NEURAL INFORMATION PROCESSING SYSTEMS

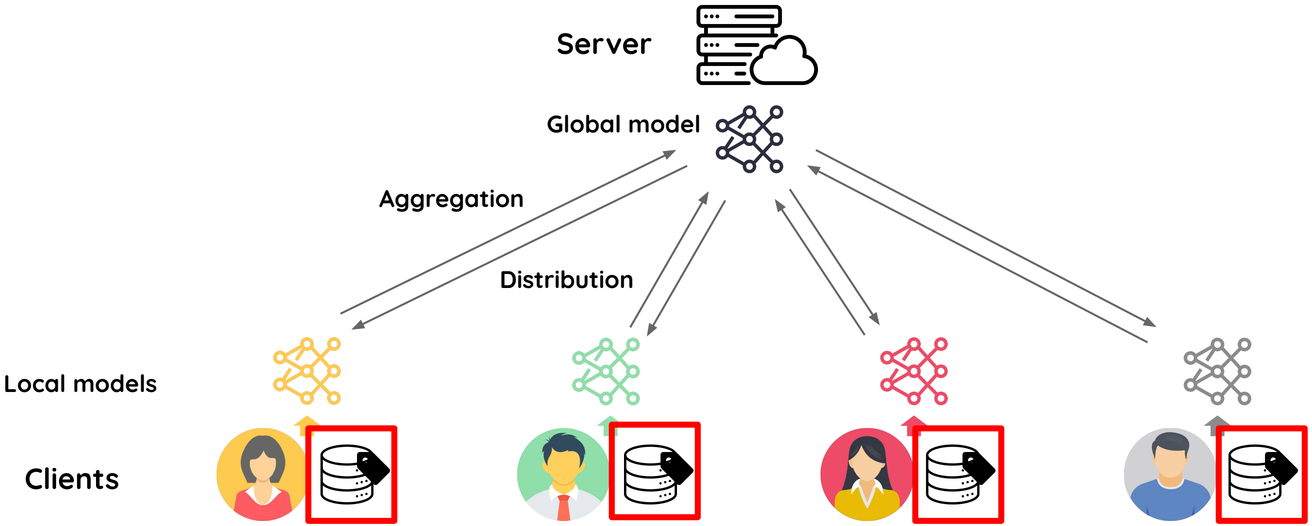
# (FL)$^2$: Overcoming Few Labels in Federated Semi-Supervised Learning




Seungjoo Lee | Thanh-Long V. Le | Jaemin Shin | Sung-Ju Lee

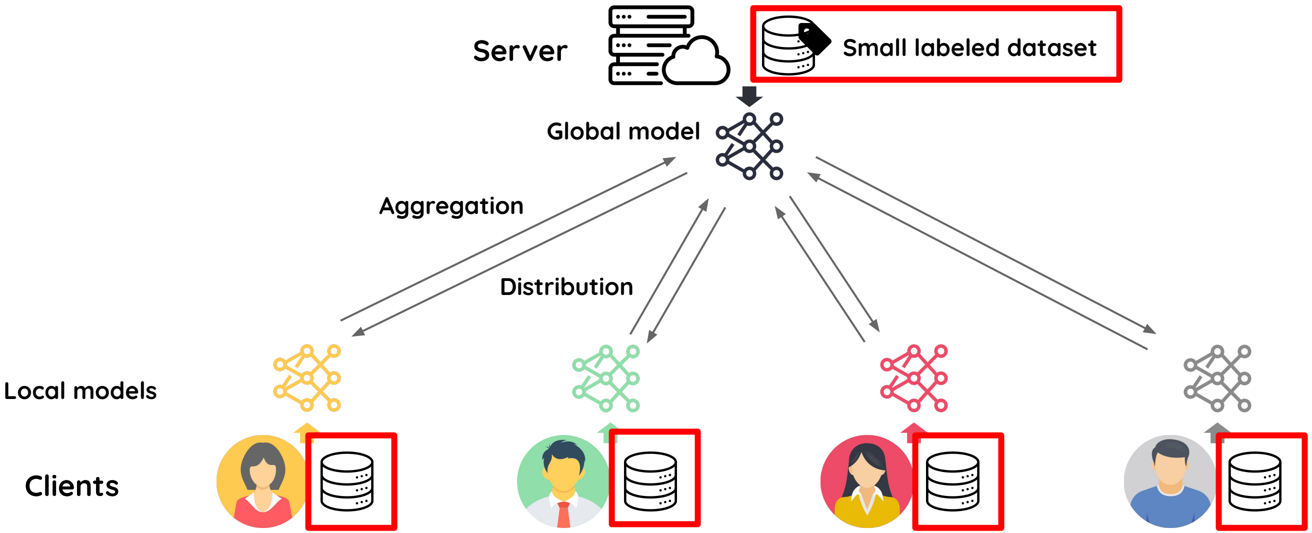
KAIST

# Unrealistic Assumption in Federated Learning (FL)




- FL collaboratively trains accurate global model while keeping clients' **privacy-sensitive data unshared**
- Most FL studies assume that clients have **labeled data**
  - **Unrealistic assumption** in practical scenarios
    - Clients are **reluctant** or **lack of motivation** to label data
    - Certain data types require **domain expertise** (e.g., medical data, sensor data)

