IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 April 2025, accepted 8 May 2025, date of publication 12 May 2025, date of current version 19 May 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3569073

RESEARCH ARTICLE

Federated Knowledge Distillation With 3D Transformer Adaptation for Weakly Labeled Multi-Organ Medical Image Segmentation

TAREQ MAHMOD ALZUBI[1] AND HAMZA MUKHTAR[2,3]

[1]Department of Computer Science, Prince Abdullah bin Ghazi Faculty of Information and Communication Technology, Al-Balqa Applied University, As-Salt 19117, Jordan
[2]Department of Computer Science, University of Engineering and Technology Lahore (UET Lahore), Lahore 39161, Pakistan
[3]Intelligent Criminology Laboratory, National Center of Artificial Intelligence, Al-Khawarizmi Institute of Computer Science, UET Lahore, Lahore 39161, Pakistan

Corresponding author: Tareq Mahmod Alzubi (tareqzubi@bau.edu.jo)

ABSTRACT The increasing reliance on medical image segmentation for disease diagnosis, treatment planning, and therapeutic assessment has highlighted the need for robust and generalized deep learning (DL)-based segmentation frameworks. However, existing models often suffer from task-specific limitations, catastrophic forgetting, and poor scalability due to their dependency on narrowly annotated datasets. This creates a significant gap in developing unified, multi-organ segmentation systems that leverage distributed and partially labeled datasets across diverse clinical institutions. To address these challenges, we propose the Federated 3D Knowledge Distillation Network (Fed3D-KDNet), a hybrid federated learning (FL) framework that integrates both global and local knowledge distillation mechanisms. Our model adapts the Segment Anything Model (SAM) for volumetric medical imaging by introducing architectural enhancements, including 3D spatial feature adapters and an Auto Prompt Generator (APG), to optimize spatial representation and reduce reliance on manually crafted prompts. Fed3D-KDNet employs a dual knowledge distillation strategy to mitigate catastrophic forgetting and improve cross-client knowledge transfer, ensuring robust generalization across heterogeneous datasets. The proposed methodology was evaluated on multi-organ CT datasets, including the BTCV benchmark, under centralized and federated settings. Experimental results demonstrate that Fed3D-KDNet achieves state-of-the-art performance with an average Dice score of 80.53% and an average Hausdorff Distance (HD) of 11.43 voxels in federated experiments involving seven clients, showing 5.04% improvement in Dice accuracy and a 4.35 voxel reduction in HD. Moreover, our model demonstrates superior efficiency with a computational cost of 371.3 GFLOPs, 26.53 million tuned parameters, and an inference time of 0.058 seconds per iteration. These results validate the efficacy, scalability, and computational efficiency of Fed3D-KDNet, positioning it as a robust solution for multi-organ medical image segmentation in federated environments.

INDEX TERMS Federated learning, medical image segmentation, transformer, knowledge distillation.

I. INTRODUCTION

Medical image segmentation plays a pivotal role in artificial intelligence (AI)-assisted clinical decision support systems, enabling applications such as disease diagnosis, treatment planning, and therapy response evaluation. However, the development and deployment of accurate segmentation models heavily depend on manual annotations provided by domain experts, such as radiologists, which are both time-intensive and resource-demanding. As a result, many deep learning (DL)-based segmentation models in the existing literature are often tailored to address specific anatomical structures or medical conditions, reflecting the narrow scope of individual research objectives. For example,

The associate editor coordinating the review of this manuscript and approving it for publication was Ines Domingues.

a segmentation model trained exclusively to identify pneumonia in lung CT scans might lack the capability to detect other abnormalities, such as lung tumors. This fragmented and task-specific approach leads to the creation of numerous isolated models, each designed for a specific task, ultimately limiting their scalability and interoperability in clinical practice [1].

One of the fundamental challenges in leveraging SAM for medical imaging lies in its original architecture, which is inherently optimized for 2D natural image processing. This design fails to effectively capture and utilize the 3D spatial context that is integral to volumetric medical imaging modalities such as CT scans and MRI volumes. Similar architectural challenges have been addressed in other domains through lightweight and attention-based modules, such as the LEG-Net and RGS-Net frameworks used in defect detection for steel strip surfaces [53], [54]. These networks demonstrate how efficient spatial feature extraction combined with attention-guided multiscale fusion can significantly enhance performance in resource-constrained environments. Analogously, integrating such efficiency-driven modules in medical segmentation frameworks could bridge the gap between general-purpose foundation models and domain-specific medical applications.

The deployment of such specialized models in real-world clinical settings would require managing and maintaining hundreds of isolated systems, posing significant logistical and computational challenges. Therefore, there is a pressing need for collaborative training strategies that enable the development of generalized global segmentation models by integrating insights from distributed and decentralized datasets curated by various research groups [2]. Such an approach to knowledge aggregation can address several key challenges. Firstly, it would reduce redundancy by avoiding repeated manual annotations for overlapping tasks. Secondly, it would allow research groups to leverage each other's labeled datasets without the need for direct data sharing, thus preserving data privacy and adhering to regulatory constraints. Finally, a federated learning paradigm for collaborative training holds the potential to create versatile and robust segmentation models, capable of handling a wide range of tasks within a unified framework, thereby improving the clinical applicability and scalability of AI-driven medical image analysis systems [3].

Methods for automatic organ segmentation from abdominal CT scans primarily rely on supervised learning models [4], [5]. However, training a single model capable of segmenting all abdominal organs and associated tumors remains a significant challenge. This limitation arises from the lack of publicly available datasets containing a large volume of CT scans annotated with comprehensive segmentation labels. Existing datasets are typically partially labeled, focusing only on a subset of organs or tumor types [6]. To address the constraint of partially annotated datasets, several models have been proposed that are specifically designed to handle partially labeled CT data collected from multiple sources [7]. One naïve approach to address multi-organ segmentation involves training individual segmentation models separately on each partially labeled dataset [8]. However, this strategy is computationally expensive and yields suboptimal results, particularly when datasets are small and lack diversity.

A more refined alternative is incremental learning [9], where a single model is sequentially trained across multiple datasets. Despite showing promise, these models are highly susceptible to catastrophic forgetting, wherein previously acquired knowledge from earlier datasets is overwritten or lost during subsequent training cycles [10]. This limitation hinders their ability to generalize effectively across diverse datasets. A solution to these challenges is federated learning (FL) [11], where models are trained locally on distributed datasets across multiple sites, and the resulting parameters are aggregated via a central server. While FL has been applied to medical image segmentation tasks, existing methods generally assume that all participating nodes (or client sites) possess homogeneous datasets annotated with the same set of organs [12]. In reality, however, datasets across sites often exhibit significant heterogeneity in terms of CT acquisition protocols and annotated organ structures.

To address this inconsistency, an adaptation of federated averaging (FedAvg) [13] has been explored, where each client node trains a local encoder while sharing a common decoder across nodes [14]. Although this approach improves certain aspects of cross-site generalization, it still faces several limitations. Firstly, the approach requires tuning a large number of parameters, which increases computational overhead. Secondly, the local encoders, being trained on limited datasets, struggle to achieve optimal performance. Thirdly, catastrophic forgetting remains a critical issue during local training, where the knowledge necessary for segmenting specific organs is partially lost over iterative updates. Lastly, the isolated and independent nature of local model updates can degrade the global model's accuracy across all participating nodes.

Recent advancements in computer vision foundation models, Segment Anything Model (SAM) [15], have advanced the capabilities of image segmentation tasks. SAM has demonstrated exceptional performance and adaptability across a wide range of semantic segmentation challenges [16], offering new opportunities for medical image segmentation, a field often constrained by the limited availability and variable quality of segmentation masks. Unlike traditional task-specific transformer-based architectures, such as UNETR [17], SwinUNETR [18], and FocalUNETR [19], which are typically trained on small-scale datasets containing limited patient samples and segmentation masks, foundation models like SAM are pre-trained on massive datasets consisting of billions of images and millions of masks. This extensive pre-training provides them with a broad knowledge base and cross-domain generalization capabilities. However, when applied to medical image segmentation tasks,

foundation models often show low performance compared to specialized medical image segmentation models [20]. Efforts have been made to adapt SAM for medical image segmentation [20], [21]. However, these adaptations have highlighted persistent challenges, including reduced precision, instability, and poor segmentation quality—particularly in tasks involving small anatomical structures, irregular boundaries, and low-contrast imaging data, which are common in medical image volumes [20].

One of the fundamental challenges in leveraging SAM for medical imaging lies in its original architecture, which is inherently optimized for 2D natural image processing. This design fails to effectively capture and utilize the 3D spatial context that is integral to volumetric medical imaging modalities such as CT scans and MRI volumes. Therefore, significant architectural modifications are required to align SAM's capabilities with the spatial complexities of 3D medical data. Addressing these challenges necessitates innovative methodologies that can bridge the inherent gap between natural image segmentation workflows and the unique demands of medical image segmentation tasks. Future developments must focus on optimizing 3D spatial representations, refining pre-trained weight adaptation strategies, and incorporating domain-specific enhancements to ensure that foundation models like SAM can deliver robust and precise segmentation results in medical imaging applications.

In this paper, we propose Federated 3D Knowledge Distillation Network (Fed3D-KDNet), a novel federated learning framework specifically designed for multi-organ medical image segmentation across partially labeled and heterogeneous datasets. Our approach integrates global knowledge distillation (global-KD) and local knowledge distillation (local-KD) strategies [22] to address challenges such as catastrophic forgetting, non-independent and identically distributed (non-IID) data heterogeneity, and inefficient cross-client knowledge aggregation. By leveraging a transformer-based architecture adapted for 3D volumetric medical imaging, our model effectively captures spatial dependencies across slices, optimizing segmentation precision and consistency. The proposed Auto Prompt Generator (APG) eliminates reliance on manually defined prompts, dynamically generating context-aware representations for multi-organ segmentation tasks. Additionally, our method introduces an efficient parameter-efficient fine-tuning strategy, ensuring computational efficiency without compromising performance.

The key contributions include:

1) Our proposed hybrid Fed3D-KDNet integrates both global-KD and Local-KD mechanisms to address catastrophic forgetting and ensure consistent knowledge aggregation across distributed client datasets.
2) We introduce 3D spatial adaptation for transformer network for the enhancement of Segment Anything Model (SAM) network to effectively handle 3D volumetric medical imaging data, overcoming the limitations of conventional 2D slice-based processing.
3) Our Auto Prompt Generator (APG) automates prompt generation mechanism, eliminating reliance on manual prompt crafting, thus streamlining multi-organ segmentation and improving model scalability.

II. RELATED WORK

A. WEAKLY LABELED LEARNING

Training a unified model to handle multiple partially labeled datasets presents a significant challenge across diverse tasks, including classification [23], object detection [24], and segmentation [8]. In the context of medical image segmentation, this challenge becomes even more pronounced due to the presence of multiple anatomical structures within a single medical scan, which makes it infeasible to ensure accurate pixel-level annotations for all regions of interest. To address this issue, various methods have been proposed. One common strategy involves using U-Net architectures [25] tailored to generate predictions specific to each dataset. For instance, Chen et al. [26] designed a framework with a shared encoder and multiple decoders, each specialized for segmenting different organs. In contrast, conditional U-Net architectures [27] aim to avoid redundant encoders and decoders by adopting a single shared encoder-decoder structure for all segmentation tasks.

Another area of research focuses on the development of customized loss functions to overcome the limitations of the standard multi-class cross-entropy loss, which is not inherently suitable for partially labeled datasets. For instance, a prior-aware loss function was introduced under the assumption that a subset of the dataset contains fully labeled samples [28]. This approach regularizes model predictions to align with the distribution patterns observed in these fully labeled subsets. In addition, a marginal loss function was proposed to merge background regions with unlabeled regions, while an exclusive loss function was designed to maximize the separation between labeled and unlabeled anatomical regions [29]. These combined loss functions address the challenges posed by ambiguous boundaries in partially labeled datasets. Furthermore, a leaf-Dice loss was developed specifically to handle missing annotations, ensuring more robust segmentation performance in datasets with sparse labeling [30]. These tailored loss functions contribute significantly to improving model accuracy and generalization in complex medical image segmentation tasks.

Despite their contributions, these methods are predominantly designed for centralized training settings, where the model has access to the entire dataset at once. In contrast, our proposed approach addresses these limitations by enabling multi-organ predictions through a single forward pass, significantly reducing computational redundancy. Furthermore, our method is designed to operate seamlessly in both centralized and federated learning environments, ensuring scalability and efficiency across diverse deployment scenarios. This dual capability allows for efficient knowledge aggregation and

improved performance, even when working with partially labeled decentralized datasets.

B. FEDERATED LEARNING

FL) [8], [31] is a decentralized machine learning paradigm designed to train models collaboratively across distributed nodes while ensuring data privacy by sharing model parameters instead of raw data. This approach has gained substantial attention in medical image analysis due to its ability to address data privacy regulations and facilitate cross-institutional collaboration. A well-established method in FL is Federated Averaging (FedAvg) [13], which operates by conducting stochastic gradient descent (SGD) updates locally on each client node, followed by global aggregation of model parameters on a central server. Despite its simplicity and effectiveness, FedAvg often struggles with performance inconsistencies across participating clients, primarily due to data heterogeneity. Moreover, strategies such as the Deep Soft Threshold Feature Separation (DSTFS) network have shown effectiveness in preserving task-specific representations under adverse imaging conditions, such as thermal diffusion in infrared handprint recognition tasks [55]. This dual-task network employs disentangled feature learning via attention mechanisms, which parallels the challenges of maintaining organ-specific segmentation precision and temporal consistency in federated settings. Drawing inspiration from DSTFS, our framework applies knowledge separation and selective feature refinement to address the challenges of feature drift and catastrophic forgetting in medical FL systems.

To address these challenges, personalized federated learning (PFL) has emerged as an alternative, emphasizing the development of client-specific models tailored to the unique characteristics of each local dataset. Instead of training a unified global model, PFL optimizes models individually while still leveraging global knowledge aggregation. For example, a personalized federated learning framework [30] utilizes a partially shared model structure to accommodate varying data distributions across different client nodes. This approach maintains a shared global model while allowing localized adaptations for each client. Similarly, [32], [33] proposed an encoder-decoder architecture within a federated learning framework, specifically targeting magnetic resonance image (MRI) reconstruction tasks. In this design, a globally shared encoder is maintained on the server to learn domain-invariant representations, ensuring consistency across clients. Meanwhile, client-specific decoders are fine-tuned locally using proprietary datasets, enabling the model to capture domain-specific nuances present in individual client data.

These advancements highlight the importance of hybrid FL architectures, where global feature consistency and local adaptability are simultaneously prioritized. By balancing global parameter sharing with client-specific optimizations, these frameworks effectively address data heterogeneity challenges, offering more robust and scalable solutions for medical image segmentation tasks in federated environments.

It is important to highlight that certain federated learning (FL) frameworks have been specifically developed to address the challenges posed by partially labeled datasets in medical image segmentation tasks. For instance, [34] applied the FedAvg algorithm to a U-Net-inspired segmentation architecture (C2FNAS) [35]. This approach leveraged federated parameter aggregation to optimize segmentation performance across decentralized datasets. In another study, a federated model MENU-Net incorporates [14] client-specific encoders and a shared auxiliary decoder. In this architecture, each client utilizes a dedicated encoder to extract feature representations, while the shared decoder integrates and processes these features to generate binary segmentation predictions. MENU-Net employs feature map concatenation from multiple encoders as the input for the shared decoder, enabling a collaborative knowledge-sharing process.

