^t) \cdot \rho|\) of \(\mathbf{E}_b^t\), where \(\rho\) is selection proportion for all old classes. The value of \(\rho\) is initialized as 20\%, and adds 2\% for each epoch empirically as training process. We set the maximum selection proportion \(\rho\) as 50\%. Given a mini-batch \(\{\mathbf{x}_{ij}^t, \mathbf{y}_{ij}^t\}_{i=1}^{B} \subset \mathcal{T}_i^t\), we generate class-balanced pseudo labels \(\{\mathbf{x}_{ij}^t, \mathbf{y}_{ij}^t\}_{i=1}^{B} \subset \mathcal{T}_i^t\) adaptively via considering class-balanced selection proportion \(\rho\) in Eq. (2) for all old classes. These confiding pseudo labels provide strong guidance for the forgetting-balanced semantic compensation loss (Section 4.2) and forgetting-balanced relation consistency loss (Section 4.3) to surmount intra-client heterogeneous forgetting among different old classes.

### 4.2. Forgetting-Balanced Semantic Compensation

To address heterogeneous forgetting speeds of different old classes within local client \(\mathcal{S}_t \in \mathcal{S}_o \cup \mathcal{S}_n\), we propose a forgetting-balanced semantic compensation loss \(\mathcal{L}_{FS}\), as shown in Figure 2. It considers balanced gradient propagation between different old tasks for intra-client heterogeneous forgetting compensation. Specifically, the loss \(\mathcal{L}_{FS}\) employs gradient propagation means of different old tasks to measure the forgetting heterogeneity of old classes, and then reweights segmentation loss \(\mathcal{L}_{SE}\) in Eq. (1) to normalize heterogeneous forgetting speeds brought by background shift. For a given sample \(\{\mathbf{x}_{ij}^t, \mathbf{y}_{ij}^t\} \subset \mathcal{T}_i^t\) with generated pseudo label, we first obtain its probability \(\mathbf{P}_l^t(x_{ij}^t, \Theta_l^{t-1})\) predicted via local model \(\Theta_l^{t-1}\). Motivated by [45], we then formulate gradient scalar \(\Gamma_{ij}^t\) of the \(j\)-th pixel with respect to the \(k\)-th output neuron \(N_k^t\) of pixel classifier in \(\Theta_l^{t-1}\) as follows:

\[
\Gamma_{ij}^t = \frac{\partial D_{CE}(\mathbf{P}_l^t(x_{ij}^t, \Theta_l^{t-1}), (\hat{\mathbf{y}}_{ij}^t)_{ij})}{\partial N_k^t} = \mathbf{P}_l^t(x_{ij}^t, \Theta_l^{t-1}) k_j - 1, \tag{3}
\]

where \(\mathbf{P}_l^t(x_{ij}^t, \Theta_l^{t-1})\) is probability of the \(k\)-th class at the \(j\)-th pixel of \(x_{ij}^t\), and \(k = \{\hat{\mathbf{y}}_{ij}^t\}_{ij}\) denotes pseudo label of the \(j\)-th pixel in \(x_{ij}^t\). Considering that intra-client heterogeneous forgetting of old classes changes dynamically as continual learning tasks, we expect gradient scalar \(\Gamma_{ij}^t\) of old classes to be adaptive in the FISS, and reformulate Eq. (3) as:

\[
\Gamma_{ij}^t = \frac{|\hat{\Gamma}_{ij}^t|^{\nu + \kappa_j} \cdot \mathbf{I}((\hat{\mathbf{y}}_{ij}^t)_{ij} \in \mathcal{Y}_{ij}^t)}, |\hat{\Gamma}_{ij}^t|^{\nu + \kappa_j} + |\hat{\Gamma}_{ij}^t| \cdot |\hat{\mathbf{y}}_{ij}^t| \in \mathcal{Y}_{ij}^t\}, \tag{4}
\]

where \(\mathcal{Y}_{ij}^t\) is background label space of the \(t\)-th local client \(\mathcal{S}_t\). When pseudo label \((\hat{\mathbf{y}}_{ij}^t)_{ij}\) of the \(j\)-th pixel in \(x_{ij}^t\) belongs to old classes from previous \(t-1\) tasks, \(\Gamma_{ij}^t = |\Gamma_{ij}^t|^{(K_j + K_t)}\); otherwise, \(\Gamma_{ij}^t = |\Gamma_{ij}^t|\) for new classes and background.