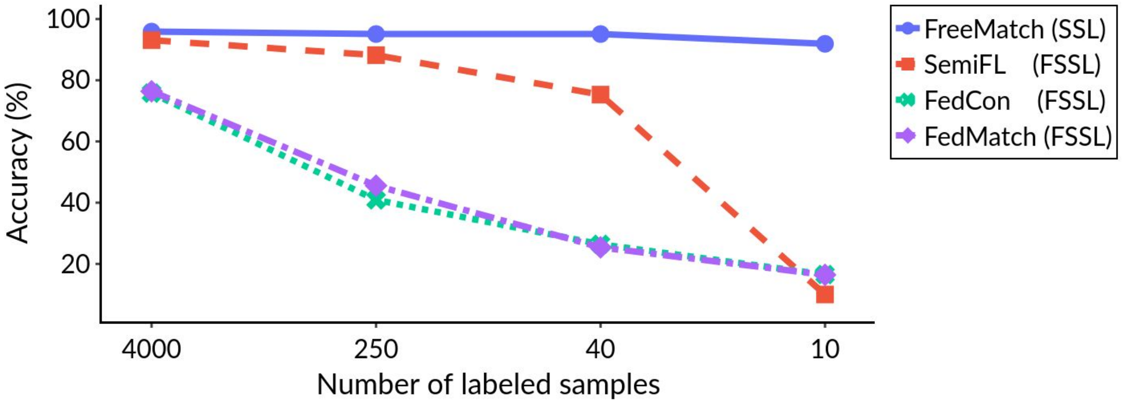
# Federated Semi-Supervised Learning (FSSL)




- ***Labels-at-server*** scenario
  - Server owns **small labeled dataset**
  - Clients remain **unlabeled**