However, a notable limitation across these approaches lies in their reliance on FedAvg, which often results in catastrophic forgetting during local training updates. Specifically, clients tend to lose global model knowledge when performing localized fine-tuning, leading to inconsistent performance across nodes. In contrast to these prior works, our approach introduces a carefully designed baseline architecture that is capable of generating multi-task predictions efficiently without the need to repeatedly perform the feed-forward computation process for each individual task. This efficiency significantly reduces computational redundancy and accelerates model inference across tasks. To address the catastrophic forgetting issue, we incorporate both global knowledge distillation (KD) loss and local KD loss mechanisms. These losses serve as regularization terms during local training, ensuring that the client models retain global knowledge while still adapting effectively to local data distributions. This dual-KD strategy mitigates the loss of essential global features during local updates, resulting in a more stable and high-performing federated segmentation model across diverse and partially labeled datasets.

C. EFFICIENT FINE-TUNING

The rise of foundational models in machine learning has led to significant interest in parameter-efficient fine-tuning techniques for domain-specific tasks. These fine-tuning methodologies can generally be categorized into three primary approaches [36]. The first approach, known as addition-based methods, involves integrating lightweight adapter modules [37], [38] or prompt-based adjustments [39], [40] into the pre-trained model architecture. Fine-tuning in this method focuses exclusively on optimizing the newly added parameters while keeping the majority of the pre-trained weights frozen. This approach reduces computational complexity and allows for efficient task-specific adaptation without retraining the entire model.

The second category, specification-based methods, emphasizes selecting a targeted subset of the model's original parameters for fine-tuning [41], [42]. By focusing only on the

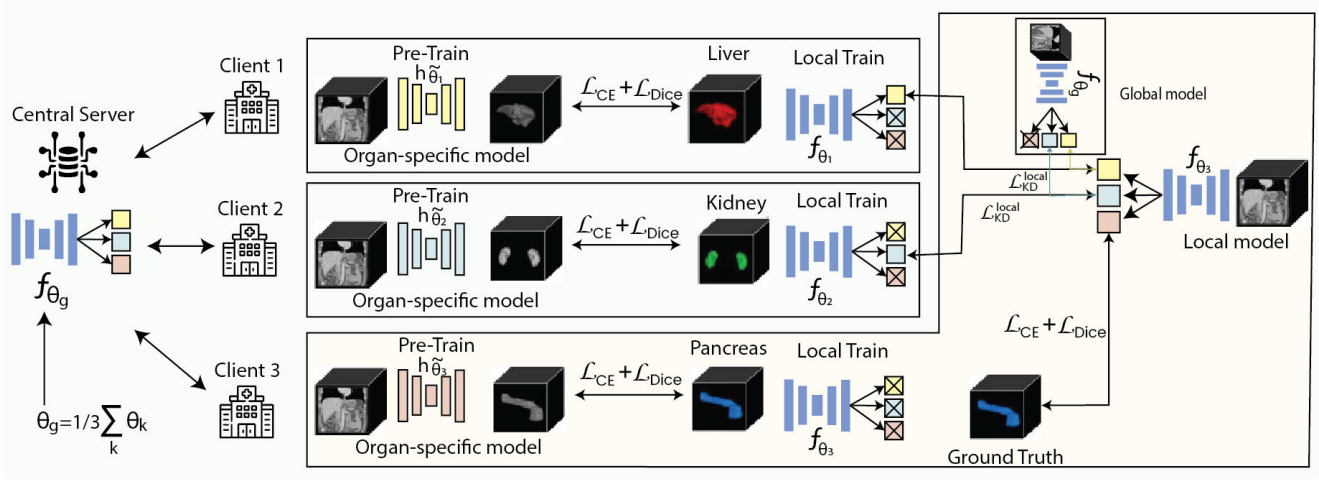

FIGURE 1. Proposed federated learning approach that begins with organ-specific pre-training at each client node (e.g., liver, kidney, pancreas) using cross-entropy ($\mathcal{L}_{CE}$) and Dice loss ($\mathcal{L}_{Dice}$) to optimize segmentation performance. Each client trains a local model ($f_{\theta_1}, f_{\theta_2}, f_{\theta_3}$) using its respective weakly labeled dataset while preserving organ-specific knowledge. The pre-trained organ-specific models ($h_{\tilde{\theta_1}}, h_{\tilde{\theta_2}}, h_{\tilde{\theta_3}}$) are used for local training. During federated aggregation, knowledge distillation loss ($\mathcal{L}_{KD}^{local}$) ensures that local updates retain global knowledge shared via the central server. The global model (f_{θ_g}) aggregates parameters across clients using weighted averaging, creating a unified segmentation model capable of multi-organ predictions. This framework addresses challenges associated with heterogeneous, weakly labeled datasets and catastrophic forgetting, enhancing both global consistency and local adaptability in federated medical image segmentation tasks.

most impactful parameters, this strategy achieves a balance between computational efficiency and fine-tuning accuracy. Instead of updating the entire parameter space, specification-based methods optimize only those parameters most relevant to the target task, making them particularly effective when computational resources are limited.

The third category, reparameterization-based methods, introduces low-rank matrix approximations to represent and optimize parameter updates [43]. By decomposing parameter matrices into low-dimensional representations, this approach significantly reduces the number of trainable parameters while preserving the expressive power of the original model. This reparameterization strategy has shown promise in maintaining fine-tuning performance while minimizing resource overhead.

Recent advancements have extended pre-trained image models to address tasks beyond traditional image processing, including video analysis [38] and volumetric medical image segmentation [44]. However, these adaptations often treat the additional temporal or spatial dimension as a sequential token group, similar to a "word group," and employ specialized aggregation modules to capture contextual dependencies along this additional dimension. While effective in certain scenarios, these methods often overlook the inherently isotropic nature of 3D spatial data present in medical imaging modalities like CT and MRI.

In contrast, our proposed approach fundamentally diverges from these conventional adaptation strategies. Instead of isolating dimensions or relying on external aggregation mechanisms, we treat all three spatial dimensions (X, Y, Z) as isotropic components. This ensures that the transformer blocks are directly adapted to capture and represent 3D spatial patterns without requiring specialized modules for dimensional aggregation. By doing so, the model can inherently learn complex spatial relationships across the volumetric medical data, improving its ability to generalize across diverse segmentation tasks. Furthermore, our methodology aligns with the principles of parameter-efficient fine-tuning, ensuring that computational resources are used optimally without compromising the expressive capabilities of the model.

This integration of 3D spatial adaptation into the transformer architecture addresses many limitations of traditional 2D-to-3D adaptation strategies, offering a more scalable and computationally efficient solution. As a result, our approach holds significant potential for advancing federated knowledge distillation frameworks in medical image segmentation, enhancing both model accuracy and real-world applicability in healthcare environments.

III. METHODOLOGY

A. PROBLEM FORMULATION

Our objective is to develop a unified segmentation model leveraging federated learning across multiple distributed datasets, denoted as $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^{N_c}$, where N_c represents the total number of client nodes or dataset sources. Each client node c^j (where $j = 1, 2, \ldots, N_c$) is associated with a partially annotated dataset $\mathcal{D}^j$. Here, the index j corresponds to both the dataset and its respective client.

Each dataset $\mathcal{D}^j$ may exhibit a unique distribution pattern and comprises M^j 3D medical imaging volumes, denoted as $\mathbf{X}_i^j \in \mathbb{R}^{Z \times H \times W}$ for $i = 1, 2, \ldots, M^j$. In this representation, Z corresponds to the axial dimension, H represents the coronal dimension, and W indicates the sagittal dimension.

Accompanying these images are voxel-wise annotations represented as $\mathbf{Y}_i^j \in \mathbb{R}^{Z\times H\times W}$.

Additionally, datasets may differ in terms of segmentation targets, such as various anatomical structures (e.g., organs or pathological regions). Let $\mathcal{C} = \{1, 2, \ldots, N_t\}$ represent the set of target classes, where N_t denotes the total number of target labels. Each voxel in the annotation maps $\mathbf{Y}_i^j$ corresponds to a subset $\mathcal{T}^j \subseteq \mathcal{C}$, indicating the specific target classes annotated within dataset j.

We design the Federated 3D Knowledge Distillation Network (Fed3D-KDNet), parameterized by Θ, to predict the segmentation masks $\hat{S}(\mathbf{X}_i^j, t)$ for each class t in the input images. The learning objective is to minimize the global segmentation loss $\mathcal{L}(\cdot)$ across all distributed datasets, expressed as:

$$\Theta = \arg\min_{\Theta} \sum_{j=1}^{N_c} \sum_{i=1}^{M^j} \sum_{t\in\mathcal{T}^j} \mathcal{L}\left(\hat{S}(\mathbf{X}_i^j, t), \mathbf{Y}_{i,t}^j\right) \tag{1}$$

Here, $\mathbf{Y}_{i,t}^j$ represents the voxel-level binary segmentation map for the target class t in the dataset $\mathcal{D}^j$.

B. FEDERATED AGGREGATION

Figure 1 illustrates our proposed Fed3D-KDNet strategy, inspired by the Federated Averaging (FedAvg) framework [13], which facilitates the aggregation of locally trained models from multiple distributed datasets. In our approach, the global model, denoted as f_Φ, is initialized with a parameter set Φ and then distributed across participating client nodes. Each client subsequently trains its local model f_{ϕ_k} using its respective dataset $\mathcal{D}^k$. The local model is fine-tuned by minimizing a predefined loss function that evaluates the discrepancy between the predicted segmentation map $f_{\phi_k}(\mathbf{X}_i^k, t)$ and the corresponding ground truth annotation $\mathbf{Y}_{i,t}^k$ for each target class t. Mathematically, this objective can be expressed as:

$$\arg\min_{\phi_k} \sum_{i=1}^{M^k} \sum_{t\in\mathcal{T}^k} \mathcal{L}\left(f_{\phi_k}(\mathbf{X}_i^k, t), \mathbf{Y}_{i,t}^k\right) \tag{2}$$

Existing segmentation approaches for partially labeled datasets often employ a combination of cross-entropy loss and Dice loss [7] to measure the alignment between predicted segmentation outputs and ground truth annotations. The Fed3D-KDNet leverages a hybrid loss function combining cross-entropy loss and Dice loss to address class imbalances and voxel-level inconsistencies. This combined loss can be expressed as:

$$\mathcal{L}\left(f_{\phi_k}(\mathbf{X}_i^k, t), \mathbf{Y}_{i,t}^k\right) = \mathcal{L}_{\text{CE},i} + \mathcal{L}_{\text{Dice},i} \tag{3}$$

where.

$$\mathcal{L}_{\text{CE},i} = -\frac{1}{N_V} \sum_{j=1}^{N_V} \mathbf{Y}_{ij}^k \log f_{\phi_k}(\mathbf{X}_{ij}^k, t) \tag{4}$$

$$\mathcal{L}_{\text{Dice},i} = 1 - \frac{2\sum_{j=1}^{N_V} f_{\phi_k}(\mathbf{X}_{ij}^k, t)\mathbf{Y}_{ij}^k}{\sum_{j=1}^{N_V} f_{\phi_k}(\mathbf{X}_{ij}^k, t) + \sum_{j=1}^{N_V} \mathbf{Y}_{ij}^k} \tag{5}$$

where, $N_V = Z \times H \times W$ represents the total number of voxels in each 3D medical image, $\mathcal{L}_{\text{CE}}$ represents the voxel-wise cross-entropy loss, a commonly used classification loss and $\mathcal{L}_{\text{Dice}}$ measures the overlap between the predicted segmentation mask and the ground truth, which is particularly effective for handling class imbalances.

Upon completing local training, each client shares its optimized model parameters ϕ_k with the global server. The global model parameters Φ are then updated by averaging the local parameters across all participating clients, as defined below:

$$\Phi = \frac{1}{N_c} \sum_{k=1}^{N_c} \phi_k \tag{6}$$

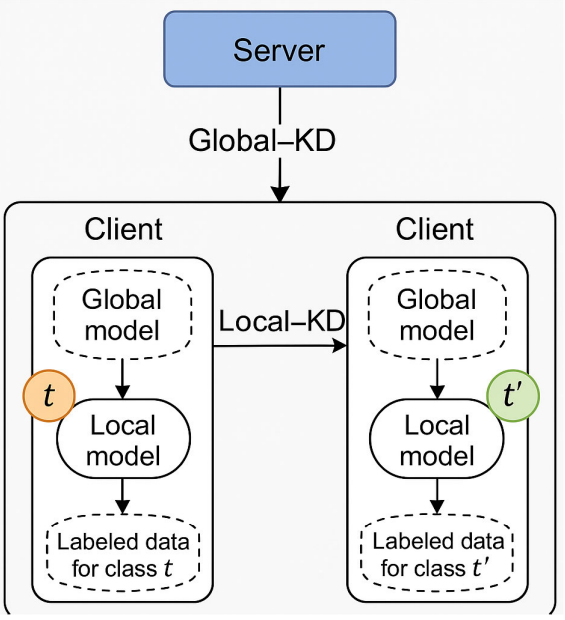

FIGURE 2. Comparative schematic of Global-KD and Local-KD in the Fed3D-KDNet framework. Global-KD aligns local model predictions with the global model outputs to retain shared knowledge, while Local-KD leverages peer models for task-specific guidance across heterogeneous clients.

While traditional FedAvg techniques weigh local contributions based on the size of client datasets, our approach maintains equal weighting due to the heterogeneity in label distributions across clients. This iterative communication and aggregation cycle continues until convergence. The final segmentation model f_Φ represents the global knowledge distilled from all client nodes. The Fed3D-KDNet ensures consistent communication and parameter alignment across all nodes, maintaining an equilibrium between global and local learning objectives.

C. GLOBAL KNOWLEDGE DISTILLATION

Global-KD refers to the mechanism where each client receives predictions from the global model on unannotated regions of their local data and aligns its local model predictions accordingly. This promotes global consistency and mitigates catastrophic forgetting by ensuring each client benefits from the aggregate knowledge of the entire federation, even if their data does not contain all organ annotations. A comparative schematic illustration of the global and local knowledge distillation processes is presented in Figure 2, highlighting their operational differences and interconnections within the Fed3D-KDNet framework. Despite the effectiveness of Federated Averaging (FedAvg) in aggregating client models, the global model often suffers from suboptimal parameter convergence. This issue arises because client-specific models may overfit to locally annotated organ classes, causing them to forget segmentation knowledge related to other organs during localized training.