As a result, given mini-batch samples \(\{\mathbf{x}_{ij}^t, \mathbf{y}_{ij}^t\}_{i=1}^{B} \subset \mathcal{T}_i^t\) in the \(t\)-th segmentation task, we denote gradient propagation means \(\Gamma_b\) and \(\Gamma_n\) for the background and foreground classes learned from the \(\eta\)-th task (1 \(\leq\) \(\eta\) \(\leq\) \(t\)) as follows:

\[
\Gamma_b = \frac{1}{Z_b^t} \sum_{i=1}^{B} \sum_{j=1}^{HW} \hat{\Gamma}_{ij}^t, \quad \Gamma_n = \frac{1}{Z_n^t} \sum_{i=1}^{B} \sum_{j=1}^{HW} \hat{\Gamma}_{ij}^t \tag{5}
\]

where the quantity of pixels belonging to background and the \(\eta\)-th task are denoted as \(Z_b = \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathbf{I}(\mathbf{y}_{ij}^t) \in \mathcal{Y}_b^{\eta}\) and \(Z_n = \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathbf{I}(\mathbf{y}_{ij}^t) \in \mathcal{Y}_n^{\eta}\). The gradient propagation means \(\Gamma_b\) and \(\{\Gamma_{ij}^t\}_{\eta=1}^{t}\) in Eq. (5) reflect gradient-imbalanced propagation between old and new classes. Thus, these gradient means can effectively measure intra-client forgetting heterogeneity of different old classes, and evaluate updating speeds of new classes and background to some extent. Under the guidance of pseudo labels \(\mathbf{y}_{ij}^t\), we employ \(\{\Gamma_{ij}^t\}_{\eta=1}^{t}\) and \(\Gamma_b\) to reweight semantic segmentation loss \(\mathcal{L}_{SE}\), and formulate the proposed forgetting-balanced semantic compensation loss \(\mathcal{L}_{FS}\) as follows:

\[
\mathcal{L}_{FS} = \frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{HW} \frac{\Gamma_{ij}^t}{\bar{\Gamma} \cdot \mathbf{I}((\mathbf{y}_{ij}^t) \in \mathcal{Y}_b^t)} \cdot D_{CE}(\mathbf{P}_l^t(x_{ij}^t, \Theta_l^{t-1})_j, (\hat{\mathbf{y}}_{ij}^t)_j), \tag{6}
\]

where \(\bar{\Gamma} = \sum_{\eta=1}^{t} \mathbf{I}(\mathbf{y}_{ij}^t) \in \mathcal{Y}_n^t + \Gamma_b \cdot \mathbf{I}(\mathbf{y}_{ij}^t) \in \mathcal{Y}_b^t\) denotes different normalization weights for background, old and new classes. \(\mathcal{L}_{FS}\) can address intra-client heterogeneous forgetting of different old classes via reweighting segmentation loss \(\mathcal{L}_{SE}\) to achieve class-balanced gradient propagation.

### 4.3. Forgetting-Balanced Relation Consistency

The intrinsic relations between old and new classes are immutable in purely semantic space, independent of background shift [1, 11] and availability of training data of old classes. In light of this, consistent semantic relations between old model \(\Theta_l^{t-1}\) and current local model \(\Theta_l^{t-1}\) plays an important role in tackling intra-client heterogeneous forgetting on old classes. However, most existing ISS methods [1, 27, 36] only consider underlying relationships among old classes via performing knowledge distillation [18] on an individual sample, which can be severely affected by noisy predictions on old classes brought by background shift. In addition, forgetting heterogeneity of old classes within local clients enforces most ISS methods [1, 11] to suffer from heterogeneous inter-class relations distillation, thus aggravating imbalanced gradient propagation across incremental tasks.