# Limitation of Previous FSSL Studies

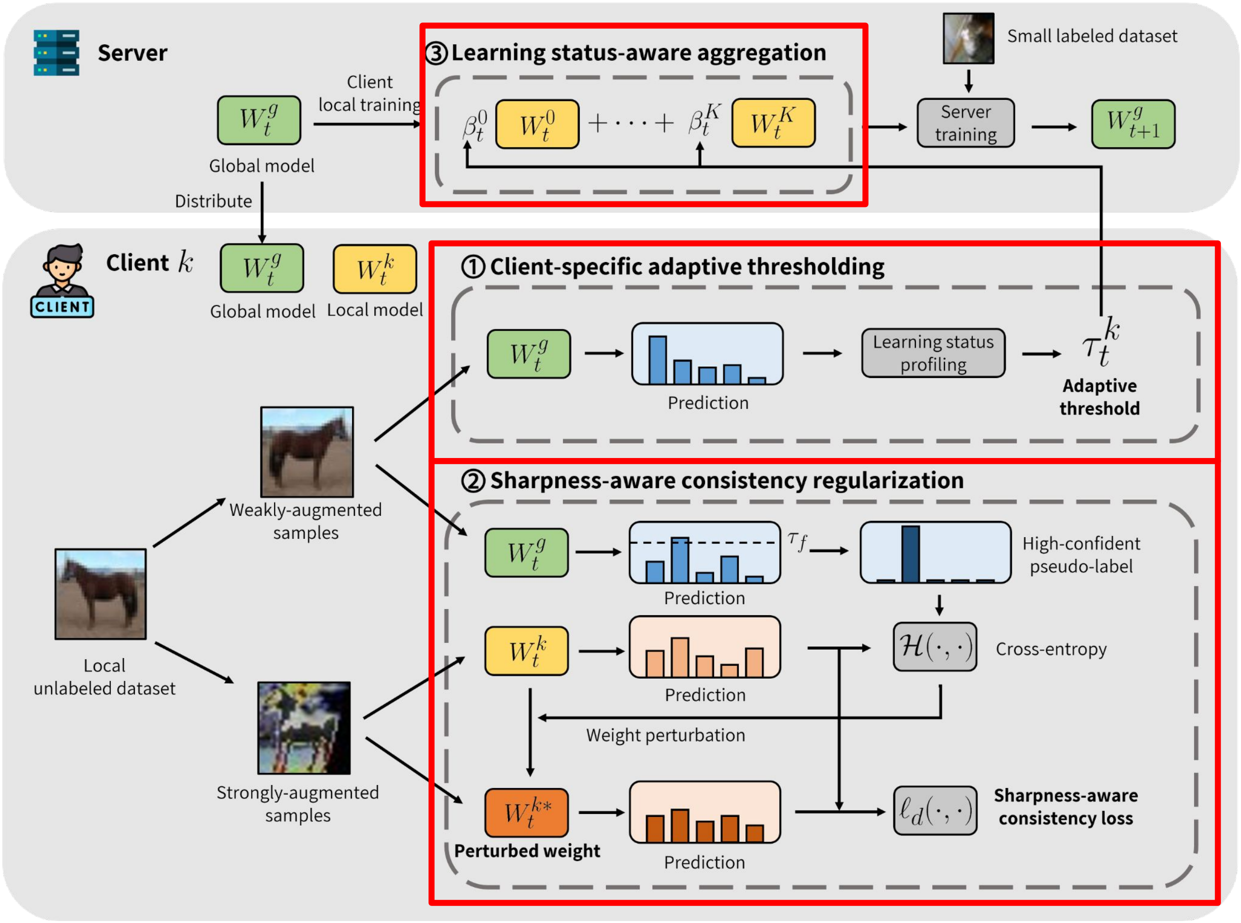
**Test accuracy on CIFAR10 dataset**



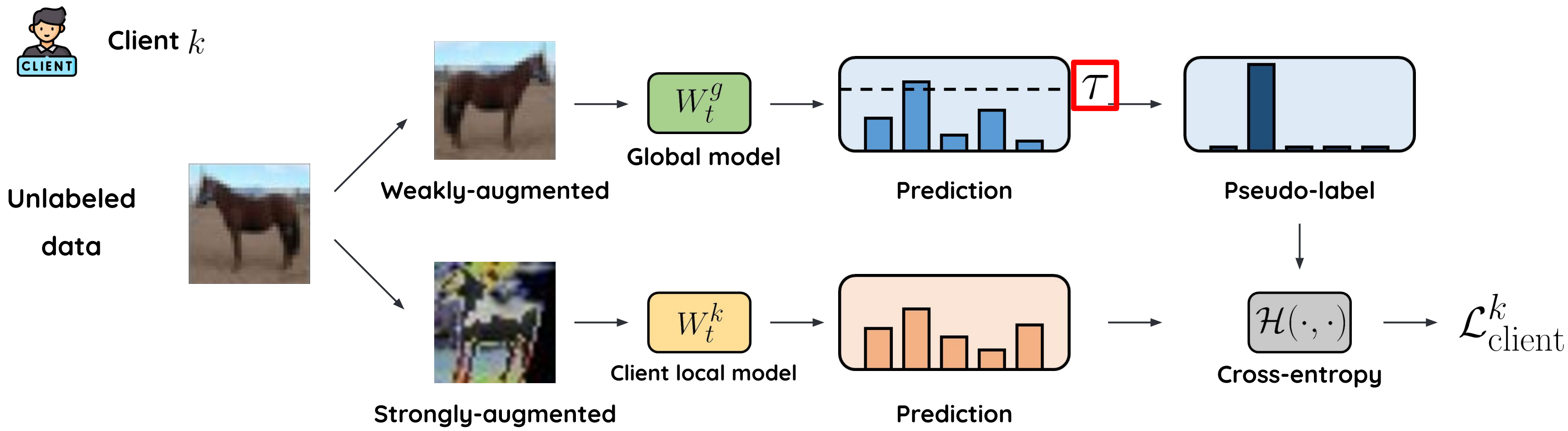

- **Large performance gap** between SSL and FSSL
  - Especially when the given labeled data is **scarce**
- **Confirmation bias** is the primary cause*
  - Overfits to **easy-to-learn samples** or **incorrectly pseudo-labeled** data

*Khanh-Binh Nguyen and Joon-Sung Yang. Boosting semi-supervised learning by bridging high and low-confidence predictions. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1028–1038, 2023.

