International Journal of Medical Informatics 195 (2025) 105806

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

FL-W3S: Cross-domain federated learning for weakly supervised semantic segmentation of white blood cells

Hussain Ahmad Madni [a,*,1], Rao Muhammad Umer [b,1], Silvia Zottin [a], Carsten Marr [b], Gian Luca Foresti [a]

[a] Department of Computer Science and Artificial Intelligence, University of Udine, 33100, Italy
[b] Institute of AI for Health, Helmholtz Zentrum München, 85764 Munich, Germany

ARTICLE INFO

Keywords:
Federated learning
White blood cell
Weakly supervised semantic segmentation
Transformer attention

ABSTRACT

Background: Segmentation models for clinical data experience severe performance degradation when trained on a single client from one domain and distributed to other clients from different domain. Federated Learning (FL) provides a solution by enabling multi-party collaborative learning without compromising the confidentiality of clients' private data.

Methods: In this paper, we propose a cross-domain FL method for Weakly Supervised Semantic Segmentation (FL-W3S) of white blood cells in microscopic images. We perform model training on multiple clients with different data distributions to obtain a global aggregated model using only image-level class labels for semantic segmentation of white blood cells. A multi-class token transformer model learns the relationship between patch tokens and class tokens during collaborative learning and generates class-specific localization maps for mask predictions. To rectify the localization maps, we use patch-level pairwise affinity obtained from patch-to-patch transformer attention.

Results: We evaluate performance of the proposed semantic segmentation method on two different datasets of white blood cells from different domains. Our experimental results show that for two datasets, there is 2.56% and 1.39% increase in performance of the proposed method over existing state-of-the-art methods.

Conclusion: The combination of federated learning for collaborative model training while preserving data privacy, alongside white blood cell segmentation techniques for precise cell identification, enhances diagnostic accuracy and personalized treatment strategies in clinical applications, particularly in hematology and pathology. More specifically, it involves isolating white blood cell from blood smear for further analysis such as automated blood cell counting, morphological analysis, cell classification, disease diagnosis and monitoring.

1. Introduction

The segmentation of white blood cells in stained peripheral blood and bone marrow samples is a key step for the automated classification of these cells and the AI-based diagnosis of hematologic malignancies [1–4]. Manual segmentation has been adopted as a standard procedure for WBCs, but this process is time-consuming, labor-intensive, and requires some level of expertise. The use of semi- or fully automatic segmentation methods can considerably reduce the time and labor needed, enhance consistency, and facilitate the analysis of large-scale datasets. Thanks to the advancements in machine learning, deep learning-based models have achieved excellent performance for accurate segmentation results across a diverse range of medical imaging tasks [5–9]. However, existing methods such as [10–15], significantly drop in their performance when trained on a single client (e.g., hospital data) and distributed to other clients. Since in clinical practice, clients' data often come from different domains (i.e., data distributions), the lack of model generalizability across domains poses a substantial obstacle to a wider application in practice.

Furthermore, getting manually annotated ground-truth segmentation masks is expensive, thus, weakly supervised or unsupervised deep segmentation models [16,17] are attractive alternatives. The crucial step involves generating pseudo segmentation ground truth labels based on the provided weak labels [18]. Semi- or Weakly Supervised Seman-

* Corresponding author.
E-mail address: hamadnig@gmail.com (H.A. Madni).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ijmedinf.2025.105806
Received 4 May 2024; Received in revised form 10 January 2025; Accepted 21 January 2025

Available online 23 January 2025

tic Segmentation (WSSS) only rely on image-level labels [18]. These image-level labels solely convey the presence of a particular object class without providing any details about the localization information within the image. However, there is an important detail in Vision Transformer (ViT) [19] features associated with semantic segmentation as disclosed by a recent method, DINO [20]. Attention maps from a class token produce semantic scene layouts. Thus, attention maps lead to an effective unsupervised segmentation. While in ViT attention, various heads address distinct semantic regions within an image, it is still not clear how a head can be associated with an accurate corresponding semantic class. Thus, attention maps in ViT remain class-agnostic so far. Exploiting class-attention maps using transformers is challenging due to tokens from one class causing precise localization for multiple objects within a single image.

Moreover, the trained segmentation models lack generalizability due to domain-shift in data among different clients based on various factors such as image resolution, color, and class labels. In this case, the traditional FL methods [21,18] cannot help a client learn a robust segmentation model for the segmentation of white blood cells.

To solve aforementioned problems, we propose FL-W3S for cross-domain FL for weakly supervised semantic segmentation of white blood cells. For WSSS, instead of one-class tokens, we leverage multi-class tokens that are useful for various object classes to learn their representations, as inspired by Xu et al. [18]. A strong relation between a class token and its corresponding label is developed that leads to an advantage of class-to-patch attention, utilized as a localization map specific to each class. Moreover, learning patch-to-patch attention is used for patch-level pairwise affinity that rectifies class-specific attention map tweaking the localization. The proposed FL-W3S uses Class Activation Mapping (CAM) [22] strategy with patch tokens by acquiring the ability to classify through representations based on class-token and patch-token simultaneously. Thus, there is consistency in patch tokens and class tokens improving the discriminative attribute concerning WBC localization maps. Furthermore, we perform cross-domain FL training to obtain a global aggregated model on a server from the parameters received from local models.

