

Mitigating Resolution-Drift in Federated Learning: Case of Keypoint Detection

Taeheon Lim, Joohyung Lee *Senior Member, IEEE*, Kyungjae Lee *Member, IEEE*, Jungchan Cho *Member, IEEE*

Abstract—The Federated Learning (FL) approach enables effective learning across distributed systems, while preserving user data privacy. To date, research has primarily focused on addressing statistical heterogeneity and communication efficiency, through which FL has achieved success in classification tasks. However, its application to non-classification tasks, such as human pose estimation, remains underexplored. This paper identifies and investigates a critical issue termed “resolution-drift,” where performance degrades significantly due to resolution variability across clients. Unlike class-level heterogeneity, resolution drift highlights the importance of resolution as another axis of not independent or identically distributed (non-IID) data. To address this issue, we present resolution-adaptive federated learning (RAF), a method that leverages heatmap-based knowledge distillation. Through multi-resolution knowledge distillation between higher-resolution outputs (teachers) and lower-resolution outputs (students), our approach enhances resolution robustness without overfitting. Extensive experiments and theoretical analysis demonstrate that RAF not only effectively mitigates resolution drift and achieves significant performance improvements, but also can be integrated seamlessly into existing FL frameworks. Furthermore, although this paper focuses on human pose estimation, our t-SNE analysis reveals distinct characteristics between classification and high-resolution representation tasks, supporting the generalizability of RAF to other tasks that rely on preserving spatial detail.

Index Terms—Federated learning, high-resolution regression, multi-resolution, and knowledge distillation.

I. INTRODUCTION

THE rapid increase in edge devices’ computational power has enabled local training on Internet of Things (IoT) devices using locally collected data [1], [2]. This paradigm shift facilitates machine learning without transmitting raw data to a central server, thereby overcoming the limitations of traditional centralized learning approaches, which often rely on constrained public datasets. Consequently, distributed machine learning [3]–[5] has become an essential technology

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) A grant funded by the Korean government (MSIT) (No. 2022-0-00871, Development of AI Autonomy and Knowledge Enhancement for AI Agent Collaboration) and a National Research Foundation of Korea (NRF) grant funded by The Korean Government (MSIT) (No. RS-2023-00211357).

Taeheon Lim is with the Department of Artificial Intelligence, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, 06974, Republic of Korea (e-mail: icecoffee2500@cau.ac.kr).

Joohyung Lee and Jungchan Cho are with the Department of Computing, Gachon University, Seongnam 13120, Rep. of Korea (e-mail: j17.lee@gachon.ac.kr; thinkai@gachon.ac.kr).

Kyungjae Lee is with the Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, Republic of Korea (e-mail: kyungjae_lee@korea.ac.kr).

(Corresponding authors: Kyungjae Lee and Jungchan Cho)

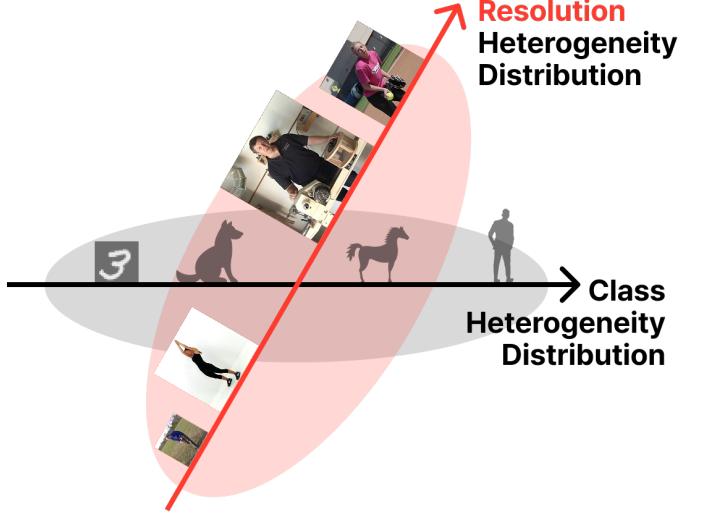


Fig. 1: Multiple axes defining statistical heterogeneity: the class axis represents the dataset’s class distribution, and the resolution axis represents the distribution based on image resolutions.

in edge computing and Internet of Things (IoT) environments. Federated Learning (FL) [6], [7] embodies this paradigm while preserving user data privacy. FedAvg [6] enables multiple clients’ models to undergo decentralized training while preserving their private data and aggregating local updates into a shared global model.

Federated learning often suffers substantial performance degradation when client data distributions are heterogeneous. FedProx [8] and SCAFFOLD [9] are representative methods for handling the statistical heterogeneity in FL. However, these advances address non-IID data primarily in classification settings. Real-world applications often require capabilities beyond classification, such as object detection, human pose estimation, and depth estimation, client datasets vary regarding class labels and image resolution (see Figure 1). Therefore, a pressing need arises to explore the applicability of FL to these non-classification tasks, an area that remains significantly under-investigated. Specifically, high-resolution regression problems such as landmark or keypoint detection (e.g., Human Pose Estimation [10]–[12]) require fundamentally different network architectures, often adopting encoder-decoder structures. The decoder must recover the encoded features back to the input resolution, requiring the preservation of spatial information throughout the network. By contrast,

classification models use an encoder-only architecture and discard most spatial details because the input image’s spatial information need not be reflected in a single-class label.

In such high-resolution regression scenarios, all the participating clients are unlikely to possess data with the same resolution. For example, the server aggregates the locally trained model weights derived from both low- and high-resolution images. Multi-resolution aggregation is a fundamentally different issue to that which has been extensively studied in existing FL research on statistical heterogeneity [8], [9], where heterogeneity is often implicitly confined to class distribution skew. A critical challenge for high-resolution regression in such settings is the feature representation mismatch caused by information loss or distortion owing to resolution differences, rendering these tasks highly sensitive to variations in data resolution. As discussed in Section III, high-resolution regression tasks demand much deeper spatial features than classification. Therefore, heterogeneity along axes other than the class can cause severe performance drift. Previous studies have noted multiple axes of non-IIDness [13]; however, resolution heterogeneity remains under-investigated. This new distribution shift axis requires dedicated methods for mitigating its effects. Although there are other forms of distribution heterogeneity, this paper focuses specifically on resolution heterogeneity as a critical factor in high-resolution regression tasks.

To our knowledge, this multi-resolution issue has not been explicitly addressed in FL literature. We introduce the term “**resolution-drift**” to describe this performance degradation phenomenon. Resolution drift occurs when clients train on data with different resolutions, causing the global model to overfit to certain resolutions and lose its ability to generalize across others. Existing FL methods designed to address data heterogeneity [8], [9] fail to address the core issue, since they focus solely on the global aggregation step and overlook local resolution-induced overfitting. For FL to be widely adopted in practical settings, its effectiveness must extend beyond class-level predictions to supporting pixel- or coordinate-level tasks, which are central to high-resolution regression problems such as analyzing CCTV footage collected from cameras with varying resolutions. Therefore, new FL methodologies must be developed, which can maintain a stable performance across diverse resolution inputs. Addressing this issue requires novel approaches beyond the conventional FL techniques.

This paper proposes **Resolution Adaptive Federated Learning (RAF)** to mitigate the resolution-drift problem for non-classification tasks. Our system architecture assumes that clients possess data with different resolutions (Figure 4). To counteract resolution drift, RAF employs a heatmap-based Knowledge Distillation (KD) strategy, where a KD loss function minimizes the distance between outputs generated from higher-resolution inputs (teacher) and those from lower-resolution inputs (student). By serving as soft targets, the teacher’s output acts as a regularizer, preventing overfitting to any single resolution, thereby, enhancing robustness across multiple resolutions. This is consistent with the findings in the literature [14]–[16]. However, designing a regularizer based on knowledge distillation across multiple resolutions using a transformer backbone introduces additional complications.

The self-attention mechanism in transformers is inherently permutation-invariant; therefore, positional embeddings are required to inject spatial order information. Vision Transformer (ViT), our selected backbone relies on Absolute Positional Embedding (APE), whose shape is fixed by the input resolution and cannot adapt during training. Consequently, training ViT on multi-resolution inputs is complicated. To address this limitation, we drew inspiration from recent work, ResFormer [17], which replaces APE with convolution-based positional embeddings. In ResFormer, the convolutional kernels are learned, allowing positional embedding to adapt dynamically to different input resolutions and inject a smooth spatial context. Since ResFormer demonstrated this convolution-based positional embedding approach only in the context of classification tasks, its effectiveness for high-resolution regression remains unclear. Our experiments confirm that these embeddings are also effective for non-classification tasks, enabling robust feature extraction across varying resolutions.

Main Contributions

Our main contributions are summarized as follows:

- We identify the “resolution-drift” phenomenon in multi-resolution federated learning and formally define its impact on non-classification tasks such as high-resolution regression.
- We extend convolution-based positional embeddings to high-resolution regression, showing that they enable Vision Transformer backbones to train effectively on multi-resolution inputs.
- We present RAF, a novel framework that integrates multi-resolution knowledge distillation as a resolution-regularizer within standard FL, and provide a theoretical analysis of its convergence.
- Through extensive experiments, we demonstrate that RAF mitigates resolution-drift and allows single-resolution clients to benefit from a globally trained model that remains robust across all resolutions.
- RAF is modular and orthogonal to existing FL aggregation schemes (e.g., FedProx), facilitating easy integration into a wide range of FL pipelines.

II. RELATED WORKS

A. Federated Learning (FL)

Federated Learning is a distributed-learning paradigm that enables training across multiple clients without transferring their data to a central server, thereby preserving data privacy [6]. FL has become an essential technology in edge computing and IoT environments because it facilitates collaborative learning while addressing privacy concerns. Existing FL research has primarily focused on solving the challenges related to statistical heterogeneity and communication efficiency. FedProx [8] has extended FedAvg by adding a proximal term to each client’s objective, thereby enabling more stable convergence under non-IID data distributions in classification settings. SCAFFOLD [9] further enhances this by introducing control variates to correct client drift during local updates, significantly reducing communication rounds, and improving

the accuracy in distributed classification tasks. However, these studies [8], [9] mainly concentrated on class-level heterogeneity and did not address resolution-level heterogeneity. Unlike classification tasks, which predict class-level labels, high-resolution regression tasks require pixel-level label predictions, making them inherently more challenging.