To this end, we propose a forgetting-balanced relation consistency loss \(\mathcal{L}_{FR}\) to tackle intra-client heterogeneous forgetting via compensating heterogeneous relation distillation. Specifically, we propose relationship prototype of each class instead of an individual sample to better characterize underlying relations between old and new classes, and consider gradient means in Eq. (5) to balance heterogeneous relation distillation. As shown in Figure 2, given \(\{\mathbf{x}_{ij}^t, \mathbf{y}_{ij}^t\}_{i=1}^{B} \subset \mathcal{T}_i^t\), we feed it into old model \(\Theta_l^{t-1}\) and local model \(\Theta_l^{t-1}\) of \(\mathcal{S}_t\) to obtain probabilities \(\mathbf{P}_l^{t-1}(x_{ij}^t, \Theta_l^{t-1}) \in \mathbb{R}^{t \times (1+K^o)}\) and \(\mathbf{P}_l^t(x_{ij}^t, \Theta_l^{t-1}) \in \mathbb{R}^{t \times (1+K^o)}\). Then we substitute the first \(1 + K^o\) channel dimensions of one-hot pseudo label \(\mathbf{Y}_{ij}^t \in \mathbb{R}^{t \times (1+K^o)}\) (\(\mathbf{Y}_{ij}^t\) is one-hot encoding of \(\mathbf{y}_{ij}^t\) with \(\mathbf{P}_l^{t-1}(x_{ij}^t, \Theta_l^{t-1})\), and abbreviate this variant as
relationship label $\mathbf{Y}_t^k = \{x_t^k | \Theta_t^{-1}\} \in \mathbb{R}^{K \times W \times (1 + K^o + K^t)}$ indicating underlying relations among old and new categories.

For the $k$-th class, the relationship prototype $\mathbf{P}_k^t$ and its label $\mathbf{Y}_k^t$ are written as:

$$
\mathbf{P}_k^t = \frac{1}{Z_k} \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathbf{P}_i^t(x_i^k, \Theta_t^{-1})_j \cdot \mathbb{I}(x_i^k)_j = k,
$$

$$
\mathbf{Y}_k^t = \frac{1}{Z_k} \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathbf{Y}_i^t(x_i^k, \Theta_t^{-1}) \cdot \mathbb{I}(x_i^k)_j = k,
$$

where $Z_k = \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathbb{I}(x_i^k)_j = k$ is pixel number of the $k$-th class. The class-wise gradient mean $\Gamma_k$ for the $k$-th class is formulated as $\Gamma_k = \frac{1}{Z_k} \sum_{i=1}^{B} \sum_{j=1}^{HW} \mathbf{P}_i^t(x_i^k, \Theta_t^{-1})_j \cdot \mathbb{I}(x_i^k)_j = k$, which is then used to reweigh heterogeneous distribution gains. As a result, the forgetting-balanced relation consistency loss $\mathcal{L}_{FR}$ is concretely written as follows:

$$
\mathcal{L}_{FR} = \frac{1}{K^o + K^t} \sum_{k=1}^{K^o + K^t} \Gamma_k \cdot \mathcal{D}_{KL}(\mathbf{P}_k^t, \mathbf{Y}_k^t),
$$

where $\mathcal{D}_{KL}(\|\cdot\|)$ is Kullback-Leibler divergence. $\Gamma_{cls} = \sum_{k=1}^{K} \Gamma_k \cdot \mathbb{I}(x_i^k)_j = k$ denotes gradient normalization mean.

In summary, the major objective of the $t$-th local client $S_t$ to learn the $t$-th segmentation task $\mathcal{T}^t_t$ is expressed as:

$$
\mathcal{L}_{obj} = \mathcal{L}_{FS} + \lambda_1 \mathcal{L}_{FR} + \lambda_2 \mathcal{L}_{POD},
$$

where $\lambda_1, \lambda_2$ are trade-off parameters, and $\mathcal{L}_{POD}$ denotes the local POD loss proposed in PLOP [11] to perform feature distillation. When $t \geq 2$, we set $\lambda_1 = 0.5$ and $\lambda_2 = 0.3$ in Eq. (10) to train local model $\Theta_t^{-1}$; otherwise, we utilize $\mathcal{L}_{SE}$ in Eq. (1) to optimize $\Theta_t^{-1}$. To learn the $t$-th segmentation task $\mathcal{T}^t_t$, local clients belonging to $S_c$ and $S_n$ share the same objective function (i.e., Eq. (10)).