Our main contributions are:

- We propose a novel method for cross-domain FL training for the segmentation of white blood cells using only image-level ground truth labels for weak supervision.
- We exploit a multi-class token transformer to learn the relationship between patch tokens and class tokens in cross-domain FL setting.
- We use patch-level pairwise affinity obtained from patch-to-patch transformer attention to refine the localization maps in local and global models. Extensive experiments validate the effectiveness of our proposed method, rather it goes beyond state-of-the-art.

2. Methodology

We train the proposed model in a cross-domain FL environment for weakly supervised semantic segmentation of single white blood cells, explained in sections 2.1 and 2.2. There are two main components of our method; one for the multi-class token transformer and the other for cross-domain FL.

2.1. Multi-class token transformer

We exploit transformer attention for class-specific object localization maps, where an input image I is transformed into $P \times P$ (non-overlapping) patches that are converted into patch tokens $T_p \in \mathbb{R}^{N \times E}$, where E represents the embedding dimension and $N = P^2$, the number of patches as shown in Fig. 1. In traditional non-convolutional models, a single-class token is used for transformer attention. However, in the proposed method, we use multi-class tokens $T_c \in \mathbb{R}^{C \times E}$, concatenated with patch tokens along position embedding to be used as input tokens $T_i \in \mathbb{R}^{(C+N) \times E}$, for the transformer encoder, where C represents the number of classes. In transformer encoders, there are multiple consecutive encoding layers, each containing a Multi Head Attention (MHA) and Multi-Layer Perceptron (MLP) to extract features, where LayerNorm layers are implemented prior to MHA and MLP respectively.

We leverage conventional self-attention layer to capture the dependencies among tokens by normalization and transformation of input tokens to a triplet, Query $Q \in \mathbb{R}^{(C+N) \times E}$, Key $K \in \mathbb{R}^{(C+N) \times E}$, and Value $V \in \mathbb{R}^{(C+N) \times E}$ using linear layers [19]. For the computation of attention between key and query, we apply Scaled Dot Product Attention [23]. Finally, using attention values as weights, each output token is generated as the result of a weighted sum of all tokens, formulated as:

$$\text{Attention}\,(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{E}}\right)V \tag{1}$$

Thus, we can compute a token-to-token attention map $A_{tt} = \text{softmax}\left(\frac{QK^T}{\sqrt{E}}\right)$, where $A_{tt} \in \mathbb{R}^{(C+N) \times (C+N)}$. These global pairwise attention maps A_{tt} are used to extract class-to-patch attentions $A_{cp} \in \mathbb{R}^{C \times N}$ as illustrated with pink circles in the matrix shown in Fig. 1, and mathematically represented as $A_{cp} = Att[1 : C, C + 1 : C + N]$. In matrix A_{cp}, each row (i.e., vector) denotes attention score of a specific class to all patches. These vectors are used to generate C class-related localization maps. Each encoding layer of model produces class-related localization maps, where earlier layers give low level and generic information, while late layers give high level class-discriminative representations. In the proposed method, we explore the trade-off between precision and recall of localization maps by merging class-to-patch attention computed on final encoding layer L, mathematically expressed as:

$$\tilde{A}_{cs} = \frac{1}{K}\sum_{e}^{L} \tilde{A}^{e}_{cs}, \tag{2}$$

where $\tilde{A}^{e}_{cs}$ represents class-specific transformer attention learned from e^{th} encoding layer. Finally, min-max normalization is performed on $\tilde{A}^{e}_{cs}$ attentions to produce final class-specific localization maps $\tilde{A}_{cs} \in \mathbb{R}^{C \times P \times P}$.

Unlike traditional transformers using a single class token (i.e., extracted on final layer with MLP), we use average pooling and make sure that class-discriminative information is learned for multiple class tokens, $T_c \in \mathbb{R}^{C \times E}$, formulated as:

$$z\,(c) = \frac{1}{E}\sum_{q}^{E} T_c\,(c, q)\,, \tag{3}$$

where $z \in \mathbb{R}^{C}$ represents class prediction, a class $c \in 1, 2, 3, ..., C$, and $T_c\,(c, q)$ represents an element (i.e. the q^{th} feature of a class c token) in multi class attention T_c. Finally, for a class c, soft margin loss between class $z\,(c)$ and its corresponding ground truth label is computed. This enables each class token to learn class-specific information by a strong class-aware supervision. Moreover, we combine the proposed framework of a multi class token model with a CAM module [22,24,25]. If the output of model is $T_o \in \mathbb{R}^{(C+N) \times E}$, then $T^{C \times E}_{c\_o}$ and $T^{N \times E}_{p\_o}$ are output class tokens and output patch tokens respectively. The reshaped T_{p_o} are passed through a convolution layer using output channels C, generating a $2D$ feature map $\mathcal{F}_{p\_o} \in \mathbb{R}^{P \times P \times C}$ that is transformed to class predictions using Global Average Pooling (GAP). We also use T_{c_o} to compute class scores as given in (3). Thus, the total loss $\mathcal{L}_i$ for a client i is calculated from two soft margin losses: 1) A loss $\mathcal{L}_{class-class}$ calculated between ground truth class label and the prediction from class tokens. 2) A loss $\mathcal{L}_{class-patch}$ between ground truth class label and prediction from patch tokens. It is formulated as follows.