In real-world applications, FL systems encounter devices with significantly different sensing capabilities. For example, in autonomous driving fleets, some vehicles are equipped with high-definition cameras, whereas others have low-resolution dashcams or infrared sensors. In precision agriculture, data may be collected using both high- and low-altitude drones, all of which produce images with different resolutions. Existing FL approaches [8], [9], [18]–[20] primarily address class-level heterogeneity but do not consider resolution-level heterogeneity. This limitation underscores the need for FL methodologies that can achieve robust learning in multi-resolution environments.

B. Vision Transformer in Federating Learning

Transformer architectures, originally introduced by Vaswani et al. [21] and now ubiquitous across NLP [21]–[25], vision [26]–[29], audio [30]–[32], and multimodal [33]–[35] domains, offer key advantages for FL: their self-attention mechanism captures long-range dependencies, their modular design adapts easily to diverse tasks, and they scale gracefully with data size. Recent FL research has begun to leverage these strengths, achieving improved robustness and convergence with non-IID data [36], [37].

We build on this trend by adopting a Vision Transformer (ViT) [26] as the backbone of the proposed method. ViT [26] processes images by splitting them into flattened patches. However, ViT relies on APE with its size fixed to a single training resolution, which hampers generalization to other resolutions and often requires interpolation at the inference stage [17]. To overcome this limitation, we replace APE with convolution-based positional embeddings from ResFormer [17], which dynamically injects the spatial context across varying input sizes without explicit resizing. Although prior work has demonstrated this strategy primarily for classification, we extend it to a high-resolution regression task within a federated learning setting, demonstrating that it significantly improves human pose estimation (HPE) performance and resolution robustness.

III. IN-DEPTH ANALYSIS OF RESOLUTION EFFECTS IN FEDERATED LEARNING

Although extensive FL research has addressed the challenge of statistical heterogeneity in non-IID settings, most efforts remain limited to two key aspects.

- Previous works have predominantly addressed classification tasks.
- Statistical heterogeneity is typically considered solely in terms of class distribution skew.

In this section, we discuss these limitations and underscore the need to address resolution heterogeneity in real-world FL scenarios.

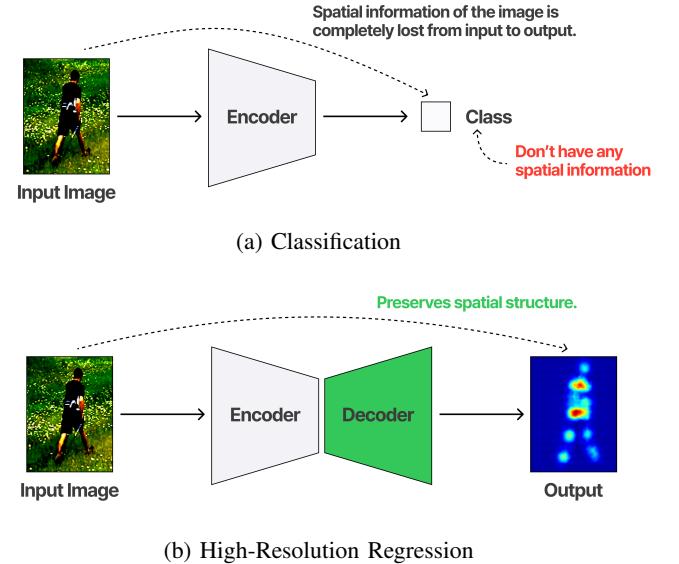


Fig. 2: The architectural difference between classification and high-resolution regression

A. Fundamental Differences Between Classification and High-Resolution Regression

In real-world FL scenarios, classification tasks represent only a small fraction of use cases. The most fundamental vision problems involve high-resolution regression, including keypoint detection, depth estimation, and super-resolution. These tasks underpin many more real-world applications; however, they pose far greater challenges than classification and have received relatively little attention from FL researchers. This gap motivated us to focus on high-resolution regression in federated settings.

Figure 2 illustrates the architectural differences between the classification and high-resolution regression models. As shown in Figure 2a, classification corresponds to image-level prediction in which a single class label is predicted for each input image. In this setting, the predicted output does not contain any spatial information even though the input is a rich spatial map. Consequently, the network progressively shrinks the spatial dimensions of its feature maps and distills all relevant information into a single vector. By contrast, high-resolution regression requires predictions that align with spatial locations in the original image, as illustrated in Figure 2b. These high-resolution regression problems must preserve spatial detail throughout the network and produce output feature maps at or near the input resolution. Consequently, architectures for these tasks often adopt an encoder-decoder design, extract deep features, and then reconstruct or upsample them to full image size.

Because the output itself must retain a spatial structure, models for these tasks are inherently more sensitive to changes in the input resolution and demand richer feature representations than models trained solely for classification. In keypoint detection models, for instance, encoder-decoder architectures are designed with skip connections specifically to preserve high-resolution information; such spatial fidelity is critical for accurate boundary localization [38].

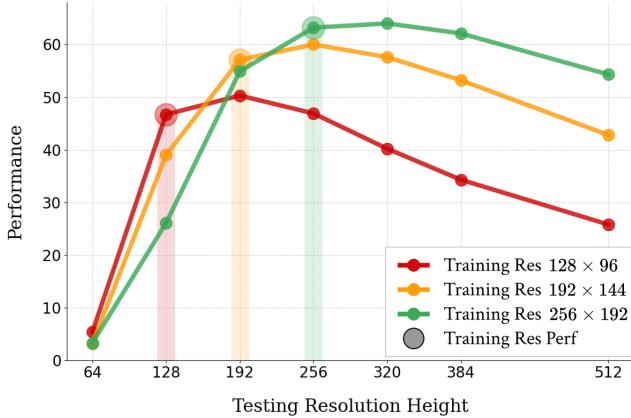


Fig. 3: For the Human Pose Estimation task in the centralized learning scenario, each model was identically trained on a single resolution across three clients and tested on various resolutions. “Training Res Perf” denotes the resolution used during model training. The x-axis shows only the height of the training resolution, and all displayed resolutions maintain a 4:3 height-to-width aspect ratio.

B. Experimental Analysis of Resolution Mismatch Effects

In realistic FL scenarios for high-resolution regression, clients inevitably possess images at different resolutions rather than at a single fixed size. As discussed in Section I, this variation introduces resolution heterogeneity, which, after model aggregation, can hinder FL by significantly degrading performance.

Figure 3 shows that the ViT-based models suffer from critical limitations: We examined the performance variation when a ViT model, trained in a centralized manner at a single resolution, was tested at different resolutions. The task was keypoint detection, a representative high-resolution regression task. This involves estimating human keypoint locations from a single image and is commonly used for detailed human analysis. As shown in Figure 3, at its trained resolution, each model achieved peak performance, but its accuracy dropped sharply when tested on unseen resolutions. This indicates that the characteristics of the learned weights vary significantly depending on resolution.

C. Resolution Drift: Performance Degradation in Federated Learning

As discussed in Section III-B, resolution heterogeneity can introduce inconsistencies in the learned representations across clients who may possess data with different resolutions. These inconsistencies, when aggregated on the server, may significantly decrease the model’s performance. To investigate this issue further, we conducted a controlled experiment to demonstrate explicitly how resolution heterogeneity can impair the effectiveness of federated learning.

In particular, we simulated an FL setting in which each client was assigned a dataset with a distinct input resolution. Despite using the same model architecture and training procedures across all clients, the heterogeneity in resolution leads to diverging updates, rendering aggregation on the server more

TABLE I
Low-resolution (128×96) test accuracy on the human pose estimation task for models trained on various resolution triplets in a federated learning scenario, illustrating the **resolution-drift** phenomenon. “Res.” denotes resolution.

Trained Resolution			Test Res.
Res. 1	Res. 2	Res. 3	128 × 96
128 × 96	128 × 96	128 × 96	52.8
128 × 96	128 × 96	192 × 144	52.9
128 × 96	192 × 144	192 × 144	52.4
128 × 96	128 × 96	256 × 192	52.3
128 × 96	192 × 144	256 × 192	51.6
128 × 96	256 × 192	256 × 192	51.3

challenging. Table I presents the performance of the global models trained under resolution-diverse scenarios.

In Table I, the top row corresponds to all the clients using low-resolution data. Moving downward, the number of clients using high-resolution data increases. Although higher-resolution data generally provide richer visual information and improve model learning, the inference performance on low-resolution inputs decreases as the training data diverge more from that resolution. Notably, only the configurations 128×96 , 128×96 , 192×144 yielded a slight improvement, possibly because of the inclusion of more informative training samples. The results in Table I reveal that federated models trained with heterogeneous resolutions consistently underperform compared to those trained in resolution-homogeneous settings. This performance drop, which we refer to as **resolution drift**, highlights the tangible risk posed by resolution heterogeneity in FL environments. Such drift arises because the model is uncertain regarding to which resolution it should adapt, with degraded performance across all resolutions. This emphasizes that federated learning frameworks must explicitly account for resolution-related variability, and ensure robustness and generalization.

IV. PROPOSED METHOD: RESOLUTION ADAPTIVE FEDERATED LEARNING (RAF)

Section III explains why federated learning in real-world settings must address high-resolution regression tasks and the resulting resolution-drift phenomenon is experimentally and visually demonstrated. When clients are trained at different image resolutions, their local updates overfit to those specific resolutions. Averaging these resolution-specific weights through FedAvg confuses the global model, causing its performance to decrease even though the input images contain rich spatial information. Consequently, clients have little incentive to participate. Why would they incur communication overhead if the aggregated model performs worse than the local training? To overcome this barrier, we propose a multi-resolution knowledge distillation framework that allows each client to leverage spatial features from its unique-resolution data and thereby effectively mitigate resolution-drift.