# $(FL)^2$ : Few-Labels Federated Semi-Supervised Learning

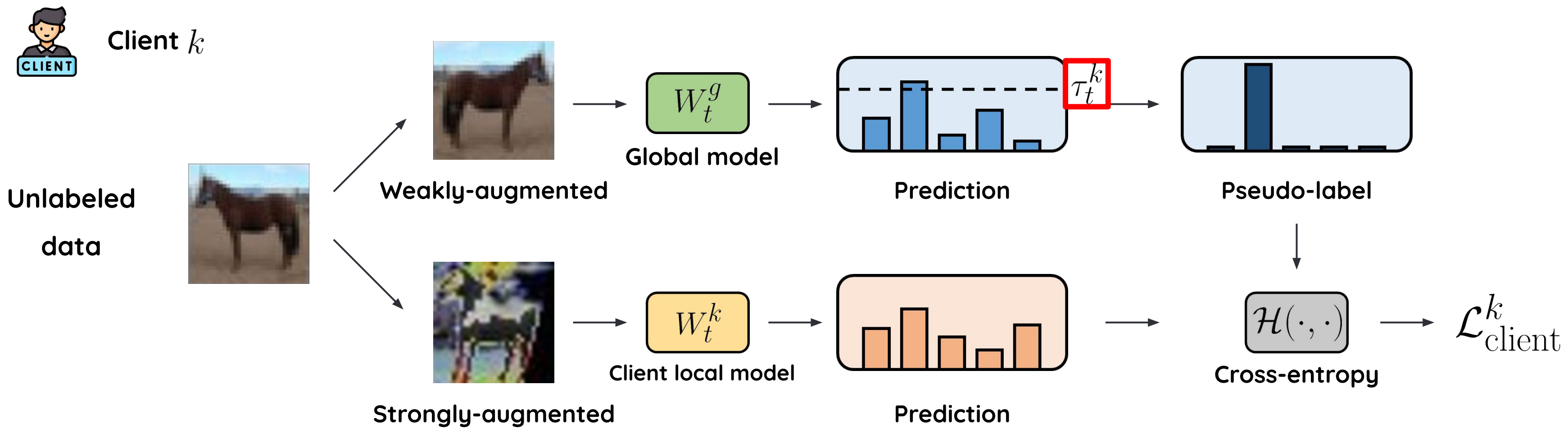


# (FL)$^2$ : Client-specific Adaptive Thresholding




- Existing FSSL approaches use **fixed, high threshold** for pseudo-labeling $(\tau = 0.95)$
  - Only **small portion** of unlabeled data is utilized at the beginning of training
    - ⇨ Prone to **overfitting**
    - ⇨ Increase **confirmation bias**

# $(FL)^2$ : Client-specific Adaptive Thresholding




- Instead, we propose **client-specific adaptive threshold** ($\tau_t^k$) based on **client's learning progress**
  - **Low threshold** at the beginning of training
    - ⇨ Utilize more unlabeled data; Prevent overfitting
  - **High threshold** at later stages
    - ⇨ Filter out wrong pseudo-labels
  - **Different threshold** for each clients according to their learning progress

# $(FL)^2$ : Learning Status-Aware Aggregation

- Existing FSSL approaches use **uniform aggregation weight**

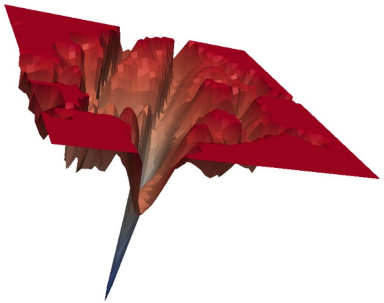
$$W^g_{t+1} = \sum_{k=1}^{K} \beta^k W^k_t, \quad \text{where } \beta^k = \frac{1}{K}$$

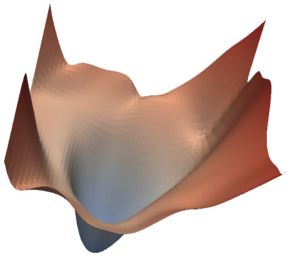
- We propose **learning status-aware aggregation**
  - Client with **low** learning status (low $\tau^k_t$ )
    - ⇨ **Increase** $\beta^k$ so that local learning is **better reflected** in the global model
  - Client with **high** learning status (high $\tau^k_t$ )
    - ⇨ **Decrease** $\beta^k$, because data is already reflected in the global model

# $(FL)^2$ : Sharpness-Aware Consistency Regularization

- SAM (Sharpness-Aware Minimization) shows **strong generalization capabilities** across various tasks
- Key idea: **'Flat' local minima** is good for generalization