### 4.4 Task Transition Monitor

When local clients segment new classes consecutively via Eq. (10), global server $S_g$ requires to automatically identify when and which local clients collect new classes, and then store the latest old global model $\Theta_t$ to perform $\mathcal{L}_{FR}$ and $\mathcal{L}_{POD}$. As a result, the accurate selection of the latest old model $\Theta_t^{-1}$ is essential to address inter-client heterogeneous forgetting across different local clients brought by Non-IID class distributions, when new foreground classes arrive. However, considering privacy preservation [54, 55], we don’t have human prior about when to obtain new classes in local clients under the FISS settings. To address this challenge, a naive method is to detect whether the labels of current training data have been observed before. Nevertheless, the Non-IID distributions across local clients make it impossible to identify whether the collected data belongs to old classes seen by other clients or new categories. Thus, inspired by [10, 14], we design a task transition monitor to automatically recognize when and which local clients collect new categories. At the $r$-th round, when $S_t$ receives global model $\Theta_r^{-t}$, it evaluates the average entropy $\mathcal{I}^{r,t}_t$ on $\mathcal{T}^r_t$:

$$
\mathcal{I}^{r,t}_t = \frac{1}{N^t} \sum_{i=1}^{N^t} \sum_{j=1}^{HW} \mathcal{H}(P_i^t(x_i^t, \Theta_r^{-t})_j),
$$

where $\mathcal{H}(P_i^t(x_i^t, \Theta_r^{-t})_j) \in \mathbb{R}^{H \times W}$ is the entropy map of $x_i^t$, and $\mathcal{H}(P_i^t(x_i^t, \Theta_r^{-t})_j)$ is entropy scalar of the $j$-th pixel. $\mathcal{H}(p) = \sum_p p \log p$, is entropy measure function. We consider local clients are collecting new classes, if there is a sudden rise for averaged entropy $\mathcal{I}^{r,t}_t: \mathcal{I}^{r,t}_t - \mathcal{I}^{r-1,t}_t \geq \tau$. We then update $t$ via $t \leftarrow t + 1$, and automatically store the latest global model $\Theta^{r-1}_t$ at the $(r-1)$-th global round as old model $\Theta^{r-1}_t$ to optimize local model $\Theta^{r,t}_t$ via $\mathcal{L}_{obj}$ in Eq. (10). We set $\tau = 0.6$ empirically in this paper. The automatic selection of old model $\Theta^{r-1}_t$ from global aspect is...
essential to tackle inter-client heterogeneous forgetting via considering Non-IID distributions across local clients.

### 4.5. Optimization Procedure

At the beginning of each global round in each incremental task, all local clients employ Eq. (11) to calculate the average entropy of local data, and then some of local clients are randomly selected by global server $S_g$ to conduct local training at each round. After these chosen clients utilize task transition monitor to accurately recognize new classes, they automatically store the global model learned at the last global round as the old model $\Theta^{r-1}$ to generate confident pseudo labels for old classes via Eq. (2), and optimize local model $\Theta^{r,t}_i$ via $L_{obj}$ in Eq. (10). Finally, the updated local models $\Theta^{r,t}_i$ of selected local clients are aggregated as $\Theta^{r+1,t}$ by $S_g$ for the next round training. The supplementary material provides optimization procedure of our FBL model.

### 5. Experiments

#### 5.1. Implementation Details

We utilize two benchmark datasets: Pascal-VOC 2012 [12] and ADE20k [59] under various experimental settings to analyze effectiveness of our FBL model. For fair comparisons with baseline ISS methods [1,11,27,36,53] under the FISS settings, we follow them to set exactly the same incremental tasks and class order, while using the identical segmentation backbone (*i.e.*, Deeplab-v3 [4] with ResNet-101 [17] pretrained on ImageNet dataset [7]). As claimed in [1, 11, 53], background pixels in the current task may belong to old classes or new classes from future tasks (*i.e.*, background has some overlap with new foreground classes in the future tasks). In the FISS, we consider more challenging settings by assigning more incremental segmentation tasks with overlapped foregrounds. On Pascal-VOC 2012 [12], 15-1, 4-4, and 8-2 settings with overlapped foregrounds respectively consist in 15 classes followed by 1 classes 5 times ($T = 6$), learning 4 classes followed by 4 classes 4 times ($T = 5$), and 8 classes followed by 2 classes 6 times ($T = 7$). Likewise, on ADE20k [59], 100-10 setting with overlapped foregrounds means 100 classes followed by 10 classes 5 times ($T = 6$).