A. Overview of RAF

Figure 4 shows a federated learning scenario in which multiple clients participate using data captured at different

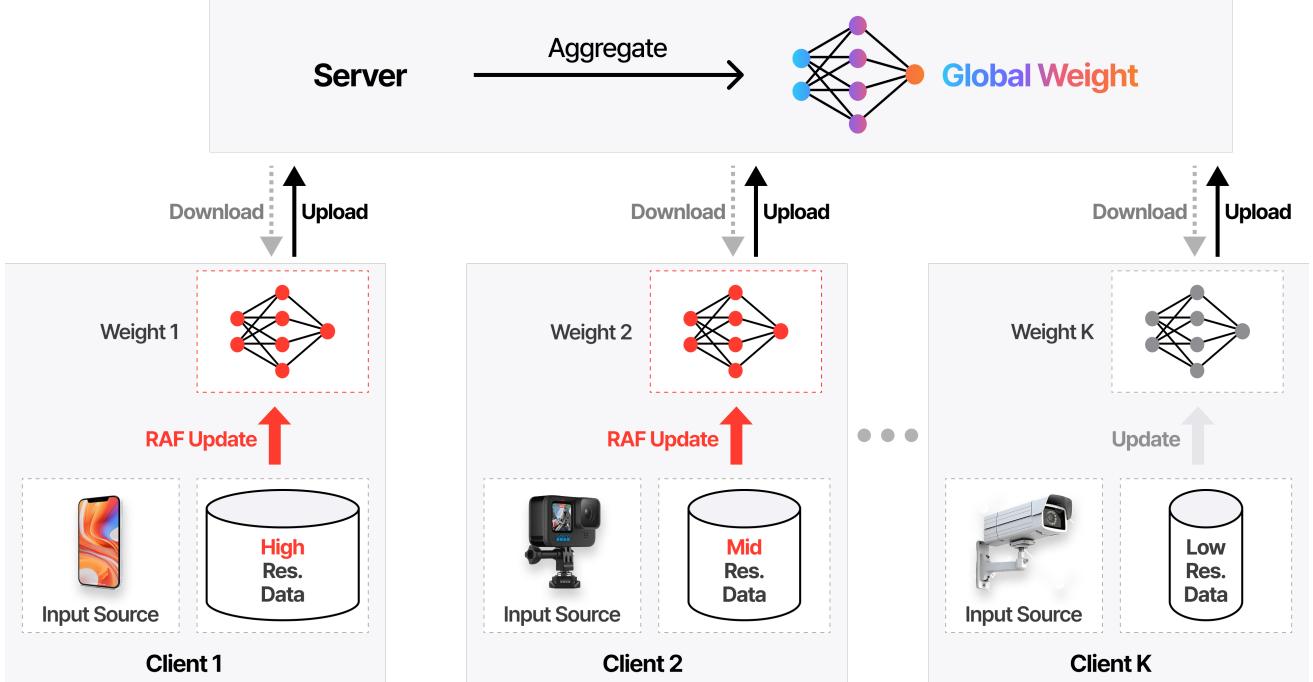


Fig. 4: Overview of RAF when applied to clients with varying-resolution data. RAF fully exploits spatial information in multi-resolution images. “Res.” denotes resolution. RAF does not update at the lowest resolution, since further downsampling is impossible.

resolutions. As noted in Section I, it is highly unlikely that all clients in a real-world FL deployment will possess datasets with identical resolutions. For instance, images captured by a smartphone are often of high resolution, those captured by a small action camera are lower, and CCTV footage may be even lower.

To model this heterogeneity, we assume that each client k holds a private dataset $D_k = \{(x_{k,j}, T_{k,j})\}_{j=1}^{n_k}$ where $x_{k,j} \in \mathbb{R}^{H_k \times W_k \times C}$ represents j -th image recorded at the client’s native resolution (H_k, W_k) , and $T_{k,j}$ is a ground-truth heat-map defined on the same pixel grid. The overall training flow followed the standard FedAvg federated learning procedure. Formally, let N be the number of clients and suppose that the k -th client holds n_k training samples $x_{k,j}$ (where $j = 1, \dots, n_k$). We seek a global objective model parameterized by w , which minimizes the aggregate objective.

$$\min_w \mathcal{L}(w) \triangleq \frac{1}{N} \sum_{k=1}^N \mathcal{L}_k(w). \quad (1)$$

In each communication round, the server broadcasts the current global weights w to all clients. Each client subsequently performs local updates on its resolution-specific dataset $\{D_k, r_k\}$, where r_k indicates the client’s input resolution. Crucially, to prevent each client’s model from overfitting to its own resolution, we augment the local training objective using a multi-resolution knowledge distillation term. This additional term encourages each client to incorporate spatial information from other resolutions, thereby counteracting the resolution drift.

After local training, each client uploads its updated weights to the server, which aggregates them using weighted averaging

(FedAvg). The updated global model is then redistributed to all the clients, and the process is repeated until convergence. Detailed derivations of the local objective explicitly incorporating multi-resolution distillation are provided in Section IV-B.

B. Maximally Utilizing Spatial Information via Multi-Resolution Knowledge Distillation

In this subsection, we describe how our proposed model and training scheme effectively mitigate resolution drift. First, we outline the backbone architecture and its modifications to support multi-resolution inputs. We then introduce our Multi-Resolution Knowledge Distillation (MRKD) method and explain how it serves as a resolution-aware regularizer.

1) Model Architecture: To leverage the strengths of transformer-based networks (as discussed in Section II-B) and ensure compatibility with recent vision-transformer models, we adopt ViTPose [11], whose encoder follows the standard ViT architecture. Specifically, it consists of a patch-embedding layer (including positional embeddings), followed by several transformer blocks, each containing a multi-head self-attention layer and a feed-forward network layer. Therefore, ViTPose relies on APE tied to a single resolution; it cannot accommodate multi-resolution training.

The ViT-based encoder requires a one-dimensional sequence of token embeddings rather than a two-dimensional image. The input 2D image $x \in \mathbb{R}^{H \times W \times C}$ is reshaped into a sequence of flattened 2D patches $x_p \in \mathbb{R}^{N_p \times (P^2 C)}$, where (H, W) is the resolution of the image, (P, P) is the resolution of each image patch, C is the number of channels, and $N_p = \frac{HW}{P^2}$ is the number of patches. Next, a learnable linear projection $E \in \mathbb{R}^{(P^2 C) \times D}$ maps every patch to a D -dimensional latent

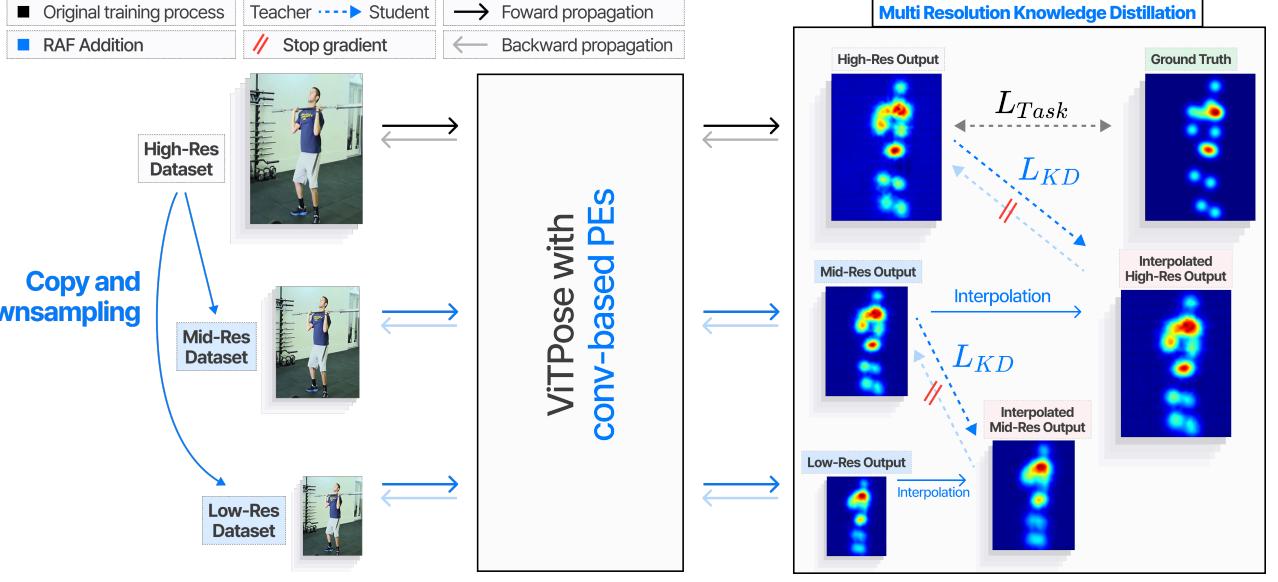


Fig. 5: Detailed illustration of RAF local training on a high-resolution client. The client’s original dataset is downsampled to create mid- and low-resolution inputs. The black arrows and boxes depict the standard ViTPose training flow, while the blue arrows and highlights indicate the additional RAF components. “Stop gradient” denotes that gradients are detached on the teacher branch during backpropagation for knowledge distillation.

vector to produce embedded patches $\mathbf{z}_p \in \mathbb{R}^{N_p \times D}$. Then, a learnable positional embedding $\mathbf{E}_{\text{pos}} \in \mathbb{R}^{N_p \times D}$ is added to each token.

Note that the shape of \mathbf{E}_{pos} depends directly on N_p , which, in turn, is determined by the input image resolution (H, W) . If the resolution changes, N_p also changes, and the learned positional embedding \mathbf{E}_{pos} must be reinitialized or resized. Consequently, a model trained with APE at one resolution cannot be directly generalized to another resolution, causing the accuracy to drop sharply on the unseen resolutions. To prevent this and enable dynamic multi-resolution learning, we replace ViTPose’s fixed APE with the convolution-based positional embeddings proposed in ResFormer [17]. First, we adopt a Global Positional Embedding (GPE) module immediately after the patch embedding layer. GPE employs a 3×3 depth-wise convolution to inject a smooth, global spatial context across the entire feature map. Next, within each multi-head self-attention block, we integrate a Local Positional Embedding (LPE) module, which also uses a 3×3 depthwise convolution, but focuses on capturing fine-grained local relationships. By substituting ViTPose’s APE with GPE and LPE, our model can process inputs at arbitrary resolutions without modifying the network architecture.

2) *Multi-Resolution Knowledge Distillation*: Figure 5 illustrates the local training procedure for clients by using high-resolution data. Unlike standard ViTPose, which operates at a single fixed resolution, RAF requires multi-resolution inputs during training to achieve strong generalization across resolutions.

Let $N_{k,\text{res}}$ be the total number of resolution levels employed when the client k undergoes local training. For every original image $x_{k,j} \in \mathbb{R}^{H_k \times W_k \times C}$ ($j = 1, \dots, n_k$), we create $N_{k,\text{res}} -$

1 additional downsampled copies, denoted

$$x_{k,j}^{(i)} \in \mathbb{R}^{H_k^{(i)} \times W_k^{(i)} \times C} \quad (i = 0, \dots, N_{k,\text{res}} - 1), \quad (2)$$

where $x_{k,j}^{(0)} = x_{k,j}$ and, for $i > 0$, $H_k^{(i)} < H_k^{(i-1)}$ and $W_k^{(i)} < W_k^{(i-1)}$ hold. Here, i is the resolution index (the larger the index, the lower the spatial resolution), and the subscripts k and j identify the client and the sample, respectively. Each image $x_{k,j}^{(i)}$ is obtained by interpolation of the native image $x_{k,j}$ to the target size $H_k^{(i)} \times W_k^{(i)}$.