Normal loss landscape





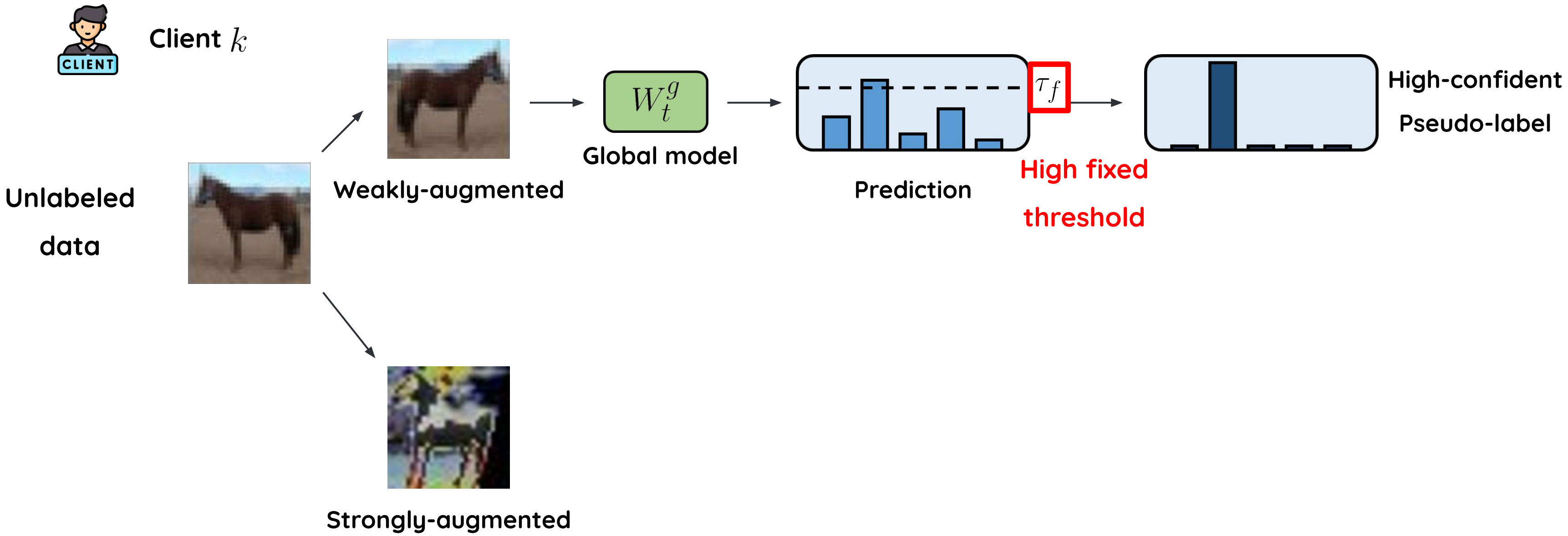
SAM loss landscape

- However, **naïve application** of SAM to FSSL is **suboptimal**

Image credit: Foret, P., Kleiner, A., Mobahi, H., & Neyshabur, B. (2020). Sharpness-aware minimization for efficiently improving generalization. arXiv preprint arXiv:2010.01412.

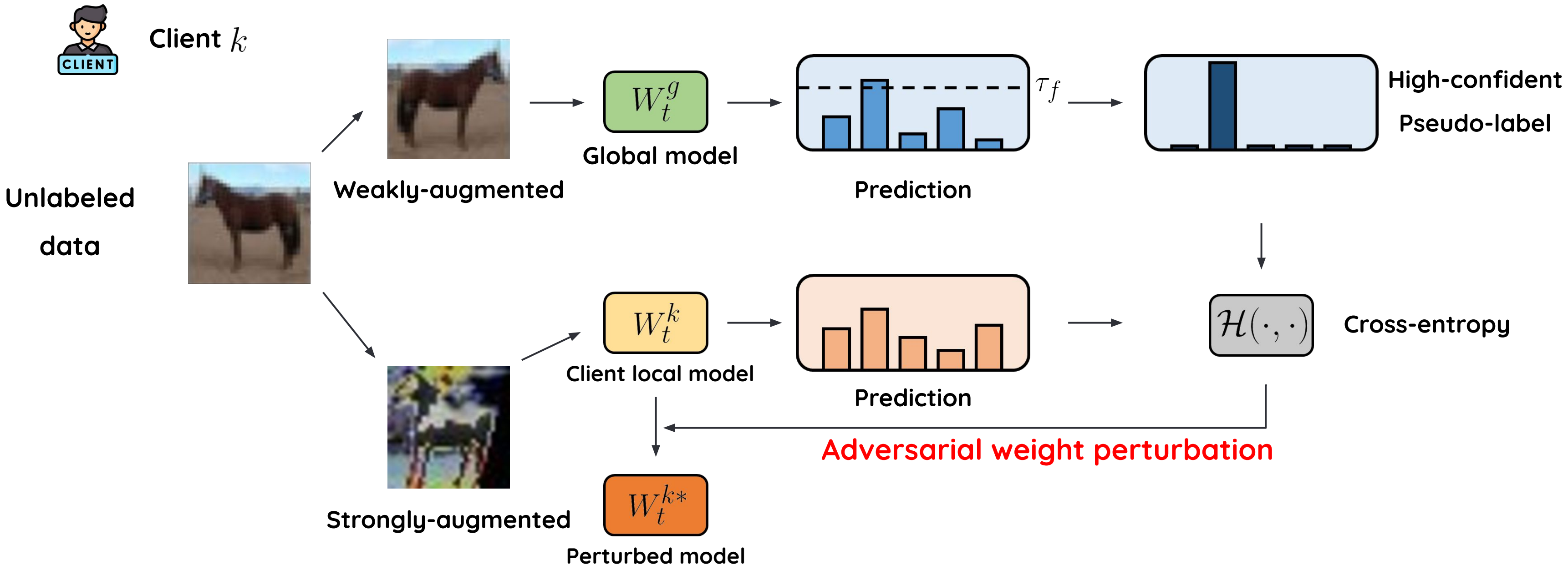
# (FL)$^2$ : Sharpness-Aware Consistency Regularization

- SAM generalizes **both correctly and incorrectly** pseudo-labeled data
- Selecting data that are **highly likely to be correct**