$$\mathcal{L}_i = \mathcal{L}_{class-patch} + \mathcal{L}_{class-class} \tag{4}$$

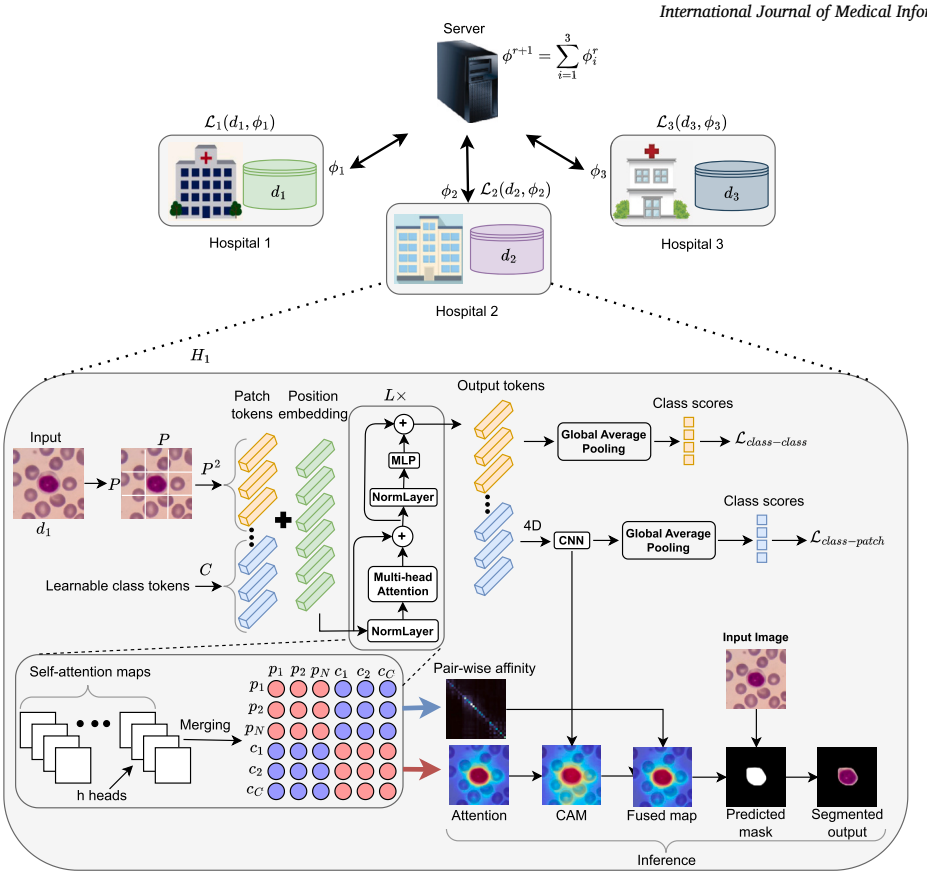


Fig. 1. FL-W3S architecture: A multi-class token transformer model training on cross-domain FL scenario produces class-specific attention maps. Pair-wise affinity obtained from patch-to-patch transformer attention is used to refine the object localization map. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

During inference, we extract CAM from final convolution layer by using min-max normalization on $\mathcal{F}_{p\_o}$ that is integrated with class-specific attention maps A_{cs} to give an object localization map A_{obj} by applying element-wise multiplication:

$$A_{obj} = A_{CAM} \circ A_{cs}. \tag{5}$$

Here, $\circ$ is the Hadamard product, and A_{CAM} represents a patch-token based CAM.

Many existing methods [26–28] used pairwise affinity for the refinement of object localization maps that need additional layer or network for the learning of an affinity map. The patch-level pairwise affinity obtained from patch-to-patch transformer attention refines localization maps by leveraging global contextual relationships between patches in an image. The attention mechanism generates an affinity matrix that captures the strength of relationships between all patches, enabling information propagation across related regions. This helps refine fragmented or noisy initial localization maps by smoothing activations and enforcing spatial coherence. Unlike traditional methods with limited receptive fields, the transformer-based approach models long-range dependencies, allowing distant but semantically related patches to influence each other. As a result, the refined localization maps are more precise, noise-free, and semantically aligned with underlying objects or regions of interest. Instead, we use patch-to-patch attention, $A_{pp} = A_{tt}[C+1:C+N, C+1:C+N]$ (i.e. a matrix with blue circles in Fig. 1) and transform it into a $4D$ tensor as $\tilde{A}_{pp} \in \mathbb{R}^{P\times P\times P\times P}$ to generate a pairwise affinity map that does not require any additional supervision or computation. The generated affinity is utilized to improve class-specific attention. Thus, for a client i, we extract patch-to-patch attention maps as a patch-level pairwise affinity from A_{obj}, given as follows.