After generating these $N_{k,\text{res}}$ multi-resolution inputs, we pass each $x_{k,j}^{(i)}$ through model \mathcal{M} parameterized by the local weight w to obtain the corresponding heatmap output

$$y_{k,j}^{(i)} = \mathcal{M}(x_{k,j}^{(i)}; w). \quad (3)$$

First, we define a task loss $\mathcal{L}_{\text{task}}$ as the highest resolution output $y_{k,j}^{(0)}$. Following ViTPose [11], we use the mean squared error (MSE) against the ground-truth heatmap $T_{k,j}$:

$$\mathcal{L}_{k,\text{task}}(w) = \frac{1}{n_k} \sum_{j=1}^{n_k} \|y_{k,j}^{(0)} - T_{k,j}\|_2^2. \quad (4)$$

To enforce *scale-consistency* and encourage the model to extract rich spatial features from low-resolution inputs, we introduce a *multi-resolution knowledge distillation* loss \mathcal{L}_{kd} . Specifically, we treat the heatmap output at resolution $i-1$ as the “teacher” and the heatmap at resolution i as the “student.” By minimizing the MSE between these output pairs, we push the model to produce low-resolution heatmaps that resemble the higher-resolution ones. Formally,

$$\mathcal{L}_{k,\text{kd}}(w) = \sum_{i=1}^{N_{k,\text{res}}-1} \frac{1}{n_k} \sum_{j=1}^{n_k} \left\| \text{sg} \left(y_{k,j}^{(i-1)} \right) - U_i^{i-1} y_{k,j}^{(i)} \right\|_2^2. \quad (5)$$

Algorithm 1 Resolution Adaptive Federated Learning (RAF)

Require: Number of rounds T ; number of local epochs E ; number of clients N ; learning rate η ; distillation weight α ; initial global model w_0 ; each client k holds private dataset D_k ; number of local data samples available on that client n_k ; total number of resolutions used in local training for k -th client $N_{k,\text{res}}$

```

1: for round  $t = 0$  to  $T - 1$  do
2:   Server broadcasts  $w_t$  to all clients
3:   for each client  $k$  in parallel do
4:     update local model parameter  $w_k$  with  $w_t$ 
5:     Let  $D_k^{(i)}$  denote the dataset downsampled to the  $i$ -th resolution level, where larger  $i$  corresponds to progressively lower resolutions.
6:     for  $i = 1$  to  $N_{k,\text{res}} - 1$  do
7:       Downsample entire  $D_k^{(0)}$  to obtain  $D_k^{(i)}$ 
8:     end for
9:     for local epoch  $e = 1$  to  $E$  do
10:    for each sample  $(x_{k,j}^{(0)}, T_{k,j}) \in D_k^{(0)}$  do
11:      Select corresponding sample  $x_{k,j}^{(i)}$  for  $i = 1, \dots, N_{k,\text{res}} - 1$ 
12:       $\triangleright$  Compute task loss  $\mathcal{L}_{k,\text{task}}(w)$  as defined in Equation (4)
13:       $\triangleright$  Compute distillation loss  $\mathcal{L}_{k,\text{kd}}(w)$  as defined in Equation (5)
14:       $\triangleright$  Compute total loss  $\mathcal{L}_k(w)$  as defined in Equation (6)
15:       $\triangleright$  Backward and update
16:       $w_k \leftarrow w - \eta \nabla_{w_k} \mathcal{L}_k(w)$ 
17:    end for
18:  Client  $k$  sends  $w_k$  back to the server
19: end for
20: Server aggregates

$$w_{t+1} = \frac{1}{N} \sum_{k=1}^N w_k$$

21: end for
22: Output: Final global model  $w_T$ 

```

Note that during backpropagation, we detach the teacher output $(y_{k,j}^{(i-1)})$ from the computational graph with sg operator (as in BYOL [39]), which denotes the stop gradient, so that the gradients flow only through the student branch $y_{k,j}^{(i)}$. The matrix $U_i^{i-1} \in \mathbb{R}^{H_k^{(i-1)}W_k^{(i-1)} \times H_k^{(i)}W_k^{(i)}}$ is a fixed linear upsampling operator that maps low-resolution predictions at level i to higher resolution $i - 1$. Finally, the combined local objective for client k is:

$$\mathcal{L}_k(w) = \mathcal{L}_{k,\text{task}}(w) + \alpha \mathcal{L}_{k,\text{kd}}(w) + \gamma \mathcal{L}_{k,\text{reg}}(w), \quad (6)$$

where $\alpha > 0$ balances task accuracy and scale-consistency regularization, $\mathcal{L}_{k,\text{reg}}$ indicates an ℓ_2 regularization, and γ is its coefficient. Algorithm 1 summarizes the RAF procedure. By replicating each original image $N_{k,\text{res}} - 1$ times at progres-

sively lower resolutions and applying knowledge distillation in a teacher-student hierarchy, RAF ensures that the model learns to generalize across all $N_{k,\text{res}}$ resolutions, thus mitigating the resolution drift discussed earlier.

3) *MRKD as a Resolution Regularizer*: Our simple, yet effective MRKD approach acts as a resolution-aware regularizer. By aligning the teacher and student outputs across different resolutions, MRKD prevents the model from overfitting to any single resolution. This regularization effect improves generalization across all scales, thereby boosting the accuracy and robustness for both low and high unseen resolutions. In Section V-E, we show that MRKD-trained models can outperform their baselines even when evaluated on interpolated inputs at resolutions higher than those accessible during inference. This indicates that MRKD helps the network extract and utilize spatial information more effectively.

C. Convergence Analysis

This section establishes that the proposed RAF algorithm enjoys the same asymptotic convergence rate as standard FEDAVG. Our analysis concentrates on the *late* phase of training, where a deep network is empirically observed, and theoretically justified [40]–[42], to behave like its linearisation. Specifically, once the shared feature extractor has *converged* (i.e., its weights evolve only marginally), the optimisation signal is absorbed almost exclusively by the last affine layer. We therefore regard the backbone up to the penultimate layer as a fixed feature map and study the remaining optimisation as that of a *linear model with frozen features*, a viewpoint often called the *lazy-training* or *post-neural-collapse* regime.

Under the feature-converged assumption, we linearize the original network objective with respect to the final-layer weights; the resulting expected local loss for client k at round t is given by

$$\mathcal{L}_k(w) = \frac{1}{n_k} \sum_{j=1}^{n_k} \|(\psi_{k,j}^{(0)})^\top w - T_{k,j}\|_2^2 \quad (7)$$

$$+ \alpha \sum_{i=1}^{N_{k,\text{res}}-1} \frac{1}{n_k} \sum_{j=1}^{n_k} \|(\psi_{k,j}^{(i-1)})^\top w_t - U_i^{i-1} (\psi_{k,j}^{(i)})^\top w\|_2^2 \quad (8)$$

$$+ \frac{\gamma}{2} \|w\|_2^2 \quad (9)$$

Throughout this section we adopt the following concise notation. For client k and sample j the expression $(\psi_{k,j}^{(i)})^\top w$ ($i = 0, \dots, N_{k,\text{res}} - 1$) denotes the heat-map obtained by projecting the feature vector $\psi_{k,j}^{(i)}$ with the weight of the last layer, w . In late rounds, the backbone is assumed *feature-converged*, so only the last-layer weight is updated.

This linearized objective includes both a supervised loss term for the highest-resolution predictions and a multi-resolution knowledge distillation loss that encourages consistency between consecutive resolutions through the fixed upsampling operators U_i^{i-1} . We now introduce an alternative

formulation of the loss:

$$\bar{\mathcal{L}}_k(w) = \frac{1}{n_k} \sum_{j=1}^{n_k} \|(\psi_{k,j}^{(0)})^\top w - T_{k,j}\|_2^2 \quad (10)$$

$$+ \alpha \sum_{i=1}^{N_{k,\text{res}-1}} \frac{1}{n_k} \sum_{j=1}^{n_k} w^\top (\psi_{k,j}^{(i)})^\top (U_i^{i-1})^\top \times \left[(\psi_{k,j}^{(i-1)})^\top - U_i^{i-1} (\psi_{k,j}^{(i)})^\top \right] w \quad (11)$$

$$+ \frac{\gamma}{2} \|w\|_2^2. \quad (12)$$

Although $\mathcal{L}_k(w)$ and $\bar{\mathcal{L}}_k(w)$ differ globally, they are locally equivalent at the current iterate $w = w_t$. Specifically,

$$\mathcal{L}_k(w_t) = \bar{\mathcal{L}}_k(w_t), \quad (13)$$

$$\nabla_w \mathcal{L}_k(w_t) = \nabla_w \bar{\mathcal{L}}_k(w_t). \quad (14)$$

Therefore, gradient-based optimization of $\mathcal{L}_k(w)$ in any neighbourhood of the current iterate w_t yields exactly the same update direction as would be obtained from $\bar{\mathcal{L}}_k(w)$. Both objectives thus attain the same critical point, and their local trajectories are indistinguishable. Moreover, because the alternative loss $\bar{\mathcal{L}}_k(w)$ is a quadratic form whose Hessian is augmented by the positive-definite matrix $\frac{\gamma}{2} I_d$ with $\gamma > 0$, it is strictly convex; consequently, it possesses a unique and non-trivial minimiser w^* rather than a degenerate solution. Hence, the two formulations induce identical convergence behaviour, and $\bar{\mathcal{L}}_k(w)$ can legitimately be used as a surrogate objective in theoretical analysis.

To analyze the convergence property, we first introduce the following assumptions.