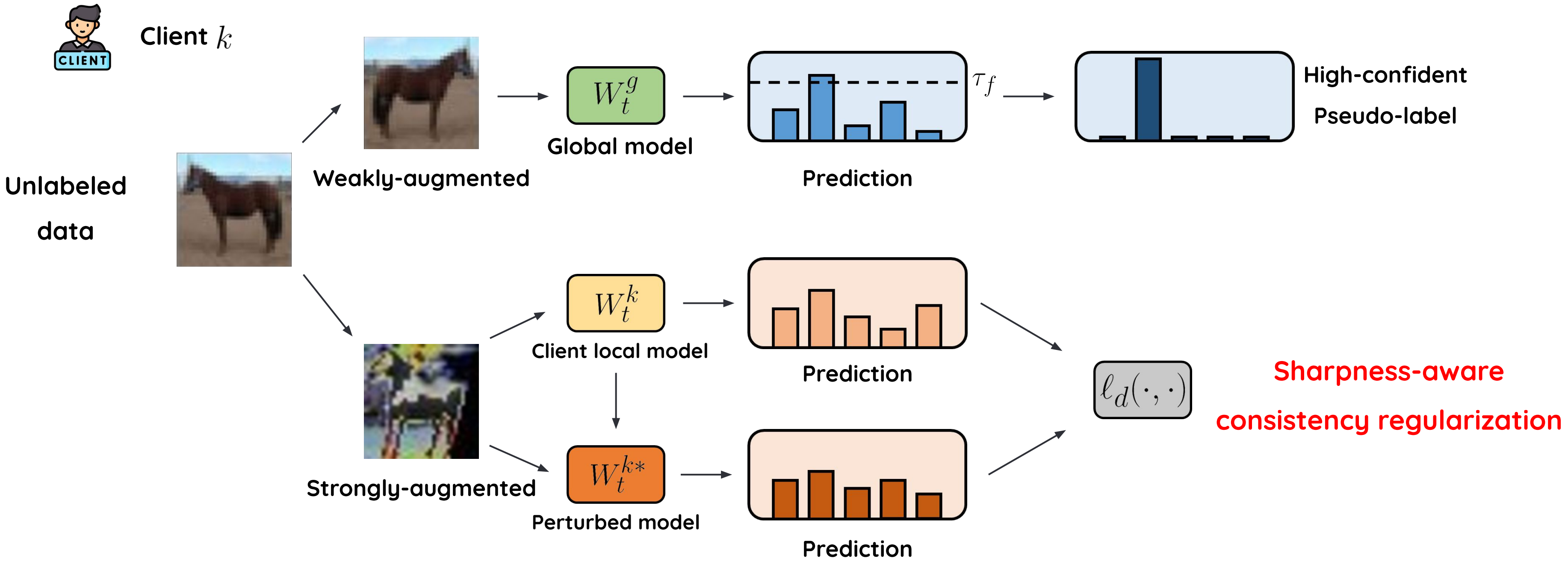
# $(FL)^2$ : Sharpness-Aware Consistency Regularization

- **Adversarial weight perturbation** with the pseudo-labeling loss



# $(FL)^2$ : Sharpness-Aware Consistency Regularization

- SAM objective is **less effective in FSSL** compared to other tasks
- **Consistency regularization with perturbed model** instead of standard SAM objective



# Performance Comparison with Existing FSSL Algorithms

- Average accuracy(%) and standard deviation across three different seeds
- **Bold**: best result / <u>underline</u>: second-best result

| Dataset | | CIFAR10 | | SVHN | | CIFAR100 | |
|---|---|---|---|---|---|---|---|
| # of labeled data samples ($N_L$) | | 10 | 40 | 40 | 250 | 100 | 400 |
| Unbalanced Non-IID, Dir(0.1) | FedMatch | 16.0(2.3) | <u>25.6(2.2)</u> | <u>20.7(2.7)</u> | 70.1(2.2) | 6.3(0.3) | 10.0(1.8) |
| | FedCon | <u>16.6(2.1)</u> | 25.4(2.3) | 20.5(1.4) | 73.1(2.0) | 4.0(0.4) | 8.2(0.6) |
| | SemiFL | 10.0(0.0) | 19.9(7.5) | 18.0(2.6) | <u>82.3(1.8)</u> | <u>9.8(2.4)</u> | <u>13.5(5.0)</u> |
| | *(FL)*$^2$ | **19.2(5.7)** | **36.4(1.4)** | **21.5(3.3)** | **88.0(1.0)** | **10.4(1.3)** | **23.5(1.2)** |
| Unbalanced Non-IID, Dir(0.3) | FedMatch | 15.3(1.3) | 25.2(3.5) | 22.3(0.7) | <u>72.3(3.0)</u> | 5.5(1.5) | 9.8(1.1) |
| | FedCon | <u>16.9(2.4)</u> | 26.5(2.1) | 21.6(1.7) | 68.7(2.7) | 5.8(0.6) | 13.3(0.9) |
| | SemiFL | 10.0(0.0) | <u>38.0(2.7)</u> | <u>26.3(2.5)</u> | 42.7(40.1) | **12.4(1.2)** | <u>18.9(9.7)</u> |
| | *(FL)*$^2$ | **24.3(4.5)** | **43.5(7.5)** | **31.0(4.2)** | **92.6(0.5)** | <u>12.1(1.1)</u> | **25.4(1.0)** |
| Balanced IID | FedMatch | 16.2(1.9) | 25.4(2.8) | 18.4(4.7) | 66.2(0.8) | 6.4(0.6) | 10.0(1.7) |
| | FedCon | <u>16.7(2.0)</u> | 23.3(6.2) | 20.3(1.0) | <u>71.6(1.5)</u> | 5.7(0.6) | 12.4(1.6) |
| | SemiFL | 10.0(0.0) | <u>75.3(2.8)</u> | <u>53.4(13.3)</u> | 43.3(41.0) | <u>13.9(3.3)</u> | <u>27.9(6.1)</u> |
| | *(FL)*$^2$ | **38.9(11.1)** | **81.5(7.4)** | **75.3(2.4)** | **94.6(1.1)** | **14.4(2.3)** | **28.1(2.2)** |