$$A_i(c,p,q) = \sum_{l}^{P}\sum_{e}^{P} \tilde{A}_{pp}(p,q,l,e) \cdot A_{obj}(c,l,e), \tag{6}$$

2.2. Cross-domain federated learning

Due to data diversity in clients, a global model in FL suffers performance degradation while extracting knowledge from clients' data. FL model is converged after a particular global communication rounds, as server and each client are updated after every global communication round. Although a pre-trained model does not perform well for all clients, it can gain knowledge from most of the collaborative clients, making it a generalized and robust model. Thus, the proposed method acquires extensive global knowledge through cross-domain FL.

In FL, if there are H clients (e.g. hospitals), a global model ϕ is trained on each client $i \in H$ having local data d_i during a local communication round r, and produces a loss $\mathcal{L}_i$. The objective function of each client is given as

$$\phi_i = \arg\min \mathcal{L}_i(f(d_i, \phi_i), y_i), \tag{7}$$

where local model ϕ_i is trained on local dataset d_i with labels y_i. Each local model is updated after each local communication round, and an average loss $\mathcal{L}_i(d_i, \phi)$ of a client $i \in H$ is computed. After a given number of local rounds, a global communication round is executed (see Fig. 1) to perform a global update on the server computed as follows:

$$\phi^{r+1} = \sum_{i=1}^{H} \phi_i^r. \tag{8}$$

Here, ϕ^r is a global model on the server, while ϕ_i^r represents the model of a client $i \in H$ for a communication round r. Thus, after a given number of local rounds, all clients send their model parameters to the server

Table 1
A summary of datasets used in the experiments of the proposed methodology.

Dataset	Classes	No. of Training Samples	No. of Validation Samples	No. of Test samples	Total Samples
Raabin [33]	5	687	229	229	1145
Matek19 [1]	15	771	257	257	1285
INT_20 [2]	13	900	300	300	1500

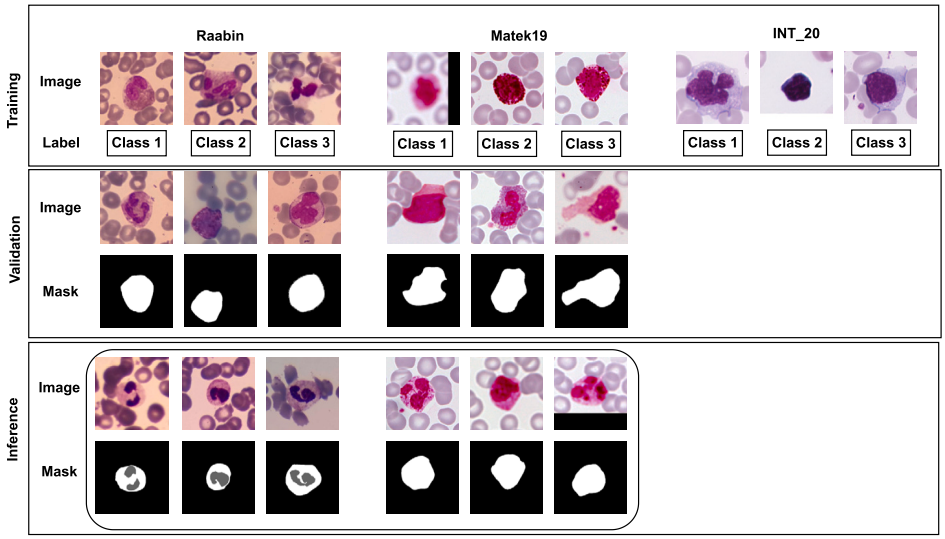


Fig. 2. Data distribution over training, validation and inference steps. In the training step, RGB images and class labels used. In the validation and inference steps, RGB images and their ground truth masks are used. Raabin, Matek19 and INT_20 datasets are used for training, while INT_20 is not used during validation and inference stage due to unavailable ground truth masks.

for aggregation step. Finally, the server aggregates the parameters and broadcasts the updated model to each client for the next global communication round.

We use a common federated averaging method, FedAvg [29] as used and discussed in [30–32], to perform the aggregation of model parameters obtained from local clients. Thus, the objective of global model is to minimize the cost function expressed as

$$\min_{\phi \in R}\{\mathcal{L}(\phi, D) = \sum_{i=1}^{H} \frac{b_i}{H_i}\mathcal{L}_i(d_i, \phi_i)\}, \tag{9}$$

where D represents the collective data from all domains on which the global model ϕ is trained through local models, b_i denotes the batch size and $\mathcal{L}_i$ represents individual loss of a local client $i \in H$ trained on the local dataset d_i.