Assumption 1 (Boundedness). *There exist positive constants M_ϕ , M_U , and M_T such that, for any client k , for all resolution level i and every sample j pairs,*

$$\|\psi_{k,j}^{(i)}\|_2 \leq M_\phi, \quad \|U_i^{i-1}\|_2 \leq M_U, \quad \|T_{k,j}\|_2 \leq M_T. \quad (15)$$

Assumption 2 (Unbiased Stochastic Gradients). *For any client k and parameter vector w with $\|w\|_2 \leq R$, let ξ denote a random index drawn uniformly from the local sample set. The sample gradient $\nabla \mathcal{L}_k(w; \xi)$ satisfies $\mathbb{E}_\xi[\nabla \mathcal{L}_k(w; \xi)] = \nabla \mathcal{L}_k(w)$.*

Note that Assumption 2 is generally used in convergence analysis of federated learning [43]. Under these conditions we shall prove three auxiliary propositions that bound (i) the gradient's Lipschitz constant, (ii) the objective's strong-convexity modulus, and (iii) the variance of stochastic gradients. Taken together, these propositions deliver the smoothness, strong-convexity, and bounded-variance conditions required for the FedAvg convergence result of [43].

Proposition 1 (Smoothness of $\bar{\mathcal{L}}_k$). *Under Assumption 1, the gradient of $\bar{\mathcal{L}}_k(w)$ with respect to w is L -Lipschitz with $L = 2(1 + \alpha(r-1)(M_U+1)^2)M_\phi^2 + \gamma$.*

Proof. We compute the gradient of each term in $\bar{\mathcal{L}}_k(w)$. Let us denote the gradient of the first term as

$$\nabla_w \mathcal{L}_{k,\text{task}}(w) = \frac{2}{n_k} \sum_{j=1}^{n_k} \psi_{k,j}^{(0)} ((\psi_{k,j}^{(0)})^\top w - T_{k,j}). \quad (16)$$

This is a linear function of w , hence, its Lipschitz constant is computed as

$$L_{k,\text{task}} = \frac{2}{n_k} \sum_{j=1}^{n_k} \|\psi_{k,j}^{(0)}\|_2^2 \leq 2M_\phi^2. \quad (17)$$

Next, let us find the Lipschitz constant of the gradient of knowledge distillation term. Set $A_{ij} = U_i^{i-1} (\psi_{k,j}^{(i)})^\top$ and $B_{ij} = (\psi_{k,j}^{(i-1)})^\top$. Then, the gradient of the knowledge distillation term is written as

$$\nabla_w \mathcal{L}_{k,\text{kd}}(w) = \sum_{i=1}^{r-1} \frac{2}{n_k} \sum_{j=1}^{n_k} A_{ij}^\top (A_{ij} - B_{ij}) w. \quad (18)$$

The matrix norms satisfy $\|A_{ij}\|_2 \leq M_U M_\phi$ and $\|B_{ij}\|_2 \leq M_\phi$, hence $\|A_{ij} - B_{ij}\|_2 \leq M_\phi(M_U + 1)$. Consequently,

$$\|\nabla_w \mathcal{L}_{k,\text{kd}}(w) - \nabla_w \mathcal{L}_{k,\text{kd}}(v)\|_2 \quad (19)$$

$$\leq \frac{2}{n_k} \sum_{i=1}^{r-1} \sum_{j=1}^{n_k} \|A_{ij}^\top\|_2 \|A_{ij} - B_{ij}\|_2 \|w - v\|_2 \\ \leq 2(r-1) M_U M_\phi (M_U + 1) \|w - v\|_2. \quad (20)$$

The bound $M_U(M_U + 1) \leq (M_U + 1)^2$ gives

$$L_{k,\text{kd}} \leq 2(r-1)(M_U + 1)^2 M_\phi^2. \quad (21)$$

Finally, for the regularization term, one has

$$\nabla_w \mathcal{L}_{k,\text{reg}}(w) = \gamma w, \quad L_{k,\text{reg}} = \gamma. \quad (22)$$

Combining all Lipschitz constants, the global Lipschitz constant of $\bar{\mathcal{L}}_k$ is obtained as

$$L_k = L_{k,\text{task}} + \alpha L_{k,\text{kd}} + L_{k,\text{reg}} \\ \leq 2M_\phi^2 + \alpha[2(r-1)(M_U + 1)^2 M_\phi^2] + \gamma \\ = 2(1 + \alpha(r-1)(M_U + 1)^2) M_\phi^2 + \gamma. \quad (23)$$

This completes the proof of L -smoothness. \square

Proposition 2 (Strong convexity of $\bar{\mathcal{L}}_k$). *Under Assumption 1, $\bar{\mathcal{L}}_k$ is γ -strongly convex with respect to w .*

Proof. From the regularization term in $\bar{\mathcal{L}}_k$, $\nabla^2 \bar{\mathcal{L}}_k(W) \succeq \gamma I$ holds, and the function is γ -strongly convex. \square

Proposition 3 (Bounded Gradient Norm and Variance). *Let Assumption 1 hold and assume the current iterate satisfies $\|w_t\|_2 \leq R$. Define the constant $C = \mathcal{O}(M_\phi M_T + (rM_\phi^2 M_U^2 + \gamma)R)$. Then,*

$$\|\nabla \bar{\mathcal{L}}_k(w_t)\|_2 \leq C, \quad (24)$$

$$\mathbb{E}_\xi [\|\nabla \bar{\mathcal{L}}_k(w_t; \xi) - \nabla \bar{\mathcal{L}}_k(w_t)\|_2] \leq C. \quad (25)$$

Proof. We analyze the gradient of the local loss $\bar{\mathcal{L}}_k(w)$, which consists of three terms: the task loss, the distillation loss, and the regularization term. Each will be bounded separately.

First, the gradient of the task loss is given by

$$\nabla_w \mathcal{L}_{k,\text{task}}(w_t) = \frac{2}{n_k} \sum_{j=1}^{n_k} \psi_{k,j}^{(0)} ((\psi_{k,j}^{(0)})^\top w_t - T_{k,j}). \quad (26)$$

Using the triangle and Cauchy-Schwarz inequalities, the norm can be bounded as

$$\|\nabla_w \mathcal{L}_{k,\text{task}}(w_t)\|_2 = \left\| \frac{2}{n_k} \sum_{j=1}^{n_k} \psi_{k,j}^{(0)} ((\psi_{k,j}^{(0)})^\top w_t - T_{k,j}) \right\|_2 \quad (27)$$

$$\leq \frac{2}{n_k} \sum_{j=1}^{n_k} \|\psi_{k,j}^{(0)}\|_2 \|(\psi_{k,j}^{(0)})^\top w_t - T_{k,j}\|_2 \quad (28)$$

$$\leq \frac{2}{n_k} \sum_{j=1}^{n_k} (\|\psi_{k,j}^{(0)}\|_2^2 \|w_t\|_2 + \|\psi_{k,j}^{(0)}\|_2 \|T_{k,j}\|_2) \quad (29)$$

$$\leq 2M_\phi(M_\phi R + M_T), \quad (30)$$

where we have used the bounds $\|\psi_{k,j}^{(0)}\|_2 \leq M_\phi$, $\|T_{k,j}\|_2 \leq M_T$, and $\|w_t\|_2 \leq R$.

Next, we bound the gradient of the distillation loss. Each individual term satisfies

$$\|(U_i^{i-1} \psi_{k,j}^{(i)\top})^\top (U_i^{i-1} \psi_{k,j}^{(i)\top} w_t - \psi_{k,j}^{(i-1)\top} w_t)\|_2 \quad (31)$$

$$\leq \|\psi_{k,j}^{(i)}\|_2 \|U_i^{i-1}\|_2$$

$$\times \left(\|U_i^{i-1}\|_2 \|\psi_{k,j}^{(i)}\|_2 \|w_t\|_2 + \|\psi_{k,j}^{(i-1)}\|_2 \|w_t\|_2 \right) \quad (32)$$

$$\leq M_\phi M_U (M_U M_\phi R + M_\phi R) \quad (33)$$

$$\leq M_\phi^2 (M_U + 1)^2 R. \quad (34)$$

Summing over all indices, the gradient norm of the distillation loss is bounded as

$$\|\nabla_w \mathcal{L}_{k,\text{kd}}(w)\|_2 \leq 2(r-1)M_\phi^2(M_U + 1)^2 R. \quad (35)$$

Finally, the norm of the gradient of the regularization term is bounded by $\|\gamma w\|_2 \leq \gamma R$. Combining all three bounds, the total gradient norm is upper-bounded by

$$\|\nabla_w \bar{\mathcal{L}}_k(w)\|_2 \leq 2M_\phi(M_\phi R + M_T) \quad (36)$$

$$+ 2\alpha(r-1)M_\phi^2(M_U + 1)^2 R + \gamma R. \quad (37)$$

This expression defines an upper bound constant, which satisfies

$$\|\nabla_w \bar{\mathcal{L}}_k(w)\|_2 \leq \mathcal{O}(M_\phi M_T + \alpha r M_\phi^2 M_U^2 R + \gamma R). \quad (38)$$

Finally, for the variance of the stochastic gradients, we note that each stochastic sample gradient $\nabla \bar{\mathcal{L}}_k(w; \xi)$ is formed from a single summand and is thus bounded similarly to the full gradient. Applying Jensen's inequality, we obtain

$$\mathbb{E}_\xi [\|\nabla \bar{\mathcal{L}}_k(w; \xi) - \nabla \bar{\mathcal{L}}_k(w)\|_2] \leq 2C, \quad (39)$$

which is also bounded by C up to scalar scale. This completes the proof. \square

Theorem 1 (Global Convergence of RAF). *Choose a stepsize sequence*

$$\eta_t = \Theta\left(\frac{1}{\gamma(E + \alpha r M_U^2 M_\phi^2 / \gamma + t)}\right), \quad (40)$$

so that, once t exceeds the threshold $E + \alpha r M_U^2 M_\phi^2 / \gamma$, the rule satisfies $\eta_t = \Theta(1/(\gamma t))$. Then after T communication rounds the FedAvg optimality gap obeys

$$\mathbb{E}[\mathcal{L}(w_T)] - \min_w \mathcal{L}(w) \leq O(1/T). \quad (41)$$

Proof. Combining Propositions 1 to 3, we identify the exact constants required by the FedAvg analysis of [43]. With these choices, all four technical conditions of [43] are satisfied automatically. See Propositions 1 to 3. \square

Under the explicit constants derived for RAF, Theorem 1 recovers the classical FedAvg rate. Because the multi-resolution distillation term is convex and Lipschitz-smooth, the presence of heterogeneous resolutions does not deteriorate the convergence guarantee, while it does enhance robustness across multiple input scales.