To address this challenge, we introduce a Knowledge Distillation (KD) loss, which quantifies the divergence between predictions from the global model f_Φ and the local model f_{ϕ_k} on unlabeled regions of the dataset $\mathbf{X}_i^k$. Mathematically, the global KD loss is expressed as:

$$\mathcal{L}_{\text{KD},i}^{\text{global}} = -\frac{1}{N_V}\frac{1}{N_T - n(\mathcal{T}^k)} \sum_{t \notin \mathcal{T}^k} \sum_{j=1}^{N_V} f_\Phi(\mathbf{X}_{ij}^k, t) \log\left(f_{\phi_k}(\mathbf{X}_{ij}^k, t)\right) \tag{7}$$

where, N_V represents the total number of voxels in the image, N_T denotes the total number of target classes across all datasets and $n(\mathcal{T}^k)$ indicates the number of target classes annotated in the local dataset $\mathcal{D}^k$. This ensures that the Fed3D-KDNet maintains global coherence and mitigates catastrophic forgetting, enabling effective multi-organ segmentation across distributed datasets.

The final loss function used for model optimization integrates the KD loss, cross-entropy loss, and Dice loss as follows:

$$\mathcal{L}\left(f_{\phi_k}(\mathbf{X}_i^k, t), \mathbf{Y}_{i,t}^k\right) = \mathcal{L}_{\text{CE},i} + \mathcal{L}_{\text{Dice},i} + \mathcal{L}_{\text{KD},i} \tag{8}$$

This composite loss function ensures that even when a client has annotations for only a limited subset of target organs, the KD loss mitigates catastrophic forgetting by aligning local predictions with the global model's outputs. Consequently, the aggregated knowledge from diverse datasets is preserved across clients, despite data privacy constraints.

D. LOCAL KNOWLEDGE DISTILLATION

Local-KD, on the other hand, enables client nodes to share their organ-specific segmentation models (or distilled predictions) with other clients. Each client incorporates this transferred knowledge to guide training on unlabeled structures, improving generalization across heterogeneous datasets. Schematic KD process of local-KD with

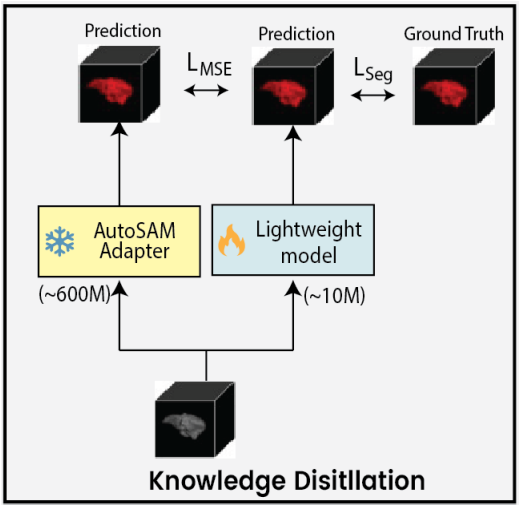

FIGURE 3. Knowledge Distillation framework for medical image segmentation. The AutoSAM model serves as the teacher model, while a Lightweight model acts as the student model. The teacher model generates predictions, which are distilled into the student model through a Mean Squared Error (MSE) loss ($\mathcal{L}_{MSE}$) for aligning feature representations. Additionally, the student model is optimized using a segmentation loss ($\mathcal{L}_{Seg}$) to ensure accurate alignment with the ground truth. This dual-loss strategy ensures that the student model effectively inherits the knowledge from the teacher while maintaining computational efficiency for deployment in resource-constrained environments.

global-KDis s shown in Figure 2. This decentralized distillation promotes collaboration without requiring direct data exchange. KD loss can also be employed at a local level, where each client trains an organ-specific segmentation model and shares it with other clients, as depicted in Figure 3. Before federated aggregation, each client optimizes its organ-specific segmentation model h_{θ_k}, parameterized by θ_k, using the local dataset $\mathcal{D}^k$. The optimized model parameters are then transmitted to all participating clients. During local training, client k integrates predictions from organ-specific segmentation models received from other clients. At the client level, the Fed3D-KDNet facilitates KD, where organ-specific segmentation models share knowledge across clients. This local KD loss function can be expressed mathematically as:

$$\mathcal{L}_{\text{KD},i}^{\text{local}} = -\frac{1}{N_V}\frac{1}{N_T - n(\mathcal{T}^k)} \sum_{t \notin \mathcal{T}^k} \sum_{j=1}^{N_V} h_{\theta_k}(\mathbf{X}_{ij}^k) \log\left(f_{\phi_k}(\mathbf{X}_{ij}^k, t)\right) \tag{9}$$

This approach, however, introduces significant computational overhead due to the increased requirement for training multiple organ-specific models and managing the resulting outputs. Although caching precomputed predictions from h_{θ_k} can partially reduce computational costs, it still imposes substantial memory requirements because of the large-scale voxel-level data involved. The computational burden escalates with an increasing number of target classes.

To alleviate this issue, we adopt a random sampling strategy, where a subset of target classes $t \notin \mathcal{T}^k$ is randomly selected for knowledge distillation. The simplified KD loss is

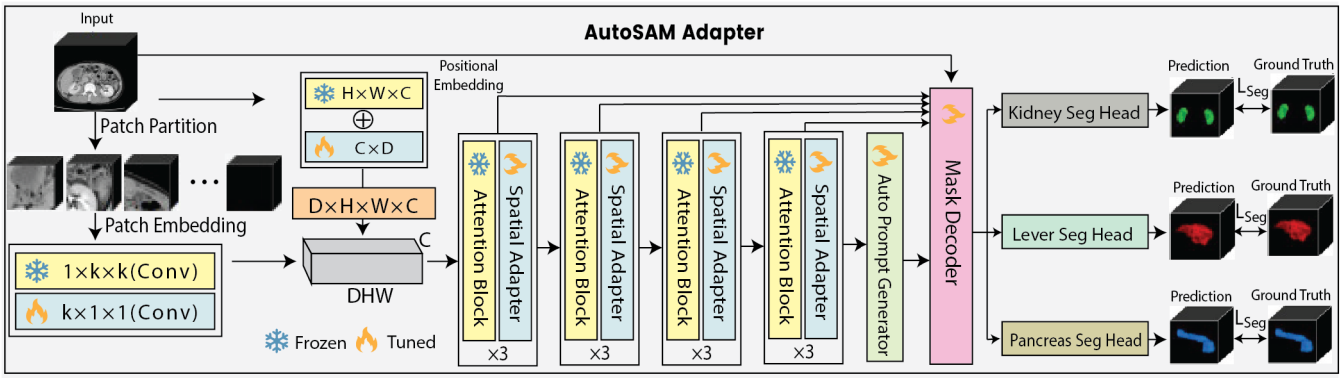

FIGURE 4. Proposed AutoSAM Adapter network architecture for medical image segmentation. The input volumetric data undergoes patch partitioning and embedding, followed by positional embedding. The architecture utilizes alternating frozen and tunable convolutional layers for efficient feature extraction. Attention blocks and spatial adapters are applied across multiple stages to refine the feature representations. An Auto Prompt Generator further enhances adaptability, feeding into a Mask Decoder that generates segmentation outputs for specific organs, including kidney, liver, and pancreas. Each segmentation head is optimized using a segmentation loss ($\mathcal{L}_{Seg}$) by comparing predictions with corresponding ground truth masks. This design balances computational efficiency and segmentation accuracy through selective parameter fine-tuning and spatial attention mechanisms.

defined as:

$$\mathcal{L}_{\mathrm{KD},i}^{\mathrm{local}} = -\frac{1}{N_V}\sum_{j=1}^{N_V} h_{\theta_k}(\mathbf{X}_{ij}^k)\log\left(f_{\phi_k}(\mathbf{X}_{ij}^k, t)\right) \tag{10}$$

Compared to the global KD loss ($\mathcal{L}_{\mathrm{KD}}^{\mathrm{global}}$), the local KD loss ($\mathcal{L}_{\mathrm{KD}}^{\mathrm{local}}$) can be more flexibly adopted in federated learning settings, particularly when data availability and memory resources vary across clients. However, $\mathcal{L}_{\mathrm{KD}}^{\mathrm{local}}$ introduces additional training complexity and memory requirements due to the necessity of maintaining multiple organ-specific models. Conversely, clients with substantial local datasets benefit from localized KD strategies, as they allow more precise organ-specific predictions. Therefore, the local KD loss proves particularly advantageous in scenarios where clients possess abundant annotated data for their target classes, effectively transferring knowledge across models during federated aggregation.

E. GLOBAL-LOCAL DISTILLATION STRATEGY

The integration of both global and KD mechanisms within the Fed3D-KDNet framework is designed to address fundamental shortcomings observed in prior federated medical segmentation approaches, including suboptimal generalization, catastrophic forgetting, and inefficient knowledge transfer in non-IID and partially labeled data settings.

Global-KD serves to preserve cross-site semantic consistency by aligning the local client's predictions with those of the global model on unannotated organ classes. This helps mitigate catastrophic forgetting a common issue in incremental and federated learning by ensuring that local models retain shared knowledge from other client sites even when those labels are absent locally. By encouraging each client to mimic the global teacher's outputs on missing classes, Global-KD bridges information gaps caused by partial annotations and ensures that global knowledge is propagated uniformly across the federation.

In contrast, Local-KD facilitates peer-to-peer knowledge transfer by allowing each client to leverage organ-specific models distilled from other clients. This strategy is especially beneficial in heterogeneous settings where local data distributions and organ annotations vary significantly. By incorporating fine-grained, client-specific insights into local training through teacher-student distillation, Local-KD enhances representation learning on structurally similar yet differently labeled datasets. It complements the coarse alignment enforced by Global-KD with more tailored, organ-specific refinement, improving segmentation quality on complex or underrepresented anatomical structures.

F. SAM FRAMEWORK

SAM [15] is a prompt-based segmentation method renowned for its exceptional accuracy and adaptability in segmenting diverse image types. SAM is structured around three core components: an image encoder, a prompt encoder, and a mask decoder. The image encoder leverages the Vision Transformer (ViT) to transform raw image data into high-dimensional feature embeddings. The prompt encoder processes a variety of user-defined prompts, converting them into compact representations through a fusion of static positional embeddings and dynamic, prompt-specific encodings.

The mask decoder integrates both self-attention mechanisms and bidirectional cross-attention layers to facilitate the alignment between image features and encoded prompts. During this process, feature maps undergo upsampling and pass through a Multi-Layer Perceptron (MLP) to produce precise segmentation masks.Despite its remarkable performance in 2D natural image segmentation, the SAM framework encounters limitations when applied to 3D volumetric medical imaging. Specifically: 1) Its slice-by-slice processing overlooks spatial correlations between adjacent slices. 2) The

domain gap between natural images and medical imaging data hampers generalization.

To address these challenges and optimize SAM for multi-organ medical image segmentation, a combination of model fine-tuning and domain-specific adaptations is essential. This ensures that the segmentation framework captures inter-slice dependencies and adapts effectively to the unique characteristics of medical imaging datasets.

G. HANDLING 3D VOLUMETRIC INPUTS

The Fed3D-KDNet extends the Segment Anything Model (SAM) to efficiently process 3D volumetric medical data, addressing inherent limitations in traditional 2D Transformer architectures. However, it faces inherent limitations when processing 3D medical imaging modalities, such as CT and MRI scans. These data types introduce challenges due to their volumetric nature, which the 2D-centric ViT framework cannot inherently address. Conventional workflows in medical imaging often rely on slice-by-slice analysis, incorporating spatial adaptors or temporal modules to establish contextual consistency. Nevertheless, the underlying architecture remains fundamentally optimized for 2D image processing.

To address these limitations, we propose a 3D adaptation strategy as illustrated in Figure 4 with two primary objectives: 1) Enable the model to directly learn 3D spatial patterns, and 2)Ensure parameter continuity by inheriting pre-trained knowledge from the existing 2D model while introducing incremental trainable parameters.

1) POSITIONAL ENCODING

The pre-trained ViT uses a positional encoding table of size $C \times H \times W$. To accommodate 3D spatial encoding, we extend this table to a tunable lookup table of size $C \times D$ and initialize it with zeros. For each 3D voxel point (d, h, w), we integrate the embeddings of the frozen lookup table corresponding to (h, w) with the embeddings from the tunable lookup table corresponding to (d).

2) PATCH EMBEDDING

We utilize a combination of $1 \times k \times k$ and $k \times 1 \times 1$ 3D convolutional kernels to approximate the functionality of a $k \times k \times k$ convolution (e.g., $k = 14$). The $1 \times k \times k$ convolution leverages pre-trained weights from the 2D model, which remain static during fine-tuning. For the newly introduced $k \times 1 \times 1$ convolution, depth-wise convolutions are applied to reduce the number of trainable parameters, ensuring computational efficiency without compromising performance.

3) ATTENTION BLOCK

The attention mechanism is modified to accommodate 3D volumetric features. In the case of 2D inputs, the query dimensions are denoted as $[B, HW, C]$. These are seamlessly adapted to handle 3D inputs with dimensions $[B, DHW, C]$ while retaining pre-trained weights. Additionally, sliding window mechanisms, akin to those utilized in SwinUNETR [18], are employed to mitigate increased memory consumption associated with 3D feature representations.

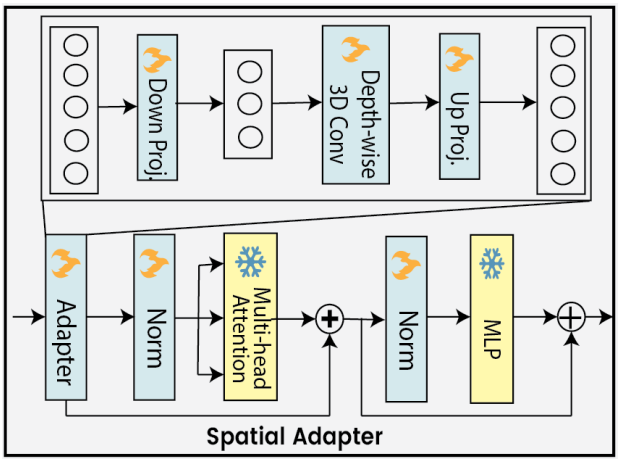

FIGURE 5. Architecture of the Spatial Adapter module. The module incorporates tunable layers for Down Projection, Depth-wise 3D Convolution, and Up Projection, enabling efficient feature refinement. Multi-head Attention enhances spatial contextual understanding, while normalization layers ensure stable training dynamics. The integration of an MLP further refines the aggregated feature representations, facilitating robust spatial feature adaptation for segmentation tasks.

4) BOTTLENECK ADAPTATION

Given the computational efficiency of convolutional layers, we replace 2D convolutions in the bottleneck with their 3D counterparts, training them from scratch to maximize adaptability and improve overall performance. To efficiently adapt the 2D ViT for 3D volumetric data, we introduce a lightweight adapter module. This module comprises of a down-projection and an up-projection linear layer.

These layers can be mathematically represented as:

$$\text{Adapter}(\mathbf{X}) = \mathbf{X} + \text{Act}\big(\mathbf{X}\mathbf{W}_{\text{down}}\big)\mathbf{W}_{\text{up}} \tag{11}$$

where,$\mathbf{X} \in \mathbb{R}^{N \times C}$ denotes the input feature representation, $\mathbf{W}_{\text{down}} \in \mathbb{R}^{C \times N'}$ and $\mathbf{W}_{\text{up}} \in \mathbb{R}^{N' \times C}$ represent down-projection and up-projection layers, respectively and $\text{Act}(\cdot)$ denotes an activation function (e.g., ReLU).