3. Experiments

3.1. Datasets

In our experiments, we use Raabin [33] dataset with 5 classes, Matek19 [1] dataset with 15 classes, and an internal dataset INT_20 [2] with 13 classes, all containing RGB images of single white blood cells. As summarized in Table 1, the Raabin dataset contains 1145 RGB images and their ground truth masks, a subset of Matek19 dataset contains 1285 RGB images with ground truth masks. Due to lack of ground truth masks, we use a subset of INT_20 dataset containing only 1500 samples for model training because the proposed model only needs class labels, and does not require ground truth masks for training as shown in Fig. 2.

We define three clients in our experiments, where each non-overlapping dataset belongs to a client participating in collaborative training of FL-W3S model. We perform five-fold cross-validation for all datasets by splitting each dataset into 60% train, 20% validation, and 20% testset. However, INT_20 dataset is used only for training due to lack of ground truth masks. Thus, each client contains its local trainset and validation to train local model. After a given number of global communication rounds, the final global model is produced that is evaluated individually using a non-overlapping testset (i.e., with corresponding ground truth masks) from Raabin and Matek19 datasets respectively.

3.2. Evaluation metrics

As used in other similar existing methods [18,34], we use Intersection over Union (IoU) score measured with testset to evaluate the segmentation performance of the model.

3.3. Implementation details

For each local client, we use DeiT-S [35,24] backbone model that is pre-trained on ImageNet [36]. More specifically, to initialize multi-class tokens, we use pre-trained DeiT-S class tokens in our approach. For the model hyperparameters, we follow [18] to train the local model. All input images are resized to 256×256 to make a standard size for all multi-shape images, later cropped into 224×224 for the input to model. Moreover, we use ResNet38 [37] based Deeplab VI by following prior works [38,39,26,27] for the semantic segmentation. In our experiments, there are 3 clients each having a local dataset d_i and a local model ϕ_i to be trained on that dataset during local communication rounds. For a global communication round, the parameters of all local models are aggregated on the server according to FedAvg [29] aggregation algorithm. We set 45 as local and global communication rounds. The hyperparameters used in the experiments are given in Table 2.

3.4. Performance comparison of the proposed method with existing methods in terms of mean IoU (%)

For the generation of pseudo masks for semantic segmentation, we employ PSA [27] on the object localization maps. We compute mean IoU (mIoU) with standard deviation between predicted masks and ground

Table 2
List of hyperparameters set in the experiments.

Hyperparameter	Value(s)
H (No. of clients)	3
C (No. of classes)	15
b_i (batch-size)	64
r (local epoch)	45
global epoch	45
input size	224
P (patch size)	16
optimizer	AdamW [40]
SGD momentum	0.9
weight decay	0.05
initial learning rate	$5e-4$

Table 3
Comparison of proposed approach with state-of-the-art methods based on Raabin and Matek19 WBC datasets. The proposed method (FL-W3S) outperforms the existing methods.

Method	Backbone	Raabin (mIoU ± SD)	Matek19 (mIoU ± SD)
AuxSegNet [26]	ResNet38	37.2 ± 0.9	44.12 ± 0.09
SEAM [41]	ResNet38	33.6 ± 0.5	41 ± 0.7
EPS [42]	ResNet101	36.7 ± 0.4	45.7 ± 0.8
Luo et al. [43]	VGG16	36.93 ± 0.39	41.9 ± 0.4
CDA [44]	ResNet38	37.18 ± 0.18	43.5 ± 0.4
MCTformer [18]	ResNet38	36.97 ± 0.79	44.8 ± 0.5
Wang et al. [28]	VGG16	31.86 ± 0.87	38.9 ± 0.2
CONTA [45]	ResNet38	33.78 ± 0.66	40.6 ± 0.5
FL-W3S (Ours)	ResNet38	**39.8 ± 0.5**	**47 ± 1.6**

truth masks to compare and evaluate the performance of the proposed method with existing methods. For the evaluation, we use a global testset extracted from both Raabin and Matek19 datasets. The performance of FL-W3S surpasses that of existing methods as given in Table 3. Moreover, it is shown that there is a remarkable difference between performance of the proposed method and existing methods. The better performance of FL-W3S is mainly because of FL enhancing the model robustness and generalization. We achieved IoU scores of 39.8% and 47% for a non-overlapping testset from Raabin and Matek19 datasets respectively.

We also compare the proposed method with a couple of competitive existing methods by the visualization of predicted masks and ground truth masks taken randomly from both Raabin and Matek19 datasets as shown in Fig. 3. For visual comparison with state-of-the-art, we select existing methods with comparatively higher performance (i.e., in Table 3). We also compare the proposed method with the SAM [5] only for the visualization because SAM requires a prompt (i.e., bounding box) for the segmentation, while other methods including FL-W3S perform segmentation without any prompt. It is clear from visual comparison that the proposed method predicts more refined masks with refined edges as compared to the predicted masks of other methods.