We employ SGD optimizer with initial learning rate as $1.0 \times 10^{-2}$ to train the first base task and $1.0 \times 10^{-3}$ to learn incremental tasks. Considering the limitation of GPU overhead, we set initial local clients as 10, and add 4 new local clients for each task. We choose 4 local clients randomly to perform local training with 6 epochs for VOC [12] and 12 epochs for ADE20k [59]. On VOC dataset [12], we randomly select 40% images for each client in each segmentation task under 15-1 setting; otherwise, we randomly sample 50% classes from current label space $Y^t$, and assign 60% samples from these classes to selected local clients under the 4-4 and 8-2 settings. For the 100-10 setting in ADE20k [59], we randomly choose 70% classes from $Y^t$, and distribute them to selected clients. Following ISS methods [1,11,27,36,53], we employ mean Intersection over Union (mIoU) as metric, and evaluate mIoU of all classes after learning the last segmentation task (*i.e.*, $t = T$). This metric evaluates the effectiveness to address heterogeneous forgetting and the ability to segment new classes continually.

### 5.2. Comparison Performance

Experiments on Pascal-VOC 2012 [12] and ADE20k [59] are introduced to analyze superiority of our model under various settings of FISS, as shown in Tables 1~4. Our model achieves large improvements over existing ISS methods [1,11,27,36,53] about 1.5% ~ 51.4% mIoU under various FISS settings. It illustrates the effectiveness of our model against other ISS methods to learn a global continual
To analyze effectiveness of each module in our model, Table 5 presents ablation experiments under various FISS settings. Ours-w/oAPL, Ours-w/oFSC and Ours-w/oFRC indicate the results of our model without adaptive class-balanced pseudo labeling (denoted as APL), forgetting-balanced semantic compensation loss $L_{FS}$ (denoted as FSC) and forgetting-balanced relation consistency loss $L_{FR}$ (denoted as FRC), where Ours-w/oAPL uses constant probability threshold for all old classes to replace adaptive class-specific entropy threshold. When compared with Ours, all ablation variants severely degrade $3.9\% \sim 11.6\%$ mIoU. It verifies importance of all modules to address the heterogeneous forgetting. The proposed APL module can effectively tackle background shift via confident pseudo labels, and some confident pseudo labels are visualized in Figure 4.

### 5.4. Analysis of Task-Wise Comparisons

As presented in Table 6, we introduce task-wise comparison results to analyze the effectiveness of our model to address FISS settings. Our model outperforms baseline ISS methods [1, 11, 27, 36, 53] for most task-wise comparisons under the overlapped 4-4 setting. The proposed FBL model encourages local clients to learn a global incremental segmentation model cooperatively under privacy preservation. Comparisons in Table 6 show large mIoU improvements of our model to address the FISS problem over other ISS methods. When segmenting new foreground classes consecutively, our model can effectively tackle intra-client and inter-client heterogeneous forgetting on different old classes.

### 6. Conclusion

In this work, we propose a Federated Incremental Semantic Segmentation (FISS) problem, and develop a novel Forgetting-Balanced Learning (FBL) model to address intra-client and inter-client heterogeneous forgetting on old classes. To tackle intra-client heterogeneous forgetting, we design a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss, under the guidance of adaptive class-balanced pseudo labeling. Meanwhile, we propose a task transition monitor to address inter-client heterogeneous forgetting. It can automatically recognize new classes and store the latest old global model for distillation. Comparison results demonstrate the superiority of our model to tackle the FISS problem. In the future, we will consider using only few samples of new classes to address intra-client and inter-client forgetting.
References

[8] Jiahua Dong, Yang Cong, Gan Sun, Zhen Fang, and Zhengming Ding. Where and how to transfer: Knowledge aggregation-induced transferability perception for unsupervised domain adaptation. IEEE Transactions on Pattern Analysis and Machine Intelligence, pages 1–1, 2021. 1
[14] Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In NeurIPS, 2022. 6
[26] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In ICLR, 2021. 2


[57] Yonggang Zhang, Xinguo Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Causaladv: Adversarial robustness through the lens of causality. 2022.