To enhance the model's spatial awareness, we incorporate an additional 3D convolution layer after the down-projection layer as shown Figure 5. During the training phase, only convolutional layers, spatial adaptors, and normalization layers are fine-tuned, while all other pre-trained parameters remain frozen. This selective adaptation minimizes memory consumption and computational overhead. Fine-tuning these targeted components bridges the performance gap between natural and medical images, enabling the model to effectively learn complex 3D spatial patterns and deliver accurate segmentation results.

H. AUTO PROMPT GENERATOR

The novel SAM employs positional embeddings to encode both image and prompt representations. This approach

ensures that both prompts and image embeddings corresponding to the same spatial location share identical positional encodings. These embeddings are subsequently aligned through cross-attention mechanisms, enabling smooth transitions from positional to semantic representations. While effective in 2D segmentation tasks, this cross-attention strategy often encounters limitations when extended to 3D volumetric feature maps. Specifically, over-smoothing effects may emerge due to higher token density in 3D and token proliferation can lead to computational overhead and flattened probability distributions, which impair effective feature learning.

Traditional prompt-based segmentation frameworks face challenges in real-world applications due to two key reasons: 1) Time-Intensive for Multi-Class Segmentation: Generating prompts for multiple anatomical classes or adjacent tissues is computationally expensive and manually intensive. 2)Prompt Quality Dependency: The performance of prompt-based models heavily depends on the accuracy and specificity of the provided prompts. Crafting precise prompts often requires domain-specific expertise, which might not always be available.

Moreover, tasks involving simultaneous multi-class segmentation demand precise, high-quality prompts for each target structure, which becomes particularly challenging for closely located organs or tissues. These limitations significantly affect non-expert users, reducing the practicality and scalability of prompt-driven segmentation in medical applications.

To address these limitations, Fed3D-KDNet introduces an Auto Prompt Generator (APG) as an alternative to static positional encodings for prompt representation. The proposed APG eliminates reliance on manually defined prompts and instead derives embeddings directly from the output feature maps produced after the final attention and spatial adaptation layers (as depicted in Figure 6. This Auto Prompt Generator adopts a fully convolutional neural network FCN architecture, akin to the design of 3D U-Net [25]. The generator integrates Lightweight design, optimizing computational efficiency and 3D convolutional layers, enabling precise volumetric feature extraction. The APG framework is capable of generating spatially-aware, context-specific prompts that are finely tuned for multi-organ medical image segmentation tasks. By eliminating the need for manual prompt crafting, the APG significantly reduces pre-processing overhead and accelerates model inference while maintaining high segmentation accuracy.

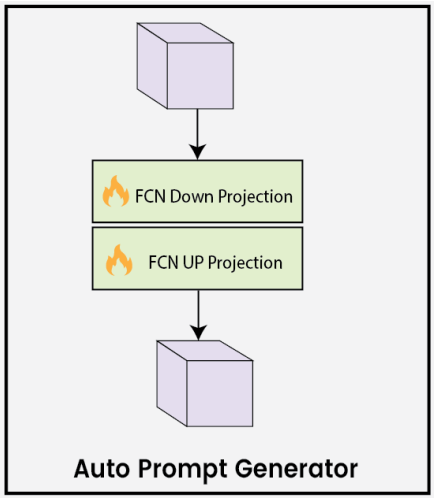

FIGURE 6. Auto Prompt Generator module consists of Fully Connected Network (FCN) Down Projection and FCN Up Projection layers, both denoted with tunable parameters. These layers process the input feature representation through dimensionality reduction (Down Projection) and subsequent expansion (Up Projection), enabling adaptive feature refinement for downstream segmentation tasks.

IV. EXPERIMENT AND RESULT

A. DATASET

1) SPLEEN DATASET

The Spleen dataset originates from the Medical Segmentation Decathlon (MSD) challenge [4]. It consists of 61 abdominal CT volumes, collected from a diverse cohort including both patients and healthy volunteers. Each volume is manually annotated with spleen segmentation masks by experienced radiologists. The dataset is divided into a training subset (N=41) and a test subset (N=20). In this study, only the training subset was utilized for model training and validation. The CT volumes exhibit significant variability in spatial dimensions and voxel spacing, with resolutions ranging from $[0.5 \times 0.5 \times 1.0]$ mm to $[1.0 \times 1.0 \times 2.0]$ mm. The image dimensions typically vary between $512 \times 512 \times [80 - 200]$ voxels. Although other organs, such as the liver, kidneys, and pancreas, are visible in the scans, annotations are exclusively available for the spleen.

2) LIVER DATASET

The Liver dataset, also sourced from the MSD challenge [47], contains 201 contrast-enhanced abdominal CT scans. These volumes were collected from multiple imaging centers, resulting in variability in imaging protocols and scanner configurations. The dataset includes manual annotations for both liver structures and associated lesions. However, in this study, only liver organ annotations were retained for training and validation. The dataset is split into training (N=131) and test (N=70) subsets. The spatial resolution ranges from $[0.6 \times 0.6 \times 1.0]$ mm to $[1.5 \times 1.5 \times 2.5]$ mm, with voxel dimensions varying between $512 \times 512 \times [100 - 300]$. While other organs, including the spleen, pancreas, and kidneys, are visible in the CT scans, annotations are provided exclusively for the liver. This selective annotation increases the complexity of multi-organ segmentation tasks.

3) PANCREAS DATASET

The Pancreas dataset is derived from the pancreas segmentation task of the MSD [47]. It includes 420 abdominal CT volumes collected from multiple institutions, each utilizing distinct scanning protocols and imaging devices. Expert

radiologists provided manual annotations for the pancreas gland, serving as the ground truth for segmentation tasks. The dataset was divided into training (N=282) and test (N=139) subsets, with only the training subset being employed in this study. Spatial resolutions vary between $[0.7 \times 0.7 \times 1.5]$ mm and $[1.25 \times 1.25 \times 2.5]$ mm, and voxel dimensions range from $512 \times 512 \times [120-300]$. Although other abdominal organs are present in the images, only the pancreas gland is annotated. This partial labeling poses challenges in generating robust segmentation models.

4) KIDNEY DATASET

The Kidney dataset originates from the Kidney Tumor Segmentation (KiTS19) challenge [48]. It consists of 210 contrast-enhanced abdominal CT scans collected from multiple clinical centers, each employing varying scanning protocols and hardware. Manual annotations are provided for both kidney structures and associated tumors. However, in this study, only kidney organ annotations were retained, and tumor labels were excluded. The dataset includes training (N=168) and test (N=42) subsets. Spatial resolution ranges from $[0.5 \times 0.5 \times 1.0]$ mm to $[1.25 \times 1.25 \times 2.0]$ mm, while voxel dimensions typically span $512 \times 512 \times [80 - 250]$. Although other organs are visible in the CT scans, annotations are restricted solely to kidney structures.

5) BTCV DATASET

Beyond The Cranial Vault (BTCV) dataset [49] is a widely utilized benchmark for multi-organ segmentation in abdominal computed tomography (CT) scans, particularly in addressing challenges posed by heterogeneous organ structures and partially labeled datasets. The dataset comprises 3D abdominal CT scans sourced from diverse clinical settings, ensuring variability in imaging protocols, anatomical structures, and scan quality. Each scan in the BTCV dataset is meticulously annotated by experienced radiologists, with segmentation masks corresponding to multiple abdominal organs. The dataset focuses on key anatomical structures, including the liver, spleen, pancreas, kidneys, stomach, aorta, inferior vena cava (IVC), portal vein, and gallbladder. These organs are selected due to their clinical significance in various diagnostic and therapeutic workflows. Furthermore, the dataset includes voxel-wise annotations, which are essential for developing high-resolution segmentation models capable of distinguishing subtle anatomical boundaries.

The BTCV dataset is employed as an external benchmark to evaluate the transferability and generalization capabilities of our proposed model. In our experimental setup, the BTCV dataset served as an independent test set to ensure an unbiased assessment of model performance across diverse imaging conditions and anatomical structures. Statistically, the BTCV dataset consists of 50 abdominal CT scans with spatial resolutions typically ranging from 0.5mm to 1.0mm voxel spacing. The scans generally include 85 to 500 axial slices per volume, depending on the acquisition protocol. Out of these, 30 scans from the training subset—annotated with 13 distinct organ labels—were selected for evaluation. These organs include the liver, spleen, kidneys, and pancreas, which were the primary focus of our segmentation task. Each scan provides detailed voxel-wise annotations, enabling precise evaluation of the model's segmentation accuracy across multiple organ structures.

B. DATA PREPROCESSING AND AUGMENTATION

To ensure consistency and compatibility across datasets, we excluded annotations related to tumors, retaining only the organ-level segmentation masks for the liver, spleen, pancreas, and kidneys. This preprocessing step resulted in a dataset characterized by heterogeneous multi-organ data with incomplete and non-overlapping labels across different sources. To standardize the data for model training, all CT volumes were resampled to an *isotropic voxel spacing of* $1.5 \times 1.5 \times 2.0$ *mm* using trilinear interpolation, ensuring uniform spatial resolution across datasets. Additionally, the volumes were reshaped to consistent dimensions of $256 \times 256 \times 128$ *voxels*, facilitating uniform input dimensions for the segmentation model. For *intensity normalization*, voxel intensities within each volume were scaled to a standardized range of *[0, 1]*. This normalization step reduced the impact of intensity variability arising from different imaging protocols and scanner settings.

To enhance the robustness and generalization capabilities of our model, a set of *data augmentation techniques* was applied during the *training phase*. These augmentations included *random flipping*, *random rotation*, and *intensity scaling*, with respective probabilities of *0.1*, *0.1*, and *0.2*. These transformations introduce variability into the training data, mitigating the risk of overfitting and enabling the model to better handle unseen data distributions. For *patch-based sampling*, the input CT volumes were partitioned into *3D patches of size* $128 \times 128 \times 32$, centered on *voxels corresponding to labeled organs*. This strategy ensured focused training on foreground regions of interest while maintaining computational efficiency. During the *patch sampling process*, a *balanced sampling strategy* was adopted, where *foreground and background patches* were randomly selected in a *1:1 ratio*. This balanced approach ensures that the model is equally exposed to both *organ-specific regions (foreground)* and *background regions*, preventing bias towards one class and improving the accuracy of segmentation predictions across the dataset.

C. EVALUATION METRICS

To comprehensively assess the performance and efficiency of our proposed Fed3D-KDNet, we employed a diverse set of quantitative evaluation metrics. These metrics enable an objective comparison of segmentation quality, computational efficiency, and resource utilization across various experimental settings.

The Dice Similarity Coefficient (DSC) measures the overlap between the predicted segmentation mask and the

ground truth mask, providing a statistical measure of spatial agreement. Mathematically, the DSC is defined as:

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \tag{12}$$

where, A represents the set of voxels in the predicted segmentation mask, B represents the set of voxels in the ground truth segmentation mask and $|A \cap B|$ denotes the number of common voxels between A and B. The DSC ranges from 0 to 1, where a higher value indicates better segmentation performance.

The Hausdorff Distance (HD) evaluates the spatial discrepancy between the predicted segmentation boundary and the ground truth boundary. It quantifies the largest distance from a point on the predicted boundary to the nearest point on the ground truth boundary and vice versa. It is mathematically expressed as:

$$HD(A, B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a, b), \sup_{b\in B}\inf_{a\in A} d(b, a)\right\} \tag{13}$$

where, $d(a, b)$ represents the Euclidean distance between two points a and b, sup denotes the supremum (least upper bound) and inf denotes the infimum (greatest lower bound). A lower HD value indicates better segmentation performance, as it reflects reduced boundary mismatch.

Along with other evaluation measures, we also evaluate the model on Inference Time (IT) that measures the average time taken by the model to process a single 3D CT volume during the prediction phase, Computational cost that is quantified using Giga Floating Point Operations per Second (GFLOPs), representing the amount of computational resources required by the model to process one input volume and model complexity that is expressed in *millions (M)* and represents the total learnable weights and biases within the model architecture. A smaller parameter count suggests a lighter model, which is advantageous for deployment on resource-constrained hardware.

D. FEDERATED TRAINING AND OPTIMIZATION STRATEGY

In our Fed3D-KDNet, each client node performs model updates over a *fixed number of iterations* rather than conventional epoch-based training. This adjustment accounts for the variability in *dataset sizes* across different clients. The *federated model aggregation* process is conducted over *1000 communication rounds*, with each local client performing *80 iterations per round* to update its local model before aggregation.

During the *encoding phase*, *feature maps of dimensions [32, 64, 128, 256]* are extracted sequentially, with *progressive spatial dimension reduction* applied at each stage. These feature maps are subsequently *decoded using skip connections* and *upsampling operations*, ensuring the preservation of spatial details necessary for accurate segmentation predictions. To ensure effective parameter initialization, *Kaiming initialization* was applied across all model architectures [50]. For optimization, we employed the *Stochastic Gradient Descent (SGD)* algorithm with an *initial learning rate of 1e-2*, a *momentum of 0.99*, and a *batch size of 2 per iteration*. The training process was conducted on an *NVIDIA RTX A5000 GPU workstation*. The *learning rate schedule* was adapted using a *polynomial decay function* defined as:

$$lr = lr_0\left(1 - \frac{e}{e_{\max}}\right)^{0.9} \tag{14}$$

where e represents the current *communication round* in the federated training cycle, and $e_{\max}$ denotes the *maximum number of rounds*.

Given the *high-resolution nature* of the CT scans, the input images were partitioned into *3D patches of size [64,128,128]*. To ensure that the cropped regions primarily contain *foreground regions of interest*, *80% of the patches were randomly sampled near these regions*. For *data augmentation*, each patch was randomly *rescaled with a ratio ranging from 0.7 to 1.4*, followed by a *50% probability of mirroring across each dimension*. These augmentation techniques aim to improve the model's robustness and generalization to unseen data. During the *inference phase*, predictions were generated using a *sliding window technique* across the entire CT volume. The resulting patch-wise predictions were *aggregated to form a comprehensive prediction map* for the CT scan. The *same patch size* and processing pipeline were maintained during the *testing phase* to ensure consistency across training and inference stages.

a: FEDERATED CONFIGURATION AND CLIENT SELECTION

To evaluate the scalability, robustness, and heterogeneity-handling capabilities of the proposed Fed3D-KDNet, we designed two distinct federated learning configurations comprising 7 and 21 clients, respectively. In the 7-client configuration, each client was assigned a unique dataset containing annotations for a single organ class (e.g., liver, kidney, pancreas, spleen). This setup emulates realistic clinical scenarios where institutions or diagnostic centers may only annotate specific anatomical structures due to domain specialization or resource constraints. Under this setting, Fed3D-KDNet operates in a partially labeled, non-IID environment, relying on its global-local knowledge distillation strategy to learn a unified segmentation model despite incomplete and disjoint label distributions.

To assess the performance of Fed3D-KDNet in a more granular and decentralized setting, the 21-client configuration was constructed by further dividing each of the original 7 datasets into three non-overlapping subsets. Each subset was then assigned to an independent client, thereby increasing the heterogeneity in terms of data volume, organ coverage, and anatomical label sparsity per client. This configuration stresses the model's ability to manage fine-grained inter-client variability, imbalanced sample distributions, and increased communication overhead.