To analyze the model convergence on Raabin dataset, we plot mean IoU up to ≈ 20 rounds (Fig. 4). Models converge after a certain number of global communication rounds. Evidently, the mIoU for the proposed FL-W3S is higher throughout all communication rounds.

Although FL has outperformed when used in multi-domain scenario, data heterogeneity affects the model performance and consistency in terms of generalizability. Moreover, model training on large data from multiple locations causes communication overhead in collaborative learning. Similarly, as the number of clients is increased, training time is exponentially increased by creating challenges for efficient model aggregation.

4. Conclusion

We propose a novel FL-W3S method for the segmentation of white blood cells, in which we perform cross-domain FL training to obtain a global aggregated weakly supervised semantic segmentation model through federated averaging on a central server, and send the aggregated global model back to each client for fine-tuning its personalized local model. We performed experiments on three different white blood cell datasets and evaluated the proposed model with existing methods based on segmentation tasks. The experimental results show the superiority of the proposed method over existing segmentation methods.

In the future, we plan to address data heterogeneity challenges by incorporating advanced federated optimization techniques, such as FedProx and Scaffold, to enhance convergence and robustness. Additionally, we aim to explore methods for reducing communication overhead, such as model compression and quantization, to improve scalability. Extending the framework to diverse segmentation tasks, including 3D imaging and multi-modal data, is another priority. Furthermore, we plan to in-

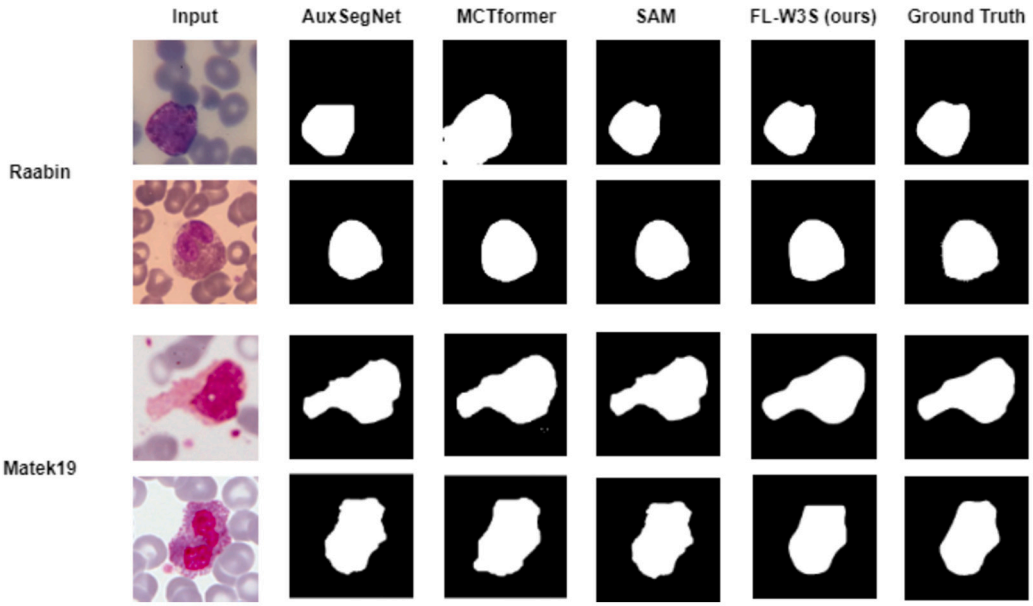
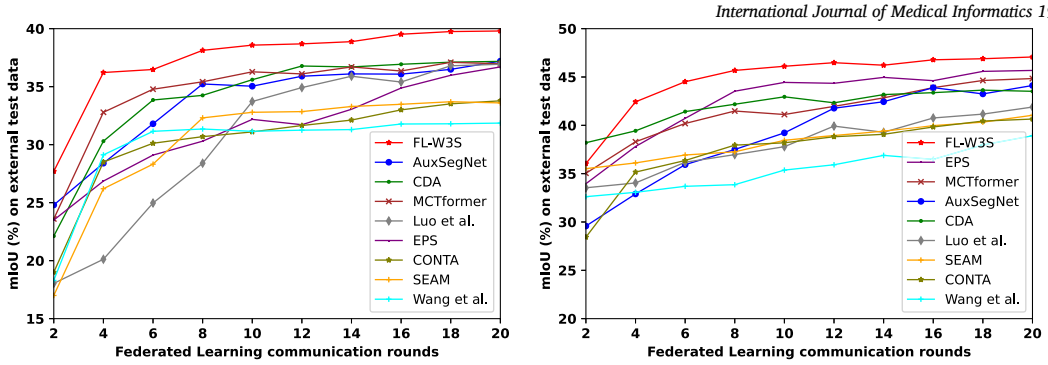


Fig. 3. Visual comparison of the proposed model with existing methods. Two random input images and their ground truth masks for each dataset (i.e., Raabin and Matek19) are selected for the comparison of the proposed method with existing competitive methods.



(a) Model convergence in terms of mIoU computed on Raabin test dataset.

(b) Model convergence in terms of mIoU computed on Matek19 dataset.