V. EXPERIMENTS

A. Experimental Setup

We evaluated the robustness of the proposed algorithm under resolution drift using a representative high-resolution regression task named Human Pose Estimation (HPE) [10]. This involves localizing keypoints corresponding to human body joints in a person-centric input image. It is formulated as a high-resolution heatmap regression problem, in which the model outputs a set of spatial heatmaps, with each channel corresponding to a specific joint, and encodes the likelihood of that joint appearing at each spatial location. The final joint coordinates are obtained by independently identifying the (x, y) position with the maximum value in each output channel, which results in a set of joint locations equal to the number of target joints.

The MPII dataset [44], which is a standard benchmark for human pose estimation, was used in the experiments. This provides predefined training and validation datasets. In our federated learning setup, each client was configured to hold 4K images sampled separately from the training data, and by default, the trained global model was evaluated using 1.5K validation images unless otherwise specified in the experimental section. As a baseline, all the clients used the same model architecture, following the ViTPose [11] configuration with a small Vision Transformer (ViT-S) backbone. The models were aggregated using the simple FedAvg. Our proposed method builds on this baseline architecture by incorporating a multi-resolution distillation mechanism for each client model. In the loss function, we set weighting coefficients to $\alpha = 1$, and $\gamma = 0.01$. The AdamW optimizer was employed to train the models with batch sizes of 32 for training and 64 for testing. The initial learning rate was set to 2.5×10^{-4} , and the other settings, such as the learning rate decay schedule and data augmentation protocol followed those described in ViTPose [11].

B. Evaluation of Generalization Beyond Training Resolutions

In this section, we demonstrate that our RAF method effectively mitigates the intrinsic limitations of ViT-based models, as discussed in Section III-B and III-C.

Figure 6 illustrates an FL scenario in which four clients are trained exclusively on high-resolution (256×192) data. To assess robustness against resolution drift, we compared our proposed RAF method with the FedAvg-only baseline by evaluating both models across a spectrum of inference

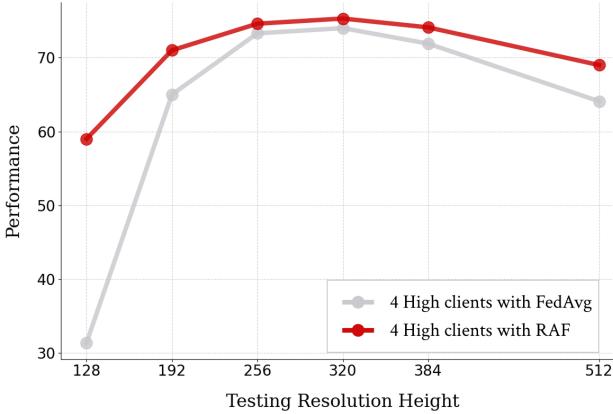


Fig. 6: Inference accuracy on the human pose estimation task at various resolutions when four clients each hold only high-resolution (256×192) datasets in a federated learning setting. The gray curve denotes the baseline using FedAvg aggregation, and the red curve denotes RAF. The x-axis indicates the height of each inference resolution; all resolutions maintain a 4:3 height-to-width aspect ratio.

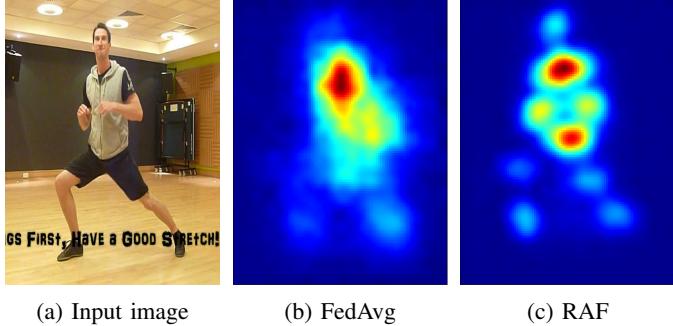


Fig. 7: Heatmap inference comparison on low-resolution (128×96) images. Figure 7a is the input; Figure 7b visualizes the heatmap inferred using weights obtained via FedAvg only from four high-resolution clients.; Figure 7c is a visualization of the heatmap inferred using weights obtained via RAF from four high-resolution clients.

resolutions. As shown, even when all client training data comprised only a single high resolution, RAF achieved substantially better accuracy at both lower and higher inference resolutions. Specifically, the baseline model suffered severe performance degradation to 128×96 and 192×144 , whereas RAF yielded improvements of 27.6% and 6.0%. These results indicate that RAF enables the model to experience multiple resolutions beyond the input resolution of each client by leveraging knowledge distillation with lower-resolution inputs. This afforded a balanced performance across all tested resolutions, including those not seen during training, even when all client input resolutions were the same.

Although Figure 6 highlights the strength of our method, we reinforce this insight using visualized inference heatmaps. Figure 7 compares the heatmaps produced by the FedAvg baseline trained on the four high-resolution clients with those from our RAF-trained model. At low resolution, the baseline model's heatmaps blur together, making it difficult to localize individual joints, whereas the RAF model clearly delineates each joint even under the same downsampling, demonstrating

its superior robustness.

C. Evaluation of the Regularization Effects of RAF

We evaluated RAF under resolution-drift conditions by comparing four configurations: Base (FedAvg), Base (FedProx), RAF (FedAvg), and RAF (FedProx). In each case, the three clients held high- (256×192) , medium- (192×144) , and low (128×96) resolution data. For the RAF settings, we applied our multi-resolution distillation to each client's local model, leaving all other aspects identical to the corresponding baseline. This design isolates RAF's impact on mitigating resolution drift. The results are summarized in Table II.

First, by comparing Base (FedAvg) with Base (FedProx), we observed similar overall performance. FedProx provides a modest +0.4 gain at the 192×144 training resolution but underperforms FedAvg at unseen scales. While FedProx effectively addresses statistical heterogeneity in classification, it fails to mitigate resolution drift and can even introduce instability. By contrast, both RAF (FedAvg) and RAF (FedProx) consistently outperform their baseline counterparts across all resolutions, with particularly large gains at both very low and very high scales. Without any additional data, RAF yields substantial gains over both aggregation schemes: RAF (FedAvg) improves the accuracy by 5.4% at 128×96 and 3.9% at 512×384 , whereas RAF (FedProx) achieves 5.3% and 4.9% improvements at the same resolutions. Our multi-resolution distillation mechanism functions as an effective regularizer by minimizing the gap between the high- and low-resolution predictions. This prevents overfitting to any single resolution and enables RAF to generalize across a broad spectrum of input scales. These results confirm that RAF's enhanced accuracy is primarily driven by the regularizing effect of multi-resolution knowledge distillation, which specifically targets and alleviates the impact of resolution heterogeneity. Furthermore, RAF's compatibility with both FedAvg and FedProx underscores its versatility and ease of integration into existing FL frameworks.

D. Analysis of Client-Side Benefits in RAF

Although RAF serves as an effective regularizer, it introduces a modest computational overhead. A client with a high-resolution dataset has little incentive to participate in FL if the performance gains are negligible. To establish the practical benefits of combining FL with our RAF method, we conducted experiments to compare each client's performance under three settings: (1) Centralized Learning (CL) on their own data, (2) CL with our knowledge distillation (KD) method, and (3) FL with KD (RAF).

Figure 8a, 8b, and 8c illustrate the performance benefits obtained by clients holding images at resolutions 256×192 , 192×144 , and 128×96 , respectively, when participating in the proposed RAF framework. In the high-resolution Figure 8a and mid-resolution Figure 8b, comparing "CL" and "CL+KD" shows that applying KD in a centralized setting improves accuracy at every tested resolution and prevents severe performance degradation on unseen scales. Furthermore, when comparing "CL+KD" with "RAF," we see similar performance gains at all resolutions. This similarity indicates that the improvement

TABLE II

Comparison of federated learning performance among three clients holding datasets with resolutions of 256×192 , 192×144 , and 128×96 . Base and RAF variants are distinguished by whether FedAvg or FedProx is used for aggregation. Red numbers indicate the improvement of “RAF (FedAvg)” over “Base (FedAvg)”, and blue numbers indicate the improvement of “RAF (FedProx)” over “Base (FedProx)”.

Inference Resolution	Base (FedAvg)	Base (FedProx)	RAF (FedAvg)	RAF (FedProx)
128×96 (Seen)	51.8	51.8	57.2 (+5.4)	57.1 (+5.3)
192×144 (Seen)	64.6	65.0	67.1 (+3.5)	67.0 (+2.0)
256×192 (Seen)	68.5	68.5	69.5 (+1.0)	69.7 (+1.2)
320×240 (Unseen)	69.0	68.7	69.9 (+0.9)	70.1 (+1.4)
384×288 (Unseen)	67.5	66.6	69.1 (+1.6)	69.0 (+2.4)
512×384 (Unseen)	60.7	59.5	64.6 (+3.9)	64.4 (+4.9)

is not merely due to having more samples (as in FL), but rather due to our multi-resolution KD method, enabling each client to exploit fully the spatial information in its images. Consequently, even in the federated setting, where the sample number increases naturally, the model’s accuracy increases proportionally because KD has already maximized the spatial feature utilization. In the low-resolution Figure 8c, there is no “CL+KD” curve because no lower-resolution dataset is available for distillation. The gap between “CL” and “RAF” becomes dramatic due to the benefits of both diverse data and resolution-aware training. In particular, as inference resolution increases, the performance advantage of “RAF” over CL grows steadily.

These experimental results demonstrate that regardless of the client-image resolution, our KD method helps extract richer spatial information from those images. Consequently, each client gains a clear performance advantage by participating in FL using the proposed method, justifying the modest computational overhead. This broadens the FL method’s applicability to heterogeneous resolutions.

E. Can Clients with Only Low-Resolution Inference Still Benefit from RAF?

In the previous experiments, we demonstrated that federated learning with the proposed multi-resolution KD method achieved high accuracy across a variety of resolutions. However, extracting the best possible performance from data that each client actually holds differs from using a model that remains robust across all resolutions.

To benefit maximally from the performance improvements observed in earlier experiments, a client trained on low-resolution data must have access to high-resolution images at the time of inference. However, in practice, clients who cannot provide high-resolution data during training typically cannot obtain high-resolution images during inference. For example, in a scenario such as CCTV surveillance, illustrated in the upper row of Figure 9, we present the standard inference process for a low-resolution image. In such cases, where only low-resolution images are available for inference, the predicted keypoint locations remain somewhat blurry because of limited resolution. Consequently, when examining the low-resolution performance at 128×96 in Table III, even a model trained using our method cannot exceed the inherent accuracy limit of 57.2 imposed by the input resolution. Otherwise expressed, notwithstanding the model’s power obtained through FL, its performance is ultimately constrained by the resolution of the data available for inference.