The client selection process was driven by maximizing anatomical and institutional diversity across clients. Datasets were chosen to reflect real-world variability in CT

acquisition protocols, voxel resolution, and segmentation annotations. Heterogeneity was intentionally enforced by ensuring that each client held a mutually exclusive or minimally overlapping subset of the full organ label space. This design challenges Fed3D-KDNet to reconcile fragmented knowledge from disparate sources using its hybrid global-local distillation mechanism and APG, ultimately enabling consistent multi-organ segmentation in federated, privacy-preserving medical environments.

1) DEVICE-SPECIFIC TRAINING HYPER-PARAMETERS

Training hyper-parameters employed on different hardware devices include resource-constrained local servers typically deployed in institutional settings (e.g., hospitals or radiology labs) and cloud-based platforms that support high-performance training. Given the non-uniform hardware resources across clients in federated learning, these hyper-parameter configurations were carefully optimized to ensure efficient training while maintaining convergence stability and model performance across all nodes.

Table 1 summarizes the primary training hyper-parameters used in both settings. On local servers, reduced batch sizes and iteration counts were chosen to accommodate memory constraints, while maintaining the learning dynamics through conservative learning rate schedules and lower momentum. In contrast, cloud platforms leveraged larger batch sizes, higher momentum, and increased local iterations per communication round to exploit the available computational throughput. Importantly, the input patch size and inference strategy were also adapted for each device to balance segmentation accuracy with computational overhead. These settings ensure harmonized training dynamics across heterogeneous clients and avoid model divergence during aggregation. Such device-aware optimization plays a critical role in real-world federated medical deployments, where resource asymmetry is prevalent.

E. EXPERIMENTAL SETUP

All experiments employed Vision Transformer-B (ViT-B) as the backbone architecture for the image encoder, leveraging its capability to effectively capture spatial features across varying image resolutions. To thoroughly assess the performance of our proposed model in a FL environment, we designed an experimental framework comprising multiple client nodes, each trained on a partially labeled dataset. Specifically, the initial experiment involved 7 client nodes, where each client was allocated a distinct subset of the dataset with partial organ annotations. To further validate the scalability and robustness of our method, we extended the evaluation to a 21-client configuration. In this setup, the data from each client was randomly partitioned into three distinct subsets, simulating a more granular and distributed federated learning scenario.

For federated optimization, we implemented and compared our approach with two widely recognized FL algorithms: FedProx [51] and FedScaffold [52]. These algorithms were integrated with our multi-head segmentation baseline architecture, serving as the backbone model for training and aggregation across distributed nodes. During the prediction phase, a binary segmentation map was generated by applying a sigmoid activation function to the output logits, followed by a thresholding operation at 0.5. For cases with a predefined number of target objects, the largest connected component was selected as the final segmentation output. This strategy ensures stability in predictions, particularly in datasets with overlapping or incomplete organ annotations.

In addition to FL experiments, we have conducted performance evaluation on centralized training specifically designed to address the challenges posed by partially labeled datasets. This dual evaluation framework allows us to assess the effectiveness, generalizability, and scalability of our model across both centralized and federated training paradigms.

F. ABLATION STUDY

1) ABLATION STUDY IN FEDERATED SETTING

To comprehensively evaluate the effectiveness of our proposed Fed3D-KDNet, we conducted an ablation study on the BTCV dataset using a federated learning setup with 7 client nodes. This study compares Local Knowledge Distillation (Local-KD), Global Knowledge Distillation (Global-KD), and Fed3D-KDNet across key performance metrics, including Dice Similarity Score (%) and Hausdorff Distance (HD, voxels). These metrics were selected to assess the segmentation accuracy and boundary precision across both organ-level structures and tumor regions in the liver, kidney, pancreas, and spleen. Local-KD focuses on client-specific training, emphasizing localized segmentation performance. However, it often suffers from catastrophic forgetting when integrating knowledge across distributed datasets. In contrast, Global-KD aggregates knowledge from multiple clients into a unified model but struggles to retain fine-grained, client-specific details, especially in regions with irregular anatomical structures such as tumors. Our proposed Fed3D-KDNet integrates both global-KD and local-KD, offering a balanced approach that effectively addresses the limitations of previous methods.

Table 2 presents a comparative analysis of Local-KD, Global-KD, and our proposed Fed3D-KDNet in a federated learning framework, evaluated using the BTCV dataset with 7 clients. In terms of Average Dice Score, Local-KD achieved an overall segmentation accuracy of 79.29%, reflecting moderate performance across different organ and tumor regions. While it performed well in organ-level segmentation tasks, such as 96.40% for the liver organ and 97.47% for the spleen organ, its performance deteriorated significantly in tumor segmentation, with 62.36% for liver tumors and 41.23% for pancreas tumors. This inconsistency indicates that Local-KD struggles to effectively generalize beyond its local training dataset, particularly in regions with complex

TABLE 1. Device-specific training hyper-parameters used in Fed3D-KDNet across heterogeneous computational environments including local servers and cloud platforms.

Device Type	Metric	Setting	Optimization		Patch Config		Inference		Init. Method
			LR	Momentum	Size	Iterations	Overlap (%)	Strategy	
Local Server	Batch Size Optimizer	1 SGD	0.005	0.95	$64 \times 128 \times 128$	50	25%	Sliding Window Poly Decay	Kaiming
Cloud Platform	Batch Size Optimizer	2 SGD	0.01	0.99	$128 \times 128 \times 32$	80	50%	Sliding Window Poly Decay	Kaiming

and irregular anatomical structures, such as tumors. On the other hand, Global-KD achieved an average Dice score of 78.05%, showing slight improvements in organ segmentation compared to Local-KD, with notable scores like 95.37% for the liver organ and 83.80% for the spleen organ. However, its tumor segmentation performance remained suboptimal, as observed in 65.06% for liver tumors and 52.99% for pancreas tumors. These findings suggest that Global-KD, while maintaining better global consistency, fails to capture localized nuances crucial for tumor delineation.

In contrast, Fed3D-KDNet achieved the highest average Dice score of 80.53%, outperforming both Local-KD and Global-KD across all evaluated organs and tumor regions. Remarkably, the model delivered segmentation accuracies of 97.95% for the liver organ, 82.19% for the pancreas organ, and 98.07% for the spleen organ, indicating its robust ability to integrate both global knowledge and client-specific localized features effectively. Furthermore, tumor segmentation accuracy showed significant improvements, with 58.58% for liver tumors and 49.95% for pancreas tumors. These results highlight the model's capacity to mitigate catastrophic forgetting and optimize cross-client knowledge transfer, ultimately leading to superior segmentation outcomes across both organ and tumor regions.

For Average Hausdorff Distance (HD), which measures boundary precision, Local-KD reported an overall average of 13.71 voxels, with significant errors observed in tumor boundaries, such as 32.27 voxels for liver tumors and 36.04 voxels for pancreas tumors. These elevated HD values suggest poor boundary delineation, especially in anatomically irregular regions where accurate edge preservation is critical. Global-KD performed slightly better with an average HD of 13.35 voxels, showing modest improvements in certain regions but still struggling with tumor boundaries, as seen in 40.60 voxels for liver tumors and 26.32 voxels for pancreas tumors. This performance gap indicates that relying solely on global aggregation limits the model's ability to capture fine-grained spatial structures precisely.

In comparison, Fed3D-KDNet demonstrated superior performance with an overall average HD of 11.43 voxels, achieving significantly lower HD values across all evaluated organs and tumor regions. The model achieved 1.26 voxels for kidney organs and 18.15 voxels for pancreas tumors, showcasing its ability to capture organ boundaries while minimizing edge inconsistencies accurately. These results suggest that Fed3D-KDNet effectively balances local geometric accuracy and global contextual alignment, ensuring robust segmentation boundary precision across heterogeneous datasets. From a technical perspective, Local-KD's limitations stem from its reliance on localized training, leading to knowledge silos and suboptimal tumor segmentation. In contrast, Global-KD enhances global consistency but struggles with fine-grained local segmentation details, particularly in tumor regions. Fed3D-KDNet effectively integrates local and global knowledge distillation, overcoming the shortcomings of both approaches. The proposed architecture mitigates catastrophic forgetting, ensures parameter aggregation efficiency, and delivers consistent segmentation accuracy across organs and tumors.

Table 3 presents the results of our ablation study conducted in a federated learning setting using 21 clients on the BTCV dataset, comparing Local-KD, Global-KD, and the proposed Fed3D-KDNet. In terms of segmentation accuracy, Local-KD achieved an average Dice score of 77.15% in the 21-client setting, showing a slight decline compared to its 79.29% score with 7 clients. This drop in performance can be attributed to increased data heterogeneity and the model's inability to efficiently aggregate knowledge across a larger number of distributed datasets. Local-KD, while effective for client-specific fine-tuning, struggles with catastrophic forgetting when attempting to generalize across diverse datasets. Global-KD, on the other hand, attained an average Dice score of 78.40%, showing moderate improvement over Local-KD but failing to match its performance in the 7-client setting (78.05%). While Global-KD benefits from aggregating global knowledge, it lacks the ability to capture fine-grained client-specific features, particularly in complex anatomical regions such as tumors. In contrast, Fed3D-KDNet demonstrated superior segmentation performance with an average Dice score of 79.52%, surpassing both Local-KD and Global-KD. Notably, the segmentation accuracy for kidney organs (97.88%), pancreas organs (81.02%), and spleen organs (96.62%) highlights the model's ability to effectively balance local and global knowledge, even under increased data heterogeneity.

TABLE 2. Ablation study in the federated setting comparing local knowledge distillation (Local-KD), Global knowledge distillation (Global-KD), and the proposed Fed3D-KDNet. evaluation conducted using 7 clients on the BTCV dataset. best value is highlighted with bold.

Method	Metric	Avg Value	Liver		Kidney		Pancreas		Spleen
			Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ
Local-KD	Dice (%)↑	79.29	96.40	62.36	**99.12**	77.55	80.90	41.23	97.47
	HD (voxel)↓	13.71	**2.10**	32.27	1.80	16.84	**5.72**	36.04	**1.13**
Global-KD	Dice (%)↑	78.05	95.37	65.06	94.77	74.51	79.84	52.99	83.80
	HD (voxel)↓	13.35	3.46	40.60	3.59	**10.03**	6.63	26.32	2.82
Fed3D-KDNet (Ours)	Dice (%)↑	**80.53**	**97.95**	**58.49**	98.23	**78.85**	**82.19**	49.95	**98.07**
	HD (voxel)↓	**11.43**	2.16	**31.78**	**1.01**	18.15	6.88	18.83	1.20

TABLE 3. Ablation study in the federated setting comparing local knowledge distillation (Local-KD), global knowledge distillation (Global-KD), and the proposed Fed3D-KDNet. evaluation conducted using 21 clients on the BTCV Dataset. Best value is highlighted with bold.

Method	Metric (Avg)	Avg Value	Liver		Kidney		Pancreas		Spleen
			Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ
Local-KD	Dice (%)↑	77.15	94.26	**60.76**	93.23	73.59	77.14	48.76	92.31
	HD (voxel)↓	18.04	8.43	74.82	3.20	8.43	**5.89**	**22.63**	2.88
Global-KD	Dice (%)↑	78.40	**94.67**	60.37	93.62	73.51	77.14	48.09	92.68
	HD (voxel)↓	17.39	5.87	**41.24**	4.69	**4.86**	6.94	52.72	5.40
Fed3D-KDNet (Ours)	Dice (%)↑	**79.52**	94.34	59.36	**97.88**	**76.24**	**81.02**	**51.18**	**96.62**
	HD (voxel)↓	**16.05**	**2.69**	44.45	**2.35**	16.15	8.83	35.91	**1.97**

Boundary precision, as measured by Hausdorff Distance (HD, voxels), provided further insights into the robustness of the three approaches. In the 21-client setting, Local-KD recorded an average HD of 18.04 voxels, reflecting poor boundary delineation, particularly in liver tumors (74.82 voxels) and pancreas tumors (22.63 voxels). These high HD values indicate that Local-KD struggles to generalize across clients, especially in anatomically irregular or low-contrast regions. Global-KD performed moderately better, with an average HD of 17.39 voxels, demonstrating reduced errors in some regions but still falling short in complex structures such as liver tumors (41.24 voxels) and pancreas tumors (52.72 voxels). In contrast, Fed3D-KDNet achieved the lowest average HD of 16.05 voxels, outperforming both Local-KD and Global-KD across all evaluated organs and tumor regions. Significant improvements were observed in kidney organs (2.35 voxels) and pancreas tumors (16.15 voxels), showcasing the model's ability to maintain fine-grained spatial consistency and precise boundary delineation despite the challenges posed by federated data heterogeneity.

When comparing the results from the 7-client and 21-client experiments, clear patterns emerge. Both Local-KD and Global-KD demonstrated a decline in performance as the number of clients increased, highlighting their scalability limitations. Local-KD, in particular, struggled with increased variability across datasets, leading to performance degradation in both segmentation accuracy and boundary precision. Similarly, Global-KD, while maintaining consistency in certain regions, failed to address client-specific variations effectively, especially in complex tumor regions. Conversely, Fed3D-KDNet showcased remarkable consistency across both experiments, maintaining high segmentation accuracy and minimizing HD values. The marginal drop in performance with the increased number of clients reflects the model's resilience and scalability in federated learning settings.

2) ABLATION STUDY IN CENTRALIZED SETTING

To comprehensively assess the performance of the Fed3D-KDNet, we conducted an ablation study in a centralized setting using the BTCV dataset, as presented in Table 4. In the centralized setting, Fed3D-KDNet achieved an average Dice score of 79.92%, reflecting robust segmentation performance across the target anatomical structures. Among the evaluated organs, the spleen (96.84%), liver (88.91%), and pancreas (80.17%) exhibited high segmentation accuracy, indicating the model's ability to generalize effectively in a controlled, non-distributed training environment. However, tumor segmentation accuracy showed variability, with liver tumors (62.10%) and pancreas tumors (65.34%) achieving lower Dice scores compared to their respective organ counterparts. This disparity can be attributed to the irregular shapes, smaller sizes, and lower contrast of tumors within CT volumes, posing challenges for accurate delineation even in a centralized setting. In terms of HD, Fed3D-KDNet demonstrated exceptional boundary precision with an average HD of 10.09 voxels across all evaluated regions. Notable improvements were observed in kidney organs

TABLE 4. Ablation study evaluating the performance of the proposed Fed3D-KDNet in a centralized setting using the BTCV dataset.

Method	Metric (Avg)	Avg Value	Liver		Kidney		Pancreas		Spleen
			Organ	Tumor	Organ	Tumor	Organ	Tumor	Organ
Fed3D-KDNet (Ours)	Dice (%)↑	79.92	88.91	62.10	88.43	77.64	80.17	65.34	96.84
	HD (voxel)↓	10.09	3.30	43.16	1.89	3.86	3.48	13.36	1.59

(1.89 voxels) and pancreas organs (3.48 voxels), reflecting the model's proficiency in accurately delineating clear and consistent boundaries for organ-level structures. However, tumor regions, particularly liver tumors (43.16 voxels) and pancreas tumors (13.36 voxels), recorded higher HD values, signifying persistent challenges in boundary precision for smaller, irregularly shaped regions.