Fig. 4. Model convergence for the comparison of proposed FL-W3S method with the existing methods in terms of mean IoU (%) vs FL communication rounds.

tegrate privacy-preserving mechanisms like differential privacy and investigate interpretability techniques to make the segmentation results more explainable. Real-world deployment on edge devices and lifelong learning capabilities will also be considered to ensure practical applicability and continuous updates. These directions will strengthen the robustness, scalability, and generalization of the proposed framework.

5. Summary table

What was already known on the topic:

- Most of the existing methods consider data splits from a single domain for segmentation model training. Traditionally, ground truth masks are used for training and inference.
- Transformer architecture is based on one-class tokens in most existing approaches.

What this study added to our knowledge:

- We perform segmentation task by model training in a multi-domain environment for multiple clients in collaborative learning using only images and class labels
- We exploit Transformer architecture for multi-class tokens in contrast to existing methods.

CRediT authorship contribution statement

Hussain Ahmad Madni: Writing – original draft, Methodology, Conceptualization. **Rao Muhammad Umer:** Writing – review & editing, Data curation. **Silvia Zottin:** Writing – review & editing, Formal analysis, Conceptualization. **Carsten Marr:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Gian Luca Foresti:** Writing – review & editing, Supervision, Project administration.

Ethics

As this work did not involve human participants and all data was obtained from publicly available sources, no ethical approval was obtained for this project.

Funding

- This work was partially supported by the FVG Project “Supporting the diagnosis of rare diseases (MR) through artificial intelligence” (2023-26) (Project A with CUP: F53C22001770002 and Project B with CUP F53C22001780002).
- This work was partially supported by project SERICS (PE00000014) under the MUR National Recovery and Resilience Plan (PNRR) funded by the European Union - Next Generation EU, Mission 4, CUP G23C24000790006 (2024-25).
- This work was partially supported by Piano Nazionale di Ripresa e Resilienza (PNRR) DD 3277 of December 30, 2021 (Mission 4, Component 2, Investment 1.5) - Interconnected Nord-Est Innovation Ecosystem (iNEST).
- Carsten Marr acknowledges funding from The European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (Grant Agreement No. 866411 & 101113551) and support from the Hightech Agenda Bayern.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Matek, S. Schwarz, K. Spiekermann, C. Marr, Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks, Nat. Mach. Intell. 1 (2019) 538–544.

[2] R.M. Umer, A. Gruber, S.S. Boushehri, C. Metak, C. Marr, Imbalanced domain generalization for robust single cell classification in hematological cytomorphology, 2023.

[3] W. Walter, C. Pohlkamp, M. Meggendorfer, N. Nadarajah, W. Kern, C. Haferlach, T. Haferlach, Artificial intelligence in hematological diagnostics: game changer or gadget?, Blood Rev. (2022) 101019.

[4] Z. Ullah, M. Usman, M. Jeon, J. Gwak, Cascade multiscale residual attention cnns with adaptive roi for automatic brain tumor segmentation, Inf. Sci. 608 (2022) 1541–1556, https://doi.org/10.1016/j.ins.2022.07.044.

[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.-Y. Lo, et al., Segment anything, arXiv preprint, arXiv: 2304.02643, 2023.

[6] J. Ma, Y. He, F. Li, L. Han, C. You, B. Wang, Segment anything in medical images, Nat. Commun. 15 (2024) 654.

[7] Z. Ullah, M. Usman, J. Gwak, Mtss-aae: multi-task semi-supervised adversarial autoencoding for covid-19 detection based on chest X-ray images, Expert Syst. Appl. 216 (2023) 119475, https://doi.org/10.1016/j.eswa.2022.119475.

[8] Z. Ullah, M. Usman, S. Latif, et al., Densely attention mechanism based network for covid-19 detection in chest X-rays, Sci. Rep. 13 (2023) 261, https://doi.org/10.1038/s41598-022-27266-9.

[9] X. Li, T. Zhou, J. Li, Y. Zhou, Z. Zhang, Group-wise semantic mining for weakly supervised semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 1984–1992.

[10] A. Raza, J. Uddin, Q. Zou, S. Akbar, W. Alghamdi, R. Liu, Aips-deepenc-ga: predicting anti-inflammatory peptides using embedded evolutionary and sequential feature integration with genetic algorithm based deep ensemble model, Chemom. Intell. Lab. Syst. 254 (2024) 105239.

[11] M. Ullah, S. Akbar, A. Raza, Q. Zou, Deepavp-tppred: identification of antiviral peptides using transformed image-based localized descriptors and binary tree growth algorithm, Bioinformatics 40 (2024) btae305.

[12] S. Akbar, A. Raza, Q. Zou, Deepstacked-avps: predicting antiviral peptides using tri-segment evolutionary profile and word embedding based multi-perspective features with deep stacking model, BMC Bioinform. 25 (2024) 102.