- (FL)$^2$ achieves **best** or **nearly the best** performance **across all settings**
  - SemiFL struggles to generalize even though performs best in few scenarios
  - (FL)$^2$ consistently maintains high performance across all tasks

# Performance Comparison with Existing FSSL Algorithms

- Average accuracy(%) and standard deviation across three different seeds
- **Bold**: best result / underline: second-best result

| Dataset | | CIFAR10 | | SVHN | | CIFAR100 | |
|---|---|---|---|---|---|---|---|
| # of labeled data samples ($N_L$) | | 10 | 40 | 40 | 250 | 100 | 400 |
| Unbalanced Non-IID, Dir(0.1) | FedMatch | 16.0(2.3) | 25.6(2.2) | 20.7(2.7) | 70.1(2.2) | 6.3(0.3) | 10.0(1.8) |
| | FedCon | 16.6(2.1) | 25.4(2.3) | 20.5(1.4) | 73.1(2.0) | 4.0(0.4) | 8.2(0.6) |
| | SemiFL | 10.0(0.0) | 19.9(7.5) | 18.0(2.6) | 82.3(1.8) | 9.8(2.4) | 13.5(5.0) |
| | $(FL)^2$ | **19.2(5.7)** | **36.4(1.4)** | **21.5(3.3)** | **88.0(1.0)** | **10.4(1.3)** | **23.5(1.2)** |
| Unbalanced Non-IID, Dir(0.3) | FedMatch | 15.3(1.3) | 25.2(3.5) | 22.3(0.7) | 72.3(3.0) | 5.5(1.5) | 9.8(1.1) |
| | FedCon | 16.9(2.4) | 26.5(2.1) | 21.6(1.7) | 68.7(2.7) | 5.8(0.6) | 13.3(0.9) |
| | SemiFL | 10.0(0.0) | 38.0(2.7) | 26.3(2.5) | 42.7(40.1) | **12.4(1.2)** | 18.9(9.7) |
| | $(FL)^2$ | **24.3(4.5)** | **43.5(7.5)** | **31.0(4.2)** | **92.6(0.5)** | 12.1(1.1) | **25.4(1.0)** |
| Balanced IID | FedMatch | 16.2(1.9) | 25.4(2.8) | 18.4(4.7) | 66.2(0.8) | 6.4(0.6) | 10.0(1.7) |
| | FedCon | 16.7(2.0) | 23.3(6.2) | 20.3(1.0) | 71.6(1.5) | 5.7(0.6) | 12.4(1.6) |
| | SemiFL | 10.0(0.0) | 75.3(2.8) | 53.4(13.3) | 43.3(41.0) | 13.9(3.3) | 27.9(6.1) |
| | $(FL)^2$ | **38.9(11.1)** | **81.5(7.4)** | **75.3(2.4)** | **94.6(1.1)** | **14.4(2.3)** | **28.1(2.2)** |

- $(FL)^2$ achieves **best** or **nearly the best** performance **across all settings**
  - SemiFL struggles to generalize even though performs best in few scenarios
  - $(FL)^2$ consistently maintains high performance across all tasks

# Performance Comparison with Existing FSSL Algorithms

- Average accuracy(%) and standard deviation across three different seeds
- **Bold**: best result / <u>underline</u>: second-best result