When compared with the federated setting, certain trends emerge. In the 7-client federated setting, the average Dice score (80.53%) slightly surpassed the centralized setting's performance (79.92%). Similarly, the average HD (11.43 voxels) in the 7-client setup was marginally higher than the centralized HD (10.09 voxels). These results suggest that the federated framework, despite being distributed and exposed to data heterogeneity, effectively aggregated knowledge from diverse clients and maintained competitive performance. On the other hand, the 21-client federated setting exhibited a slight drop in performance, with an average Dice score (79.52%) and HD (16.05 voxels), indicating that the increase in data heterogeneity and inter-client variability slightly impacted the model's ability to generalize seamlessly.

The performance disparity between centralized and federated settings can be primarily attributed to two factors: data heterogeneity and communication constraints in the federated framework. In the centralized scenario, the model benefits from direct access to the entire dataset, enabling it to learn global patterns more effectively. In contrast, federated settings must balance client-specific knowledge with global consistency, and frequent aggregation cycles may introduce minor parameter misalignments. Furthermore, while Fed3D-KDNet in the centralized setting achieved high overall accuracy and boundary precision, it did not exhibit significant gains over the 7-client federated experiment. This observation suggests that Fed3D-KDNet is capable of mitigating the challenges associated with federated learning, such as catastrophic forgetting and parameter divergence, effectively bridging the performance gap between federated and centralized training paradigms.

3) ABLATION STUDY FOR AUTO PROMPT GENERATOR

To quantify the contribution of the Auto Prompt Generator (APG) in our Fed3D-KDNet framework, we conducted a controlled ablation study by comparing three variants: (1) Manual-Prompt-SAM is the baseline model using handcrafted prompts such as 2D point prompts or bounding boxes derived from ground-truth masks, following traditional SAM usage. (2) NoPrompt-SAM is the prompt-free configuration where the SAM-based segmenter operates without any prompt input, relying solely on the frozen image encoder's features. (3) APG-SAM (Ours) is our proposed method employing the AGP to dynamically create context-aware volumetric prompts from the internal attention and adapter features. We conducted experiments using the BTCV dataset under a centralized training setup to isolate the effect of the prompt generation mechanism. The evaluation metrics included Dice Score (%) and Hausdorff Distance (HD, voxels), as shown in Table 5.

TABLE 5. Ablation study for auto prompt generator (APG) on BTCV dataset.

Method	Avg Dice (%)↑	Avg HD (voxel)↓
Manual-Prompt-SAM	74.38	13.72
NoPrompt-SAM	70.15	17.40
APG-SAM (Ours)	**79.92**	**10.09**

The results demonstrate that Manual-Prompt-SAM outperforms the prompt-free configuration by a margin of 4.23% in Dice score and achieves lower HD, indicating the value of prompt guidance in segmentation. However, our APG-SAM outperforms both baselines, achieving a 5.54% improvement over Manual-Prompt-SAM and a 9.77% gain over NoPrompt-SAM in Dice accuracy, while reducing HD by 3.63 and 7.31 voxels, respectively.

These findings validate the efficacy of the APG in automatically generating context-sensitive, volumetric prompts. Unlike handcrafted prompts, which require manual engineering and do not scale across multiple organs, the APG seamlessly adapts to varying anatomical contexts, enhancing segmentation consistency and eliminating user dependence.

4) CONTRIBUTION ANALYSIS OF CORE COMPONENTS

To explicitly connect the superior performance of Fed3D-KDNet to its key methodological innovations, we provide an in-depth analysis of how each component contributes to the improvements observed in the ablation studies.

a: AUTO PROMPT GENERATOR

The APG introduces dynamic, spatially-aware prompt embeddings derived from internal feature maps, replacing the need for manually crafted prompts. This module enables context-specific tokenization for each volumetric scan, enhancing segmentation performance in heterogeneous client settings. Unlike Global-KD and Local-KD, which rely solely on knowledge alignment through logits, APG ensures

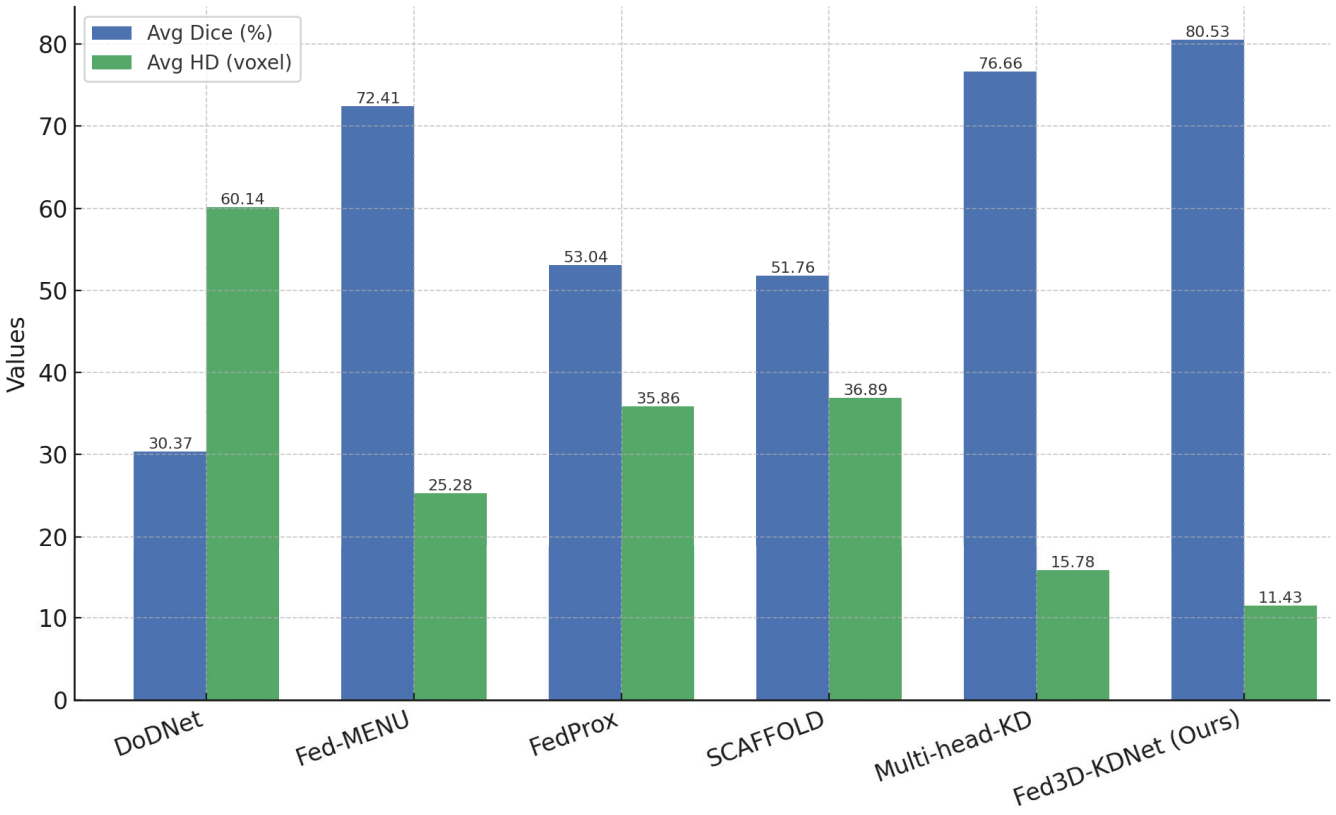

FIGURE 7. Performance comparison of Fed3D-KDNet with state-of-the-art (SoTA) models in the federated setting using 7 clients on the BTCV dataset. The Avg Dice (%) metric (higher values indicate better segmentation accuracy) and the Avg HD (voxel) metric (lower values indicate better boundary delineation) are presented for each model.

more anatomically relevant attention by utilizing features generated in later stages of the segmentation pipeline. As demonstrated in Table 5, APG-SAM achieves a 5.54% improvement in Dice score and a reduction of 3.63 voxels in Hausdorff Distance (HD) compared to manually defined prompts, highlighting its substantial role in enhancing both segmentation accuracy and boundary delineation.

b: 3D SPATIAL ADAPTATION OF SAM

Traditional KD methods optimized for 2D image domains suffer from inter-slice discontinuities when applied to volumetric medical imaging. Fed3D-KDNet addresses this through 3D extensions of SAM, including volumetric patch embeddings, sliding window-based attention, and depth-aware adapter layers. These modifications significantly enhance the model's ability to capture spatial continuity across axial slices. As shown in Tables 2 and 3, Fed3D-KDNet consistently achieves lower HD scores across all organs and tumor classes, indicating more precise anatomical boundary learning compared to standalone Global-KD and Local-KD models.

c: SYNERGISTIC GLOBAL-LOCAL KNOWLEDGE DISTILLATION

The integration of both Global-KD and Local-KD within Fed3D-KDNet enables robust knowledge transfer across clients with non-overlapping organ labels. Global-KD enforces global consistency by aligning predictions on unannotated classes, while Local-KD complements this by transferring fine-grained, organ-specific details via peer-to-peer distillation. The performance gap between the hybrid model and either distillation strategy alone becomes more evident as the federation scales. In the 21-client setting (Table 3), Fed3D-KDNet outperforms Global-KD by 1.12% in Dice and reduces HD by 1.34 voxels, demonstrating that the dual mechanism effectively bridges the trade-off between generalization and specificity.

These results collectively affirm that the performance improvements observed in Fed3D-KDNet are not incidental but stem directly from the carefully designed architectural and training innovations. The integration of APG, 3D-aware SAM adaptations, and dual-stream knowledge distillation allows for scalable, efficient, and anatomically consistent multi-organ segmentation in federated learning environments.

Empirical ablation studies (Tables 2, 3) confirm that neither Global-KD nor Local-KD alone sufficiently overcomes the limitations of heterogeneity and partial labeling. However, their hybridization within Fed3D-KDNet significantly improves both segmentation accuracy and boundary precision, demonstrating their synergistic effect. This dual mechanism enables the framework to maintain global

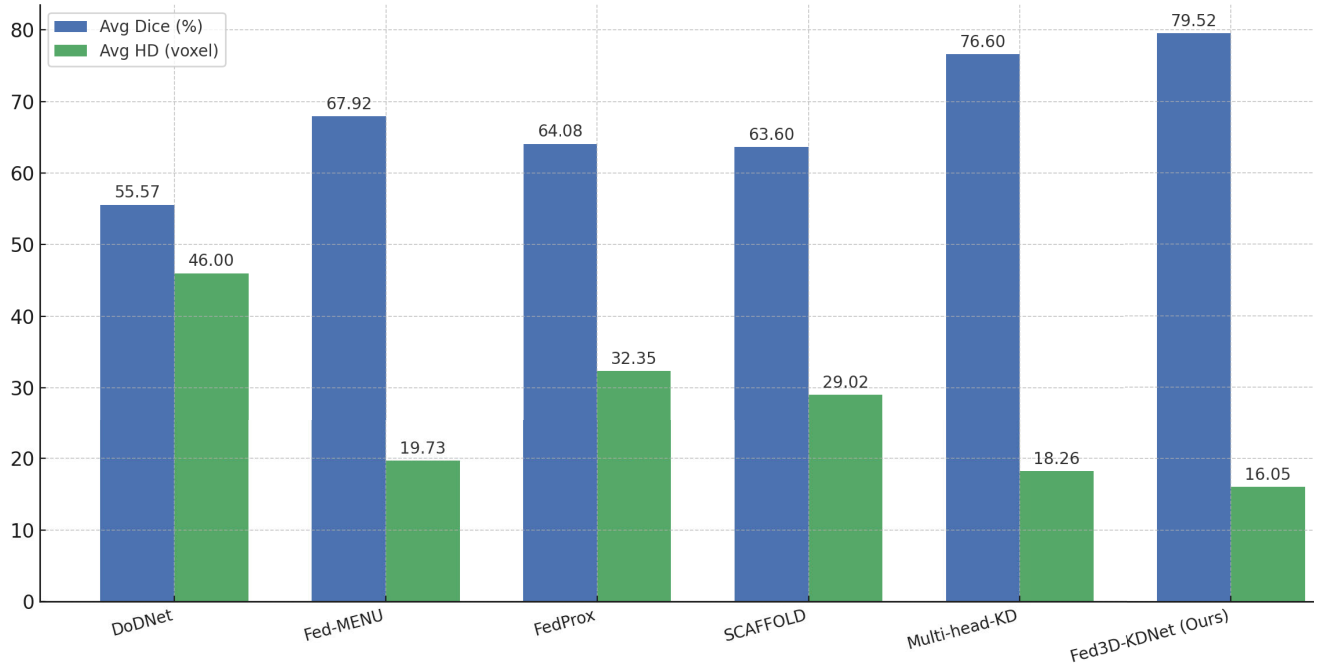

FIGURE 8. Performance Comparison of Fed3D-KDNet with State-of-the-Art (SoTA) Models in the Federated Setting Using 21 Clients on the BTCV Dataset. The Avg Dice (%) values (blue bars, higher is better) and Avg HD (voxel) values (green bars, lower is better) are compared across six segmentation models.

coherence while preserving client-specific expertise, thereby addressing both the scalability and generalization limitations highlighted in prior literature.

G. PERFORMANCE COMPARISON

We compare Fed3D-KDNet with various state-of-the-art (SoTA) segmentation models on the BTCV dataset. Dynamic On-Demand Network (DoDNet) [7] model specifically designed to address the partially labeled dataset problem through a shared encoder-decoder architecture and an adaptive dynamic head. DoDNet employs task-specific kernel generation for each segmentation task, optimizing computational efficiency and flexibility across multi-organ and tumor segmentation benchmarks. Federated Multi-Encoding U-Net (Fed-MENU) [14] employs a multi-encoder architecture for organ-specific feature extraction and an Auxiliary Generic Decoder to enhance feature distinctiveness, enabling efficient federated learning across heterogeneous datasets with non-overlapping region-of-interest annotations. FedProx [51], a widely recognized federated optimization framework designed to address statistical and system heterogeneity in distributed networks. FedProx introduces a proximal term in the local objective function, allowing for variable local updates across participating clients, thereby improving convergence stability and robustness in heterogeneous federated learning scenarios. SCAFFOLD [52] address client-drift caused by data heterogeneity across distributed clients. SCAFFOLD employs stochastic controlled averaging by introducing client-specific control variates to correct local updates, effectively reducing the variance in client contributions and improving convergence stability in non-IID federated settings. Multi-head Knowledge Distillation (Multi-head-KD) [8] integrates both global and local knowledge distillation techniques. This approach leverages a shared encoder-decoder architecture inspired by U-Net and employs lightweight segmentation heads tailored for each target organ, addressing challenges such as catastrophic forgetting and efficiency in partially labeled federated datasets.