To address this limitation, we conducted additional experiments aimed at boosting performance when only low-resolution data were available. The lower row in Figure 9 illustrates the proposed approach. To obtain a crisper output, we first interpolated the original low-resolution image to a higher resolution (e.g., 256×192) and then ran an inference on this interpolated image. The lower row of Figure 9 shows that the high-resolution heatmap produced hereby clearly pinpoints keypoint locations compared to the directly obtained low-resolution heatmap. Table III presents the quantitative results for the various interpolation methods and resolutions. When

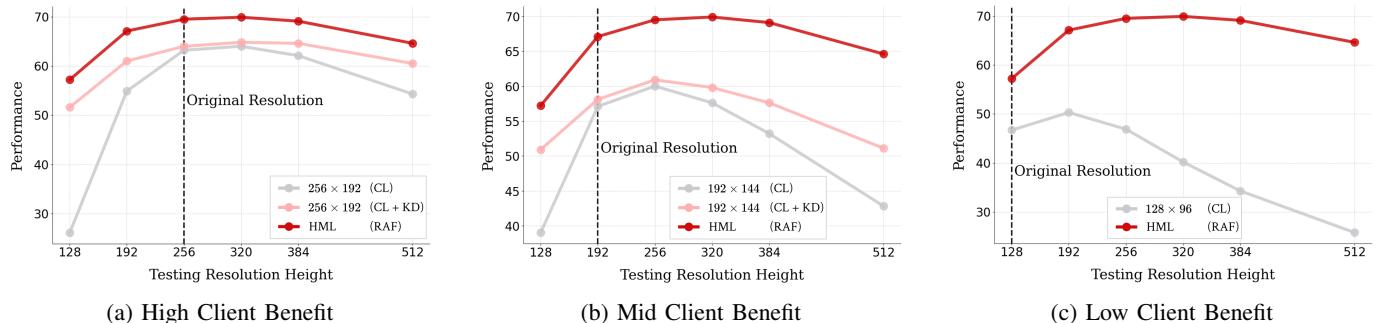


Fig. 8: Inference accuracy on the HPE task across various input resolutions for different training regimes. “CL” denotes centralized learning on a single client, “CL+KD” augments centralized training with our multi-resolution knowledge distillation, and both use 4,000 MPII samples. “HML (RAF)” denotes federated learning with three clients (high 256×192 , mid 192×144 , low 128×96), each holding 4,000 MPII samples (12,000 in total), using our RAF method. The x-axis indicates the input height, with all resolutions maintaining a 4:3 aspect ratio.

TABLE III

Comparison of inference accuracy at low resolution (128×96) using different interpolation methods. The model was trained using RAF across three clients holding high- (256×192), mid- (192×144), and low- (128×96) resolution data. The “Interpolated Res” column indicates the post-interpolation resolution, “*” denotes the original non-interpolated resolution. The results are shown for three interpolation methods: Bilinear, Area, and Bicubic.

Interpolated Res	Bilinear	Area	Bicubic
128×96 (*)	57.2	57.2	57.2
192×144	66.7	66.8	66.9
256×192	69.0	69.2	69.7
320×240	69.0	69.4	69.6
384×288	68.2	68.8	69.0
512×384	63.1	64.5	64.1

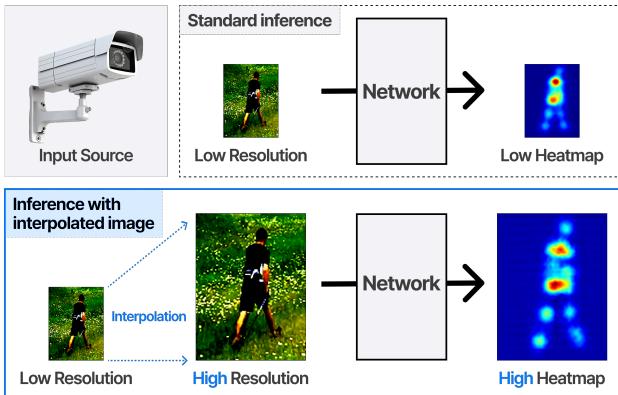


Fig. 9: Inference with an interpolated image. The top row shows standard inference, while the bottom row shows inference on a low-resolution image after increasing its resolution via interpolation for improved performance.

interpolating to 256×192 , the model achieved **69.0**, **69.2**, and **69.7** for Bilinear, Area, and Bicubic, respectively. Interpolating to 320×240 yielded similarly high scores. These accuracies far exceed the standalone centralized learning performance of 46.7 and post-FL low-resolution accuracy of 57.2. This demonstrates that interpolation combined with RAF can achieve significantly improved accuracy compared with an otherwise suboptimal low-resolution dataset.

F. Scalability of RAF: Benefit to a Single High-Resolution Client with Low-Resolution Clients

While a federated learning setup with only low-resolution clients is admittedly unrealistic, we conducted this extreme experiment to determine how far our method can “rescue” a high-resolution client’s performance even under the worst conditions. Figure 10 illustrates a scenario in which one high-resolution client is joined by an increasing number of low-resolution clients (from 0 to 11). All clients were trained on 1,000 MPII images. We report the low- (128×96) and high- (256×192) resolution inference performances as the number of low-resolution clients increases. When there are zero low-resolution clients, the high-resolution client simply trains alone in a centralized fashion. Starting from one low-resolution client, we switched to FL, with RAF applied to each client.

The low-resolution (128×96) plot in Figure 10 shows that every FL configuration (one or more low-resolution clients)

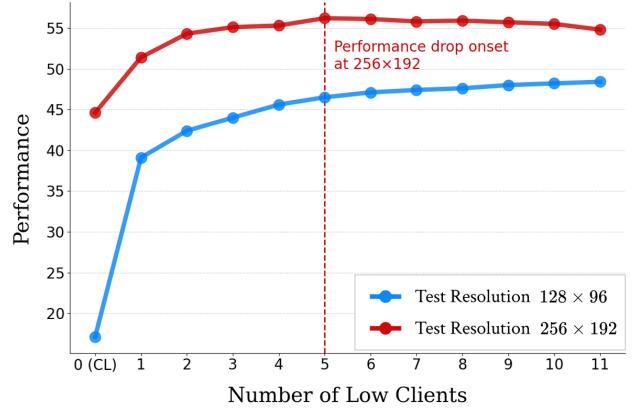


Fig. 10: RAF with one high-resolution (256×192) client and multiple low-resolution (128×96) clients. In this setup, each client has 1,000 samples for training.

significantly outperformed the centralized baseline (zero low-resolution clients) at a low resolution. Recall from Table I that standard FL without the proposed multi-resolution KD actually degrades the performance when aggregating dissimilar resolutions, such as high and low, despite the greater spatial detail in higher-resolution inputs. In contrast, RAF shows that even when only low-resolution images are available during inference, the model learns to leverage every bit of spatial information, demonstrating the effectiveness of our KD scheme. Moreover, as more low-resolution clients were added, the low-resolution performance steadily improved. This implies that a purely low-resolution FL setup still yields benefits for each low-resolution participant, guaranteeing that they gain by joining, even though a high-resolution client is alone among many low-resolution peers.

In the high-resolution (256×192) plot, we observed that all FL variants (with one or more low-resolution clients) outperformed the centralized high-resolution baseline. As we increased the number of low-resolution clients to five, the high-resolution performance continued to increase, peaking when five low-resolution participants were present. Beyond the five low-resolution clients, the performance began to dip slightly. We interpret this as follows: for up to five low-resolution clients, our KD mechanism’s resolution-generalization effect applied to a single high-resolution client dominates and continues to improve the performance. However, when there are more than five low-resolution clients, the KD module cannot fully absorb the vast increase in low-resolution data, leading to a slight overfitting of low-resolution features, and consequently, a minor drop in high-resolution accuracy. In summary, our experiments demonstrate that a single high-resolution client augmented with our KD method can be generalized across up to five low-resolution peers without overfitting. Even when faced with a much larger number of low-resolution clients, the high-resolution client still achieved a much higher performance than it would when training alone in a centralized setting. Therefore, regardless of the adverse effects in the FL scenario, a high-resolution client can guarantee a performance gain by adopting the proposed method.

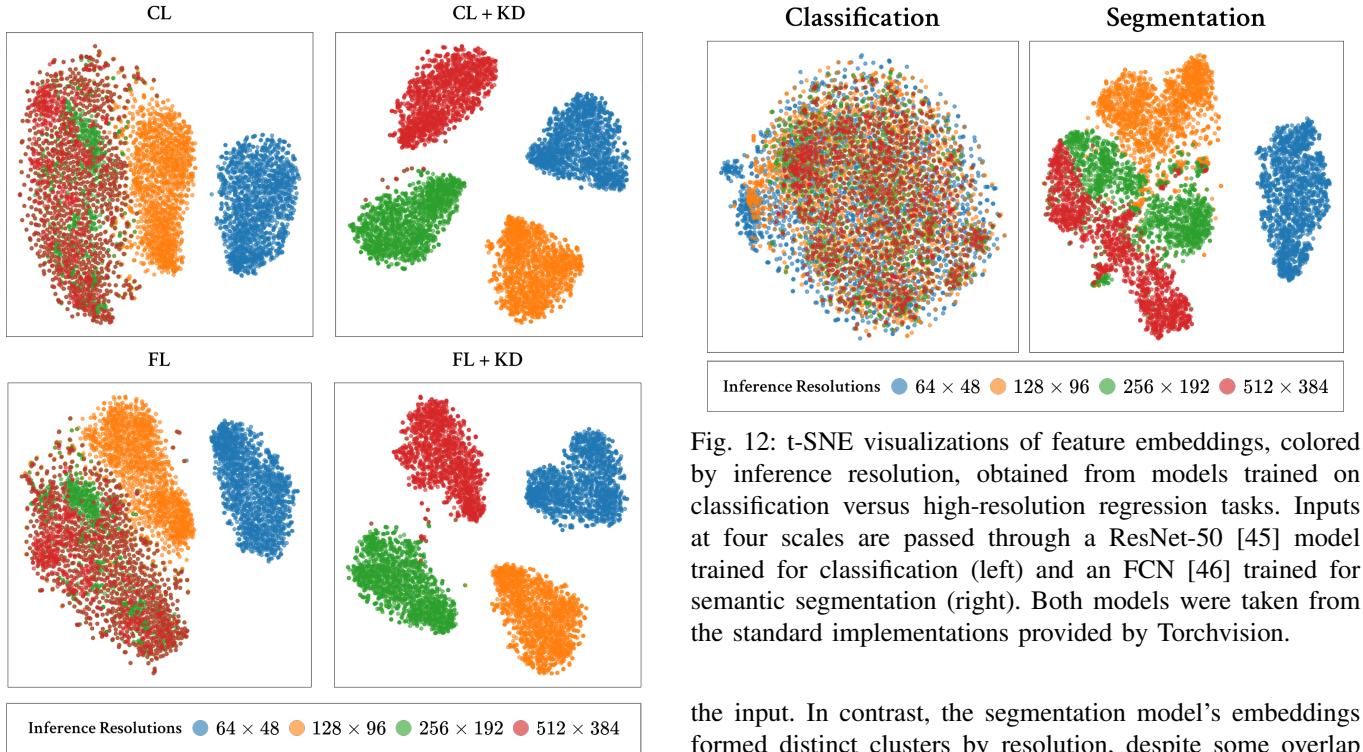


Fig. 11: t-SNE visualization of feature embeddings (from the last ViT block) colored by inference resolution. “CL” denotes centralized learning on a single high-resolution client (256×192); “FL” denotes federated learning with three clients holding high (256×192), mid (192×144), and low (128×96) resolution data. “CL+KD” and “FL+KD” apply our multi-resolution knowledge distillation to the CL and FL setups, respectively.