| Dataset | | CIFAR10 | | SVHN | | CIFAR100 | |
|---|---|---|---|---|---|---|---|
| # of labeled data samples ($N_L$) | | 10 | 40 | 40 | 250 | 100 | 400 |
| Unbalanced Non-IID, Dir(0.1) | FedMatch | 16.0(2.3) | <u>25.6(2.2)</u> | <u>20.7(2.7)</u> | 70.1(2.2) | 6.3(0.3) | 10.0(1.8) |
| | FedCon | <u>16.6(2.1)</u> | 25.4(2.3) | 20.5(1.4) | 73.1(2.0) | 4.0(0.4) | 8.2(0.6) |
| | SemiFL | 10.0(0.0) | 19.9(7.5) | 18.0(2.6) | <u>82.3(1.8)</u> | <u>9.8(2.4)</u> | <u>13.5(5.0)</u> |
| | *(FL)*$^2$ | **19.2(5.7)** | **36.4(1.4)** | **21.5(3.3)** | **88.0(1.0)** | **10.4(1.3)** | **23.5(1.2)** |
| Unbalanced Non-IID, Dir(0.3) | FedMatch | 15.3(1.3) | 25.2(3.5) | 22.3(0.7) | <u>72.3(3.0)</u> | 5.5(1.5) | 9.8(1.1) |
| | FedCon | <u>16.9(2.4)</u> | 26.5(2.1) | 21.6(1.7) | 68.7(2.7) | 5.8(0.6) | 13.3(0.9) |
| | SemiFL | 10.0(0.0) | <u>38.0(2.7)</u> | <u>26.3(2.5)</u> | 42.7(40.1) | **12.4(1.2)** | <u>18.9(9.7)</u> |
| | *(FL)*$^2$ | **24.3(4.5)** | **43.5(7.5)** | **31.0(4.2)** | **92.6(0.5)** | <u>12.1(1.1)</u> | **25.4(1.0)** |
| Balanced IID | FedMatch | 16.2(1.9) | 25.4(2.8) | 18.4(4.7) | 66.2(0.8) | 6.4(0.6) | 10.0(1.7) |
| | FedCon | <u>16.7(2.0)</u> | 23.3(6.2) | 20.3(1.0) | <u>71.6(1.5)</u> | 5.7(0.6) | 12.4(1.6) |
| | SemiFL | 10.0(0.0) | <u>75.3(2.8)</u> | <u>53.4(13.3)</u> | 43.3(41.0) | <u>13.9(3.3)</u> | <u>27.9(6.1)</u> |
| | *(FL)*$^2$ | **38.9(11.1)** | **81.5(7.4)** | **75.3(2.4)** | **94.6(1.1)** | **14.4(2.3)** | **28.1(2.2)** |

- (FL)$^2$ achieves **best** or **nearly the best** performance **across all settings**
  - SemiFL struggles to generalize even though performs best in few scenarios
  - (FL)$^2$ consistently maintains high performance across all tasks

# Performance Comparison with Existing FSSL Algorithms

- Average accuracy(%) and standard deviation across three different seeds
- **Bold**: best result / underline: second-best result

| Dataset | | CIFAR10 | | SVHN | | CIFAR100 | |
|---|---|---|---|---|---|---|---|
| # of labeled data samples ($N_L$) | | 10 | 40 | 40 | 250 | 100 | 400 |
| Unbalanced Non-IID, Dir(0.1) | FedMatch | 16.0(2.3) | 25.6(2.2) | 20.7(2.7) | 70.1(2.2) | 6.3(0.3) | 10.0(1.8) |
| | FedCon | 16.6(2.1) | 25.4(2.3) | 20.5(1.4) | 73.1(2.0) | 4.0(0.4) | 8.2(0.6) |
| | SemiFL | 10.0(0.0) | 19.9(7.5) | 18.0(2.6) | 82.3(1.8) | 9.8(2.4) | 13.5(5.0) |
| | *(FL)*$^2$ | **19.2(5.7)** | **36.4(1.4)** | **21.5(3.3)** | **88.0(1.0)** | **10.4(1.3)** | **23.5(1.2)** |
| Unbalanced Non-IID, Dir(0.3) | FedMatch | 15.3(1.3) | 25.2(3.5) | 22.3(0.7) | 72.3(3.0) | 5.5(1.5) | 9.8(1.1) |
| | FedCon | 16.9(2.4) | 26.5(2.1) | 21.6(1.7) | 68.7(2.7) | 5.8(0.6) | 13.3(0.9) |
| | SemiFL | 10.0(0.0) | 38.0(2.7) | 26.3(2.5) | 42.7(40.1) | **12.4(1.2)** | 18.9(9.7) |
| | *(FL)*$^2$ | **24.3(4.5)** | **43.5(7.5)** | **31.0(4.2)** | **92.6(0.5)** | 12.1(1.1) | **25.4(1.0)** |
| Balanced IID | FedMatch | 16.2(1.9) | 25.4(2.8) | 18.4(4.7) | 66.2(0.8) | 6.4(0.6) | 10.0(1.7) |
| | FedCon | 16.7(2.0) | 23.3(6.2) | 20.3(1.0) | 71.6(1.5) | 5.7(0.6) | 12.4(1.6) |
| | SemiFL | 10.0(0.0) | 75.3(2.8) | 53.4(13.3) | 43.3(41.0) | 13.9(3.3) | 27.9(6.1) |
| | *(FL)*$^2$ | **38.9(11.1)** | **81.5(7.4)** | **75.3(2.4)** | **94.6(1.1)** | **14.4(2.3)** | **28.1(2.2)** |

- (FL)$^2$ significantly outperforms other methods when **labeled data is extremely limited**
  - **21.9%** in IID/SVHN/40-labels, **22.2%** in IID/CIFAR10/10-labels

# $(FL)^2$ :

## Few-Labels Federated Semi-Supervised Learning

- Effectively reduces **confirmation bias** with novel methods
  - **CAT**: Client-specific Adaptive Thresholding
  - **LSAA**: Learning Status-Aware Aggregation
  - **SACR**: Sharpness-Aware Consistency Regularization
- Outperforms existing FSSL methods up to **23.0%** accuracy
- Closes gap between SSL and FSSL, especially when **labels are scarce**

Email: seungjoo.lee@kaist.ac.kr
Homepage: seungjoo.com