Figure 7 presents a comparative analysis of Fed3D-KDNet against state-of-the-art (SoTA) segmentation models, including DoDNet, Fed-MENU, FedProx, SCAFFOLD, and Multi-head-KD, evaluated on the BTCV dataset using a federated learning setup with seven clients. The evaluation metrics include Avg Dice (%)↑ for segmentation accuracy and Avg HD (voxel)↓ for boundary precision. DoDNet, despite its dynamic kernel generation for task-specific segmentation, achieved the lowest performance with 30.37% Avg Dice and 60.14 Avg HD, indicating poor segmentation accuracy and substantial boundary inconsistencies. Fed-MENU showed notable improvement with 72.41% Avg Dice and 25.28 Avg HD, benefitting from multi-encoder feature extraction. However, it struggled with computational complexity and feature consistency across clients. FedProx, designed to address statistical and system-level heterogeneity, achieved 53.04% Avg Dice and 35.86 Avg HD, reflecting moderate accuracy and suboptimal boundary precision. SCAFFOLD, which mitigates client drift using stochastic controlled averaging, demonstrated slightly lower performance with

TABLE 6. Model efficiency comparison of Fed3D-KDNet with SoTA models in a federated setting. inference time represents the model updating duration for a single iteration to generate predictions across all tasks.

Model	GFLOPs	Tuned Params (M)	Time (s)
DoDNet [7]	458.4	45.81	0.098
Fed-MENU [14]	3219.3	128.54	0.427
FedProx [51]	531.8	58.78	0.198
SCAFFOLD [52]	2254.4	83.25	0.360
Multi-head-KD [8]	461.0	47.28	0.097
Fed3D-KDNet (Ours)	**371.3**	**26.53**	**0.058**

51.76% Avg Dice and 36.89 Avg HD, highlighting challenges in fine-grained spatial feature representation. Multi-head-KD integrated global and local knowledge distillation techniques, achieving 76.66% Avg Dice and 15.78 Avg HD, marking significant improvements in both segmentation accuracy and boundary precision. In comparison, our proposed Fed3D-KDNet outperformed all baseline models with an Avg Dice of 80.53% and an Avg HD of 11.43 voxels, surpassing multi-head-KD by 5.04% in Avg Dice and 4.35 voxels in Avg HD. This substantial improvement can be attributed to the hybrid knowledge distillation strategy, which harmonizes both local and global knowledge, and the robust feature extraction capability of the model. Furthermore, Fed3D-KDNet mitigates catastrophic forgetting, ensures better adaptation to non-IID data, and maintains boundary precision across clients.

In Figure 8, the performance comparison of Fed3D-KDNet with state-of-the-art (SoTA) segmentation models is presented in a federated setting using 21 clients on the BTCV dataset. The results demonstrate that our proposed Fed3D-KDNet achieves superior performance across both evaluation metrics: Avg Dice (%) and Avg HD (voxel). Specifically, Fed3D-KDNet attains the highest Avg Dice score of 79.52%, indicating its strong capability in accurate voxel-level segmentation, and the lowest Avg HD value of 16.05 voxels, reflecting its precise boundary delineation and reduced irregularity in segmentation results. Among the comparative models, Multi-head-KD emerges as the closest competitor with an average dice of 76.60% and an Avg HD of 18.26 voxels, showcasing its efficacy in leveraging knowledge distillation techniques. However, it still falls short of the performance exhibited by Fed3D-KDNet, highlighting the latter's optimization in both accuracy and boundary regularization. Fed-MENU achieves a moderate performance with an Avg Dice of 67.92% and an Avg HD of 19.73 voxels, benefiting from its multi-encoder architecture but unable to match the holistic knowledge integration achieved by Fed3D-KDNet. FedProx and SCAFFOLD show slightly lower performance, with Avg Dice scores of 64.08% and 63.60%, and Avg HD values of 32.35 voxels and 29.02 voxels, respectively. These results indicate that while both frameworks address statistical heterogeneity and client drift, they are less effective in optimizing segmentation precision. Finally, DoDNet performs the weakest, with an Avg Dice score of 55.57% and an Avg HD of 46.00 voxels, underscoring the limitations of its dynamic task-specific kernel generation in the federated learning scenario. The results validate the effectiveness of our integrated local and global knowledge distillation approach in enhancing segmentation accuracy and structural consistency across heterogeneous datasets.

H. EFFICIENCY COMPARISON

Table 6 presents a comprehensive evaluation of model efficiency in a federated learning setting. The comparison includes three critical metrics: GFLOPs, representing computational complexity; Tuned Parameters (in millions), indicating model size and resource requirements; and Inference Time (in seconds), reflecting the time required for a single iteration to produce predictions across all tasks. Fed3D-KDNet demonstrates high efficiency, achieving the lowest computational cost with 371.3 GFLOPs, significantly outperforming resource-intensive models such as Fed-MENU (3219.3 GFLOPs) and SCAFFOLD (2254.4 GFLOPs). This reduced computational overhead highlights the optimized architecture of Fed3D-KDNet, enabling faster and more efficient processing without compromising performance. In terms of model size, Fed3D-KDNet maintains the smallest footprint with 26.53 million tuned parameters, showcasing a substantial reduction compared to Fed-MENU (128.54M) and SCAFFOLD (83.25M). This compact design reduces memory requirements, making it suitable for deployment in resource-constrained federated environments. Regarding inference time, Fed3D-KDNet achieves the fastest prediction speed at 0.058 seconds per iteration, marking a 40.2% improvement over Multi-head-KD (0.097s) and an 85.4% reduction compared to Fed-MENU (0.427s). This remarkable efficiency is attributed to the streamlined architecture and effective integration of knowledge distillation and task-specific optimization strategies. Overall, Fed3D-KDNet excels in balancing computational efficiency, model size, and inference speed, establishing itself as a highly efficient solution for multi-organ segmentation tasks in federated learning settings. These results reinforce its practicality and scalability for deployment across heterogeneous client environments.

I. QUANTITATIVE RESULT

Figure 9 illustrates the comparative segmentation performance of various federated learning approaches—namely

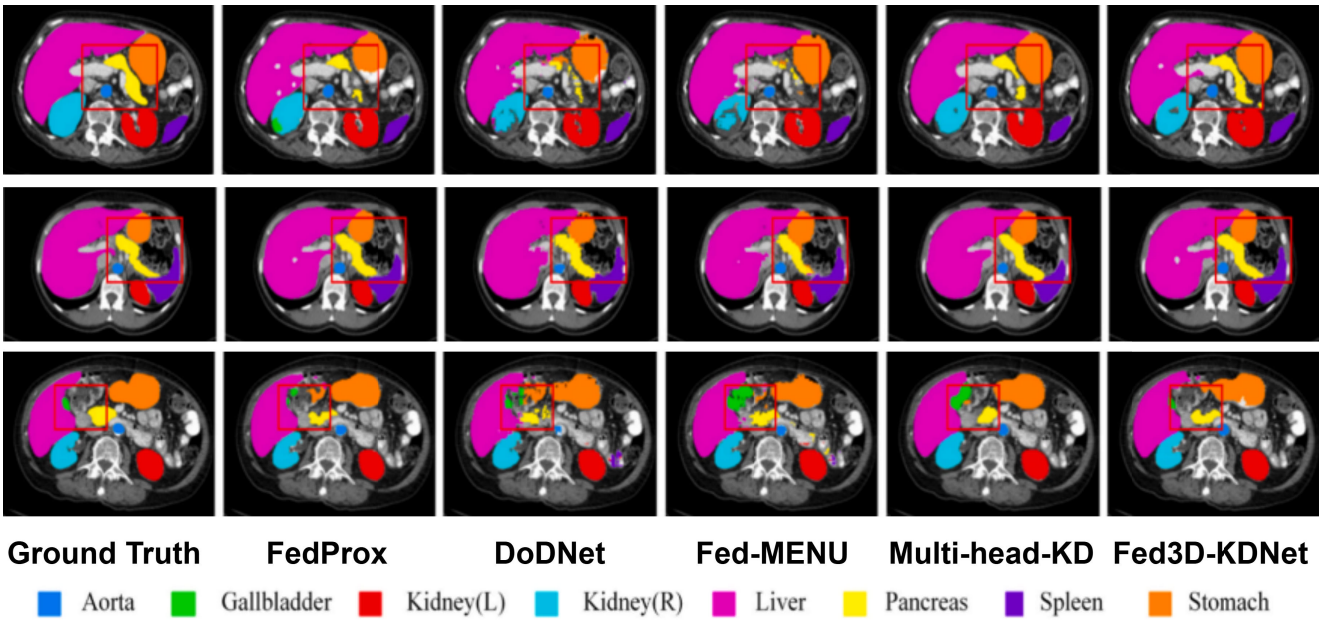

FIGURE 9. Qualitative comparison of multi-organ segmentation results on the BTCV dataset. From left to right: Ground Truth, FedProx, DoDNet, Fed-MENU, Multi-head KD, and the proposed Fed3D-KDNet. The results demonstrate that Fed3D-KDNet achieves superior boundary delineation and organ consistency, especially for challenging anatomical structures such as the pancreas and gallbladder (highlighted by red boxes). Each organ is color-coded as shown in the legend, enabling visual assessment of segmentation accuracy across methods.

FedProx, DoDNet, Fed-MENU, Multi-head-KD, and our proposed Fed3D-KDNet on the BTCV dataset. The visualization includes three representative axial CT slices from different patients. Each column presents the segmentation output of one method, with the leftmost column displaying the ground truth annotations. Organ structures such as the liver, spleen, pancreas, kidneys, stomach, aorta, and gallbladder are color-coded consistently across all rows to facilitate direct comparison.

Highlighted by red bounding boxes, regions with complex anatomical structures (e.g., pancreas, stomach, gallbladder) pose challenges for accurate delineation. The segmentation predictions generated by Fed3D-KDNet demonstrate superior anatomical consistency, clearer organ boundaries, and minimal false positives compared to all baseline methods. Specifically, Fed3D-KDNet excels in capturing fine structural details of the pancreas and gallbladder, which are often missed or poorly segmented by DoDNet and FedProx. The outputs from Fed-MENU and Multi-head-KD show partial success but still suffer from boundary ambiguity and label leakage, especially in densely packed regions.

The superior performance of Fed3D-KDNet can be attributed to its hybrid knowledge distillation strategy, which combines both global and local contextual information to enhance segmentation fidelity across clients. Additionally, the 3D spatial adaptation of the SAM backbone and the Auto Prompt Generator allow for robust volumetric representation and class-aware attention, further improving model precision in inter-organ boundaries. These qualitative observations corroborate the quantitative metrics reported in Tables 2 and 3, reaffirming the effectiveness of our method in challenging federated segmentation scenarios.

1) SCALABILITY ANALYSIS

As FL systems are deployed across increasingly distributed environments, evaluating the scalability of the proposed Fed3D-KDNet in terms of computational and memory overhead is critical. We performed extensive experiments in both 7-client and 21-client configurations, which allowed us to analyze system behavior under varying degrees of federation.

From a computational perspective, as the number of clients increases, the communication and local training workload scales linearly due to more frequent client updates and increased volume of model exchanges. However, Fed3D-KDNet mitigates this overhead by adopting parameter-efficient fine-tuning modules and a hybrid knowledge distillation strategy that avoids full-model retraining at each client. The proposed APG, designed with lightweight 3D convolutional layers and minimal learnable parameters, further contributes to reducing per-client computational cost. Empirical results in Table 6 show that Fed3D-KDNet achieves a 40.2% faster inference time and consumes 43.9% fewer GFLOPs compared to Multi-head-KD, while scaling to 21 clients without significant degradation in performance (see Table 3).

Regarding memory trade-offs, knowledge distillation modules introduce overhead due to intermediate feature storage and teacher-student prediction alignment. This is partially alleviated through (1) random sampling of distillation targets

in local-KD (Equation 13) and (2) caching logits for global-KD, both of which reduce the number of memory-intensive forward passes. The memory footprint also remains bounded since each client processes only organ-specific labels relevant to its dataset. Additionally, the model's design such as 3D adapter modules with frozen base encoders—ensures that memory consumption does not scale proportionally with the number of clients.

J. BIAS AND FAIRNESS IN FEDERATED SETTINGS

FL introduces inherent ethical and fairness challenges due to the heterogeneity of data across clients. In the medical domain, data imbalances stem from variations in imaging modalities, population demographics, clinical protocols, and anatomical prevalence. Such non-independent and identically distributed characteristics may lead to biased model behavior, where predictions favor overrepresented client populations or anatomical regions. For instance, a segmentation model trained predominantly on liver datasets may underperform on underrepresented organs such as the pancreas or spleen.

To mitigate these biases, Fed3D-KDNet incorporates a hybrid knowledge distillation strategy that integrates both global and local information during training. This dual distillation ensures equitable knowledge propagation across all client nodes, even in the presence of partially labeled or class-imbalanced datasets. Additionally, we adopt balanced patch sampling and class-aware loss functions (e.g., Dice loss) to further reduce prediction skewness. Future extensions may explore fairness-aware FL optimizers and adaptive aggregation strategies that assign dynamic weights based on client diversity and performance parity.

K. PRACTICAL APPLICATION SCOPE AND SIGNIFICANCE

The proposed Fed3D-KDNet offers a highly practical solution for real-world clinical environments where data privacy, infrastructure heterogeneity, and annotation scarcity present major barriers to AI adoption. By enabling federated multi-organ segmentation across partially labeled datasets, Fed3D-KDNet can be directly integrated into collaborative networks of hospitals and diagnostic centers without requiring patient data to leave institutional boundaries. Its hybrid knowledge distillation strategy (combining global and local KD) ensures that organ-specific expertise from one institution can benefit other sites—even in the absence of shared annotations—making it highly applicable in decentralized healthcare systems. The incorporation of a 3D spatial adaptation of the SAM allows the model to handle complex volumetric data common in CT and MRI scans, while the APG removes the need for radiologist-defined prompts, reducing dependency on expert intervention. This makes Fed3D-KDNet suitable for use in real-time workflows such as AI-assisted radiology reporting, organ localization for surgical planning, and follow-up monitoring in longitudinal imaging studies. Furthermore, the lightweight and parameter-efficient design ensures that the model can be deployed on low-resource edge devices, enabling scalable AI diagnostics in rural clinics, mobile imaging units, or telemedicine platforms where computational capacity is limited. Overall, Fed3D-KDNet bridges the gap between high-performance segmentation and practical deployability in diverse, real-world medical settings.

L. COMPLIANCE WITH MEDICAL DATA REGULATIONS

Given the sensitive nature of medical imaging data, our federated learning framework adheres to established data protection standards, such as the General Data Protection Regulation and the Health Insurance Portability and Accountability Act. By design, Fed3D-KDNet does not involve direct data sharing across institutions; instead, only model parameters are exchanged, preserving patient privacy and minimizing the risk of data leakage.

All client data remains locally stored, and model updates are transmitted through secure communication protocols. Furthermore, the use of knowledge distillation at both global and local levels ensures that private data distributions are not inferable from shared parameters. Although our work complies with general federated learning privacy standards, integrating differential privacy and secure aggregation mechanisms could offer additional layers of protection in future deployments, particularly in clinical environments involving regulatory audits and cross-border data governance.

M. MODEL LIMITATIONS AND FUTURE DIRECTIONS

Despite the demonstrated effectiveness of the proposed Fed3D-KDNet framework in federated multi-organ medical image segmentation, several limitations should be acknowledged to guide future research and practical deployments.