G. t-SNE Analysis: Resolution Robustness via RAF

a) *Analysis:* Figure 11 shows an intuitive t-SNE analysis that demonstrates the resolution robustness of the models trained using RAF. Without multi-resolution KD, the model struggles to distinguish between the inputs at 256×192 and 512×384 . Although this can separate some of the other scales, it mistakes the largest unseen resolution, (512×384). In contrast, models trained with the proposed multi-resolution KD formed four clearly distinct clusters corresponding to each resolution, both in the centralized and FL settings. This striking separation indicates that RAF both enhances resolution robustness and internally teaches the network to recognize and adapt to different input scales.

b) *Discussion:* In Section III-A, we highlight that classification models discard spatial information from the input image, whereas high-resolution regression tasks must preserve this information, and are therefore sensitive to the input resolution. Although our primary focus was human pose estimation, we further illustrated this fundamental difference using a semantic segmentation task. Figure 12 presents t-SNE visualizations of feature embeddings from models trained on classification and segmentation. The classification model collapsed the features from all resolutions into a single cluster, clearly confirming that it ignored the spatial structure of

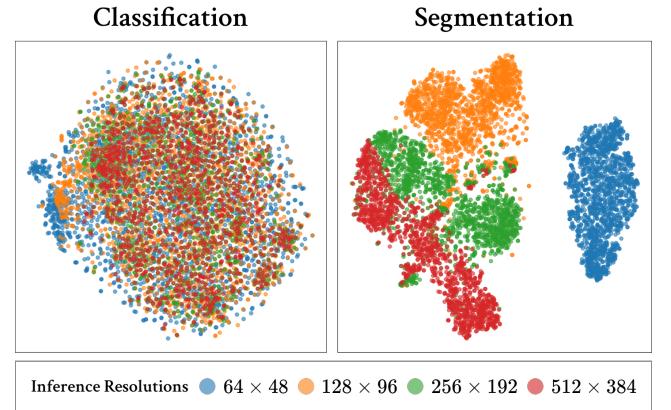


Fig. 12: t-SNE visualizations of feature embeddings, colored by inference resolution, obtained from models trained on classification versus high-resolution regression tasks. Inputs at four scales are passed through a ResNet-50 [45] model trained for classification (left) and an FCN [46] trained for semantic segmentation (right). Both models were taken from the standard implementations provided by Torchvision.

the input. In contrast, the segmentation model’s embeddings formed distinct clusters by resolution, despite some overlap between 256×192 and 512×384 , demonstrating that high-resolution representation models remain sensitive to changes in the input scale and require richer feature representations. This pattern holds for HPE and for segmentation, indicating that the resolution sensitivity we identified is a general property of high-resolution representation models. Consequently, we believe that our RAF framework is applicable beyond human pose estimation and can benefit a wide range of high-resolution representation tasks.

VI. CONCLUSION AND FUTURE WORK

This paper identified a significant performance degradation termed *resolution drift* that occurs when clients with different input resolutions collaborate on high-resolution regression tasks in an FL setting. To address this issue, we proposed and investigated RAF, a framework that augments local training using multi-resolution knowledge distillation. By acting as a resolution-aware regularizer, RAF prevents a model from overfitting to any single scale and substantially improves resolution robustness. Our extensive experiments demonstrated that RAF effectively mitigated resolution drift, and we complemented these empirical findings with a theoretical convergence analysis. Because RAF functions as a local training augmentation process, it is orthogonal to existing FL aggregation algorithms and can seamlessly integrate into a wide range of FL pipelines.

While our evaluation focused on human pose estimation, we believe that the RAF approach potentially extends to many other non-classification tasks. In future work, we will apply RAF to additional high-resolution representation problems, such as semantic segmentation, depth estimation, and super-resolution.

REFERENCES

- [1] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, “Federated learning for iot devices with domain generalization,” *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9622–9633, 2023.
- [2] J. Mills, J. Hu, and G. Min, “Communication-efficient federated learning for wireless edge intelligence in iot,” *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 5986–5994, 2020.
- [3] S. Deng, H. Zhao, W. Fang, J. Yin, S. Dustdar, and A. Y. Zomaya, “Edge intelligence: The confluence of edge computing and artificial intelligence,” *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7457–7469, 2020.
- [4] M. S. ElBamby, C. Perfecto, C.-F. Liu, J. Park, S. Samarakoon, X. Chen, and M. Bennis, “Wireless edge computing with latency and reliability guarantees,” *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1717–1737, 2019.
- [5] W. Y. B. Lim, J. S. Ng, Z. Xiong, J. Jin, Y. Zhang, D. Niyato, C. Leung, and C. Miao, “Decentralized edge intelligence: A dynamic resource allocation framework for hierarchical federated learning,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 536–550, 2022.
- [6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 2017.
- [7] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.
- [8] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” in *Proceedings of the Third Conference on Machine Learning and Systems*, 2020.
- [9] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, “Scaffold: Stochastic controlled averaging for federated learning,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020.
- [10] B. Xiao, H. Wu, and Y. Wei, “Simple baselines for human pose estimation and tracking,” in *15th European Conference on Computer Vision*, 2018.
- [11] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, “Vitpose: Simple vision transformer baselines for human pose estimation,” in *Advances in Neural Information Processing Systems 35*, 2022.
- [12] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep high-resolution representation learning for human pose estimation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- [13] K. Borazjani, P. Abdasarshali, N. Khosravan, and S. Hosseinalipour, “Redefining non-iid data in federated learning for computer vision tasks: Migrating from labels to embeddings for task-specific data distributions,” 2025.
- [14] G. E. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” *CoRR*, vol. abs/1503.02531, 2015.
- [15] H. Mobahi, M. Farajtabar, and P. Bartlett, “Self-distillation amplifies regularization in hilbert space,” in *Advances in Neural Information Processing Systems 33*, vol. 33, 2020, pp. 3351–3361.
- [16] L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng, “Revisiting knowledge distillation via label smoothing regularization,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020.
- [17] R. Tian, Z. Wu, Q. Dai, H. Hu, Y. Qiao, and Y.-G. Jiang, “Resformer: Scaling vits with multi-resolution training,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.
- [18] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, “Tackling the objective inconsistency problem in heterogeneous federated optimization,” in *Advances in Neural Information Processing Systems 33*, 2020.
- [19] Q. Li, B. He, and D. Song, “Model-contrastive federated learning,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.
- [20] C. Jin, X. Chen, Y. Gu, and Q. Li, “Feddyn: A dynamic and efficient federated distillation approach on recommender system,” in *28th IEEE International Conference on Parallel and Distributed Systems*, 2022.
- [21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5998–6008.
- [22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.
- [23] OpenAI, “Gpt-4 technical report,” *CoRR*, vol. abs/2303.08774, 2023.
- [24] H. T. et al., “Llama: Open and efficient foundation language models,” *CoRR*, vol. abs/2302.13971, 2023.
- [25] L. team et al., “Large concept models: Language modeling in a sentence representation space,” *CoRR*, vol. abs/2412.08821, 2024.
- [26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *9th International Conference on Learning Representations*, 2021.
- [27] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- [28] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *IEEE International Conference on Computer Vision*, 2023.
- [29] M. O. et al., “Dinov2: Learning robust visual features without supervision,” *Transactions on Machine Learning Research*, vol. 2024, 2024.
- [30] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems 33*, 2020.
- [31] Z. B. et al., “Audiolm: A language modeling approach to audio generation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2523–2533, 2023.
- [32] Y. Gong, Y.-A. Chung, and J. R. Glass, “Ast: Audio spectrogram transformer,” in *Interspeech 2021*, 2021, pp. 571–575.
- [33] A. R. et al., “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, 2021.
- [34] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *International Conference on Machine Learning*, 2023.
- [35] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Advances in Neural Information Processing Systems 36*, 2023.
- [36] Y. Gao, Z. Hou, C. Yang, Z. Li, H. Yu, and X. Li, “The prospect of enhancing large-scale heterogeneous federated learning with foundation models,” in *IEEE International Conference on Multimedia and Expo*, 2024.
- [37] X. Z. et al., “Fedyolo: Augmenting federated learning with pretrained transformers,” *CoRR*, vol. abs/2307.04905, 2023.
- [38] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, “Deep high-resolution representation learning for visual recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.
- [39] J.-B. G. et al., “Bootstrap your own latent – a new approach to self-supervised learning,” in *Advances in Neural Information Processing Systems 33*, 2020.
- [40] L. Chizat, E. Oyallon, and F. Bach, “On lazy training in differentiable programming,” 2019.
- [41] K. Lyu and J. Li, “Gradient descent maximizes the margin of homogeneous neural networks,” in *International Conference on Learning Representations*, 2019.
- [42] V. Papyan, X. Han, and D. L. Donoho, “Prevalence of neural collapse during the terminal phase of deep learning training,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 40, pp. 24 652–24 663, 2020.
- [43] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of fedavg on non-iid data,” in *8th International Conference on Learning Representations*, 2020.
- [44] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, “2d human pose estimation: New benchmark and state of the art analysis,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.
- [45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
- [46] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.