[13] S. Akbar, Q. Zou, A. Raza, F.K. Alarfaj, Iafps-mv-bitcn: predicting antifungal peptides using self-attention transformer embedding and transform evolutionary based multi-view features with bidirectional temporal convolutional networks, Artif. Intell. Med. 151 (2024) 102860.

[14] G. Rukh, S. Akbar, G. Rehman, F.K. Alarfaj, Q. Zou, Stackedenc-aop: prediction of antioxidant proteins using transform evolutionary and sequential features based multi-scale vector with stacked ensemble learning, BMC Bioinform. 25 (2024) 256.

[15] A. Raza, J. Uddin, A. Almuhaimeed, S. Akbar, Q. Zou, A. Ahmad, Aips-sntcn: predicting anti-inflammatory peptides using fasttext and transformer encoder-based hybrid word embedding with self-normalized temporal convolutional networks, J. Chem. Inf. Model. 63 (2023) 6537–6554.

[16] T. Wald, S. Roy, G. Koehler, N. Disch, M.R. Rokuss, J. Holzschuh, D. Zimmerer, K. Maier-Hein, Sam. md: zero-shot medical image segmentation capabilities of the segment anything model, in: Medical Imaging with Deep Learning, 2023, short paper track.

[17] A. de Vulpian, V. di Proietto, G. Roy, S.B. Hadj, R.R. Fick, A semi-supervised deep learning approach for multi-stain foreground segmentation in digital pathology, in: Medical Imaging with Deep Learning, 2022.

[18] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, D. Xu, Multi-class token transformer for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4310–4319.

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: transformers for image recognition at scale, in: ICLR, 2021.

[20] N. Araslanov, S. Roth, Single-stage semantic segmentation from image labels, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4253–4262.

[21] S. Su, B. Li, C. Zhang, M. Yang, X. Xue, Cross-domain federated object detection, in: 2023 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2023, pp. 1469–1474.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[24] W. Gao, F. Wan, X. Pan, Z. Peng, Q. Tian, Z. Han, B. Zhou, Q. Ye, Ts-cam: token semantic coupled attention map for weakly supervised object localization, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2886–2895.

[25] X. Zhang, Y. Wei, J. Feng, Y. Yang, T.S. Huang, Adversarial complementary learning for weakly supervised object localization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1325–1334.

[26] L. Xu, W. Ouyang, M. Bennamoun, F. Boussaid, F. Sohel, D. Xu, Leveraging auxiliary tasks with affinity learning for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6984–6993.

[27] J. Ahn, S. Kwak, Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4981–4990.

[28] X. Wang, S. Liu, H. Ma, M.-H. Yang, Weakly-supervised semantic segmentation by iterative affinity learning, Int. J. Comput. Vis. 128 (2020) 1736–1749.

[29] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, PMLR, 2017, pp. 1273–1282.

[30] H.A. Madni, R.M. Umer, G.L. Foresti, Blockchain-based swarm learning for the mitigation of gradient leakage in federated learning, IEEE Access 11 (2023) 16549–16556.

[31] H.A. Madni, R.M. Umer, G.L. Foresti, Swarm-fhe: fully homomorphic encryption-based swarm learning for malicious clients, Int. J. Neural Syst. (2023) 2350033.

[32] P. Dhade, P. Shirke, Federated learning for healthcare: a comprehensive review, Eng. Proc. 59 (2024) 230.

[33] Z.M. Kouzehkanan, S. Saghari, S. Tavakoli, P. Rostami, M. Abaszadeh, F. Mirzadeh, E.S. Satlsar, M. Gheidishahran, F. Gorgi, S. Mohammadi, et al., A large dataset of white blood cells containing cell locations and types, along with segmented nuclei and cytoplasm, Sci. Rep. 12 (2022) 1123.

[34] J. Lee, E. Kim, S. Yoon, Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4071–4080.

[35] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[37] Z. Wu, C. Shen, A. Van Den Hengel, Wider or deeper: revisiting the resnet model for visual recognition, Pattern Recognit. 90 (2019) 119–133.

[38] F. Zhang, C. Gu, C. Zhang, Y. Dai, Complementary patch for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7242–7251.

[39] B. Zhang, J. Xiao, Y. Wei, M. Sun, K. Huang, Reliability does matter: an end-to-end weakly supervised semantic segmentation approach, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 12765–12772.

[40] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019, https://openreview.net/forum?id=Bkg6RiCqY7.

[41] Y. Wang, J. Zhang, M. Kan, S. Shan, X. Chen, Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12275–12284.

[42] S. Lee, M. Lee, J. Lee, H. Shim, Railroad is not a train: saliency as pseudo-pixel supervision for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5495–5505.

[43] W. Luo, M. Yang, Learning saliency-free model with generic features for weakly-supervised semantic segmentation, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, 2020, pp. 11717–11724.

[44] Y. Su, R. Sun, G. Lin, Q. Wu, Context decoupling augmentation for weakly supervised semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 7004–7014.

[45] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, Q. Sun, Causal intervention for weakly-supervised semantic segmentation, Adv. Neural Inf. Process. Syst. 33 (2020) 655–666.