First, while Fed3D-KDNet exhibits stable convergence and performance across up to 21 clients, its scalability to larger federated settings—such as nationwide hospital networks or multi-institutional research collaborations—poses non-trivial challenges. Specifically, increasing the number of clients amplifies the communication bandwidth requirements, parameter synchronization latency, and global consistency maintenance during knowledge distillation. In such settings, the burden introduced by iterative local updates, parameter aggregation, and dual-stream distillation (Global-KD and Local-KD) can compromise both training efficiency and system responsiveness. Although parameter-efficient tuning and random target selection partially alleviate this, large-scale deployment may necessitate communication compression protocols, asynchronous aggregation mechanisms, or client clustering strategies to maintain scalability and fault tolerance.

Second, the computational overhead introduced by Local-KD remains a practical constraint, especially in resource-constrained environments such as rural clinics, mobile imaging units, or edge-based diagnostic systems. Since Local-KD requires each client to store and integrate multiple peer-specific organ-specific models or logits for distillation, the memory footprint grows linearly with the number of segmentation classes and participating clients.

Furthermore, as the number of distillation targets increases, so does the inference latency and training complexity, potentially impeding real-time applications. While logit caching and random distillation targets provide some mitigation, these approximations may degrade performance when anatomical variability is high. Future work could consider distillation-aware model pruning, shared representation banks, or teacher-student co-distillation strategies to minimize computational cost without sacrificing segmentation accuracy.

Third, the generalizability of the proposed framework to unseen domains, imaging modalities, and rare clinical scenarios has yet to be fully validated. The current evaluation focuses primarily on CT-based datasets with common abdominal organs (liver, spleen, kidney, pancreas). However, many real-world settings involve heterogeneous modalities (e.g., MRI, PET, ultrasound), underrepresented anatomical structures (e.g., adrenal glands, small bowel), or low-resource regions with limited training samples and differing disease distributions. Additionally, domain shift effects—such as variations in scanner vendors, acquisition protocols, and patient demographics—may introduce performance degradation when the model is deployed in unseen environments. Although the APG and 3D SAM adaptation enhance contextual awareness and prompt independence, they are still conditioned on training domain statistics. Cross-modality, cross-institutional, and cross-population generalization thus remain key challenges.

Finally, the current framework assumes that data annotations at each client site are reasonably accurate and class-specific, which may not always hold in practice. Medical annotations often vary in quality due to radiologist subjectivity, institutional guidelines, or labeling errors. The lack of explicit mechanisms to handle noisy labels or inter-annotator variability may limit robustness in diverse clinical settings.

To address these limitations, future research will focus on: (1) Communication-efficient FL strategies, including lossy gradient quantization, hierarchical aggregation, and asynchronous update schedules to improve scalability; (2) Lightweight and modular distillation mechanisms, such as dynamic teacher selection or attention-based knowledge filtering; (3) Self-supervised and semi-supervised learning techniques to reduce dependence on extensive labeled data and improve robustness in under-annotated regions; and (4) Domain-adaptive and fairness-aware learning modules, incorporating feature alignment, bias mitigation, and adaptive client weighting to support ethical and inclusive AI deployment across heterogeneous populations and clinical infrastructures.

V. CONCLUSION

In this study, we proposed Fed3D-KDNet, an innovative Federated 3D Knowledge Distillation Network designed for multi-organ medical image segmentation across heterogeneous and partially labeled datasets within the FL environment. The model addresses critical challenges, such as catastrophic forgetting, data heterogeneity, and computational inefficiency, which often limit the performance and scalability of conventional FL frameworks in medical imaging applications. The proposed Fed3D-KDNet introduces a hybrid knowledge distillation strategy, integrating both global and local knowledge distillation mechanisms. This approach ensures knowledge retention across distributed client datasets while maintaining consistency during global aggregation. Furthermore, the incorporation of 3D spatial adaptations into the SAM framework enables effective handling of volumetric medical imaging data, overcoming the limitations of traditional 2D-centric architectures. A key innovation of our methodology is the Auto Prompt Generator, which eliminates reliance on manual prompts and introduces automated spatially-aware prompt embeddings. This enhancement streamlines the segmentation pipeline and improves model adaptability across diverse anatomical regions. Additionally, the model leverages parameter-efficient fine-tuning strategies, incorporating spatial adapters and selective layer optimization to reduce computational overhead without compromising performance.

REFERENCES

[1] A. Kanhere, P. Kulkarni, P. H. Yi, and V. S. Parekh, "Privacy-preserving collaboration for multi-organ segmentation via federated learning from sites with partial labels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 2380–2387.

[2] M. Jiang, H. R. Roth, W. Li, D. Yang, C. Zhao, V. Nath, D. Xu, Q. Dou, and Z. Xu, "Fair federated medical image segmentation via client contribution estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16302–16311.

[3] J. Dai, H. Wu, H. Liu, L. Yu, X. Hu, X. Liu, and D. Geng, "FedATA: Adaptive attention aggregation for federated self-supervised medical image segmentation," *Neurocomputing*, vol. 613, Jan. 2025, Art. no. 128691.

[4] J. Ma et al., "Fast and low-GPU-memory abdomen CT organ segmentation: The flare challenge," *Med. Image Anal.*, vol. 82, Nov. 2022, Art. no. 102616.

[5] S. Kim, S. An, P. Chikontwe, and S. H. Park, "Bidirectional RNN-based few shot learning for 3D medical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 3, pp. 1808–1816.

[6] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102680.

[7] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "DoDNet: Learning to segment multi-organ and tumors from multiple partially labeled datasets," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1195–1204.

[8] S. Kim, H. Park, M. Kang, K. H. Jin, E. Adeli, K. M. Pohl, and S. H. Park, "Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets," *Med. Image Anal.*, vol. 95, Jul. 2024, Art. no. 103156.

[9] M. H. Vu, G. Norman, T. Nyholm, and T. Löfstedt, "A data-adaptive loss function for incomplete data and incremental learning in semantic image segmentation," *IEEE Trans. Med. Imag.*, vol. 41, no. 6, pp. 1320–1330, Jun. 2022.

[10] J.-W. Xiao, C.-B. Zhang, J. Feng, X. Liu, J. van de Weijer, and M.-M. Cheng, "Endpoints weight fusion for class incremental semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7204–7213.

[11] L. Li, Y. Fan, M. Tse, and K. Y. Lin, "A review of applications in federated learning," *Comput. Ind. Eng.*, vol. 149, Nov. 2020, Art. no. 106854.

[12] Y. Xia, D. Yang, W. Li, A. Myronenko, D. Xu, H. Obinata, H. Mori, P. An, S. Harmon, E. Turkbey, B. Turkbey, B. Wood, F. Patella, E. Stellato, G. Carrafiello, A. Ierardi, A. Yuille, and H. Roth, "Auto-FedAvg: Learnable federated averaging for multi-institutional medical image segmentation," 2021, *arXiv:2104.10195*.

[13] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Artif. Intell. Statistics*, Apr. 2017, pp. 1273–1282.

[14] X. Xu, H. H. Deng, J. Gateno, and P. Yan, "Federated multi-organ segmentation with inconsistent labels," *IEEE Trans. Med. Imag.*, vol. 42, no. 10, pp. 2948–2960, Apr. 2023.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[16] C. Zhang, L. Liu, Y. Cui, G. Huang, W. Lin, Y. Yang, and Y. Hu, "A comprehensive survey on segment anything model for vision and beyond," 2023, *arXiv:2305.08196*.

[17] A. Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H. R. Roth, and D. Xu, "UNETR: Transformers for 3D medical image segmentation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1748–1758.

[18] Y. Tang, D. Yang, W. Li, H. R. Roth, B. Landman, D. Xu, V. Nath, and A. Hatamizadeh, "Self-supervised pre-training of Swin transformers for 3D medical image analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20698–20708.

[19] C. Li, Y. Qiang, R. Ibn Sultan, H. Bagher-Ebadian, P. Khanduri, I. J. Chetty, and D. Zhu, "FocalUNETR: A focal transformer for boundary-aware segmentation of CT images," 2022, *arXiv:2210.03189*.

[20] S. Gong, Y. Zhong, W. Ma, J. Li, Z. Wang, J. Zhang, P.-A. Heng, and Q. Dou, "3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable tumor segmentation," 2023, *arXiv:2306.13465*.

[21] T. Shaharabany, A. Dahan, R. Giryes, and L. Wolf, "AutoSAM: Adapting SAM to medical images by overloading the prompt encoder," 2023, *arXiv:2306.06370*.

[22] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.

[23] K. Duarte, Y. Rawat, and M. Shah, "PLM: Partial label masking for imbalanced multi-label classification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 2733–2742.

[24] K. Yan, J. Cai, Y. Zheng, A. P. Harrison, D. Jin, Y. Tang, Y. Tang, L. Huang, J. Xiao, and L. Lu, "Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2759–2770, Oct. 2021.

[25] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany. Cham, Switzerland: Springer, Oct. 2015, pp. 234–241.

[26] S. Chen, K. Ma, and Y. Zheng, "Med3D: Transfer learning for 3D medical image analysis," 2019, *arXiv:1904.00625*.

[27] H. Wu, S. Pang, and A. Sowmya, "Tgnet: A task-guided network architecture for multi-organ and tumour segmentation from partially labelled datasets," in *Proc. IEEE 19th Int. Symp. Biomed. Imag. (ISBI)*, Mar. 2022, pp. 1–5.

[28] Y. Zhou, Z. Li, S. Bai, X. Chen, M. Han, C. Wang, E. Fishman, and A. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10671–10680.

[29] G. Shi, L. Xiao, Y. Chen, and S. K. Zhou, "Marginal loss and exclusion loss for partially supervised multi-organ segmentation," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101979.

[30] L. Fidon, M. Aertsen, D. Emam, N. Mufti, F. Guffens, T. Deprest, P. Demaerel, A. L. David, A. Melbourne, S. Ourselin, and J. Deprest, "Label-set loss functions for partial supervision: Application to fetal brain 3D MRI parcellation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, Strasbourg, France. Cham, Switzerland: Springer, Sep. 2021, pp. 647–657.

[31] G. Sun, M. Mendieta, J. Luo, S. Wu, and C. Chen, "FedPerfix: Towards partial model personalization of vision transformers in federated learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4965–4975.

[32] C.-M. Feng, Y. Yan, S. Wang, Y. Xu, L. Shao, and H. Fu, "Specificity-preserving federated learning for MR image reconstruction," *IEEE Trans. Med. Imag.*, vol. 42, no. 7, pp. 2010–2021, Aug. 2022.

[33] J. Liu, B. Li, and Z. Luo, "Magnetic type classification in sunspot group based on semi-supervised learning and knowledge distillation," in *Proc. 7th Int. Conf. Intell. Comput. Signal Process. (ICSP)*, Apr. 2022, pp. 1526–1529.

[34] C. Shen, P. Wang, D. Yang, D. Xu, M. Oda, P. T. Chen, K. L. Liu, W. C. Liao, C. S. Fuh, K. Mori, and W. Wang, "Joint multi organ and tumor segmentation from partial labels using federated learning," in *Proc. Int. Workshop Distrib., Collaborative, Federated Learn.* Cham, Switzerland: Springer, Sep. 2022, pp. 58–67.

[35] Q. Yu, D. Yang, H. Roth, Y. Bai, Y. Zhang, A. L. Yuille, and D. Xu, "C2FNAS: Coarse-to-fine neural architecture search for 3D medical image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4125–4134.

[36] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen, J. Yi, W. Zhao, X. Wang, Z. Liu, H.-T. Zheng, J. Chen, Y. Liu, J. Tang, J. Li, and M. Sun, "Parameter-efficient fine-tuning of large-scale pre-trained language models," *Nature Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023.

[37] J. Shen, W. Wang, C. Chen, J. Jiao, J. Liu, Y. Zhang, S. Song, and J. Li, "Med-tuning: A new parameter-efficient tuning framework for medical volumetric segmentation," 2023, *arXiv:2304.10880*.

[38] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "St-adapter: Parameter-efficient image-to-video transfer learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Dec. 2022, pp. 26462–26477.

[39] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surveys*, vol. 55, no. 9, pp. 1–35, Sep. 2023.

[40] M. Jia, L. Tang, B. C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2022, pp. 709–727.

[41] E. Ben Zaken, S. Ravfogel, and Y. Goldberg, "BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models," 2021, *arXiv:2106.10199*.

[42] D. Guo, A. M. Rush, and Y. Kim, "Parameter-efficient transfer learning with diff pruning," 2020, *arXiv:2012.07463*.

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[44] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," 2023, *arXiv:2304.12306*.

[45] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "NnU-net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Feb. 2021.

[46] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a Sequence-to-Sequence perspective with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

[47] M. Antonelli et al., "The medical segmentation decathlon," *Nature Commun.*, vol. 13, no. 1, p. 4128, Jul. 2022.

[48] N. Heller, N. Sathianathen, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, M. Oestreich, J. Dean, M. Tradewell, A. Shah, R. Tejpaul, Z. Edgerton, M. Peterson, S. Raza, S. Regmi, N. Papanikolopoulos, and C. Weight, "The KiTS19 challenge data: 300 kidney tumor cases with clinical context, CT semantic segmentations, and surgical outcomes," 2019, *arXiv:1904.00445*.

[49] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein, "MICCAI multi-atlas labeling beyond the cranial vaultworkshop and challenge," in *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault Workshop*, vol. 5, 2015, p. 12.

[50] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.

[51] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, Mar. 2020, pp. 429–450.

[52] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *Proc. Int. Conf. Mach. Learn.*, Nov. 2020, pp. 5132–5143.

[53] Y. Huang, Z. Chen, Z. Chen, D. Zhou, and E. Pan, "Lightweight defect detection network based on steel strip raw images," *Eng. Appl. Artif. Intell.*, vol. 145, Apr. 2025, Art. no. 110179.
[54] X. Yu, X. Liang, Z. Zhou, and B. Zhang, "Multi-task learning for hand heat trace time estimation and identity recognition," *Expert Syst. Appl.*, vol. 255, Dec. 2024, Art. no. 124551.
[55] X. Yu, X. Liang, Z. Zhou, B. Zhang, and H. Xue, "Deep soft threshold feature separation network for infrared handprint identity recognition and time estimation," *Infr. Phys. Technol.*, vol. 138, May 2024, Art. no. 105223.

TAREQ MAHMOD ALZUBI received the Ph.D. degree in computer graphics from the University of Santiago de Compostela, Spain. He is currently an Assistant Professor with Al-Balqa Applied University, Jordan. His research and professional endeavors encompass a range of interdisciplinary areas, including human–computer interaction, computer vision, machine learning, and artificial intelligence. His primary focus lies in the creation of advanced algorithms and systems for visual recognition, object detection, image segmentation, and understanding 3D scenes. He has significantly contributed to the development of real-time computer vision applications, especially in the domains of autonomous systems, surveillance, healthcare, and human–computer interaction. With a strong passion for both the theoretical and practical dimensions of computer vision, he has authored research papers in prestigious journals and conferences. His research interests include deep learning techniques for image and video analysis, facial recognition technologies, and the integration of multi-modal data.

HAMZA MUKHTAR received the M.S. degree in computer science from the University of Engineering and Technology Lahore (UET Lahore), Pakistan. He is currently a Deep Engineer with the Intelligent Criminology Laboratory, National Center of Artificial Intelligence, Al-Khawarizmi Institute of Computer Science, UET Lahore. His research interests include computer vision, natural language processing, machine learning, and deep learning.