This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication.
Citation information: DOI: 10.1109/JIOT.2025.3641118, IEEE Internet of Things Journal

KDET-HPFL: A Personalized Federated Learning Framework for Multimodal Pedestrian Detection With Adaptive Feature Selection

Rukai Lan, Yong Zhang, Zidong Wang, *Fellow, IEEE,* Weibo Liu, *Member, IEEE,* and Rui Yang, *Senior Member, IEEE*

***Abstract*—Pedestrian detection plays a critical role in intelligent perception systems in autonomous vehicles, which directly influences the reliability and safety of the overall system. Advanced in-vehicle sensor technology has enabled the continuous evolution of pedestrian detection systems by leveraging heterogeneous multimodal inputs such as RGB, infrared, depth, Light Detection And Ranging, and event data. Nevertheless, establishing a robust pedestrian detection system that is capable of integrating and processing such heterogeneous multimodal data effectively remains a significant challenge. At the same time, growing concerns about data privacy among automobile manufacturers have hindered further advances in detection model performance by restricting the sharing of private data within the industry. In this paper, a novel personalised federated learning framework, Kolmogorov-Arnold network-based Dual Expert Transformer Heterogeneous Personalized Federated Learning (KDET-HPFL), is proposed for multimodal pedestrian detection. To be specific, the KDET pedestrian detector is developed based on an expert feature selection module (which is designed to adaptively choose essential features from multimodal data) and a Group-Rational Kolmogorov-Arnold Network module, which enhances the feature extraction capabilities and improves the detection performance effectively. The HPFL framework is proposed for data privacy protection on heterogeneous multimodal data, where a cross-client aggregation (CCA) method is put forward by integrating different aggregation methods for certain layers in the KDET detector. With CCA, the HPFL framework achieves personalised feature retention of multimodal data pairs on multiple clients and improved model aggregation effect for each client. Experimental findings reveal that the proposed KDET-HPFL framework outperforms some existing personalised federated learning frameworks for pedestrian detection on four public datasets (i.e., LLVIP, STCrowd, InOutDoor, and EventPed) with mAP scores of 73.74%, 75.39%, 66.14%, and 79.57%, respectively.**

***Index Terms*—Pedestrian detection, multimodal fusion, privacy protection, personalized federated learning, mixture of experts.**

I. Introduction

Pedestrian detection has long been a prominent research focus in computer vision due to its broad applicability and critical importance across various applications, such as path planning, video surveillance, and autonomous driving [7]. In autonomous driving, reliable pedestrian detection plays a crucial role in ensuring human safety and enabling intelligent decision-making. System failures or anomalies during operation may lead to serious accidents, which highlights the need for accurate and reliable pedestrian detection systems.

Traditional pedestrian detection methods depend exclusively on RGB images captured by onboard cameras. Note that such RGB-based systems often exhibit degraded performance under challenging conditions, e.g., nighttime driving, backlighting, or extreme weather conditions. By integrating multiple types of sensor data, modern pedestrian detection systems are capable of addressing more diverse driving scenarios compared to single-modal systems. As illustrated in Fig. 1, Light Detection And Ranging (LiDAR) and depth cameras can provide accurate object position information in scenarios with limited visibility. Infrared radiation sensors are capable of capturing discernible silhouette features in dark environments, and event cameras facilitate the acquisition of object information with an extended dynamic range. The modalities complement RGB data by providing positional, thermal, and temporal information that single-modal systems cannot obtain.

Recent advancements in onboard sensing technology have further empowered the exploitation of multimodal data for building reliable and robust pedestrian detection systems [44]. Many efforts have been devoted to developing advanced object detection methods by focusing on multimodal data fusion, e.g., Iterative Cross-Attention Guided Feature Fusion (ICAFusion) [32], Removal and Selection Detector (RSDET) [45], and the Dual Vision Transformer (Dual-ViT) [42]. Among them, Dual-ViT has been widely accepted as a powerful model due to its strong feature fusion ability.

Similar to the original Transformer, the Multilayer Perceptron (MLP) is employed in the Dual-ViT to perform nonlinear transformations and simple feature processing. Despite its simplicity and effectiveness, MLP frequently suffers from low parametric efficiency and feature representation capacity, especially when applying to high-dimensional data for complex tasks. Comparing with MLP, the recently proposed Kolmogorov-Arnold Network (KAN) [33] exhibits competitive performance with better interpretability while handling challenging tasks due to its univariate functions parameterized as splines and learnable activation functions at the edges. However, the B-spline function employed in the KAN is not

R. Lan and Y. Zhang are with the School of Artificial Intelligence and Automation, Wuhan University of Science and Technology, Wuhan, 430081, China, and also with the Engineering Research Center of Ministry of Education for Metallurgical Automation and Inspection Technology, Wuhan, 430081, China. (emails: {iakrulan, zhangyong77}@wust.edu.cn)

Z. Wang and W. Liu are with the Department of Computer Science, Brunel University of London, Uxbridge, Middlesex, UB8 3PH, United Kingdom. (emails: {Zidong.Wang, Weibo.Liu2}@brunel.ac.uk) (*Corresponding author: Zidong Wang*)

R. Yang is with the School of Advanced Technology, Xi'an Jiaotong Liverpool University, Suzhou 215123, China. (email: R.Yang@xjtlu.edu.cn)

well-suited for parallel computing. In response to facilitating parallel computing, Group-Rational KAN (GR-KAN) has been introduced in [43]. It becomes a seemingly natural idea to integrate the GR-KAN into the Dual-ViT with the hope of further improving the feature extraction capability of the Dual-ViT on multimodal data.

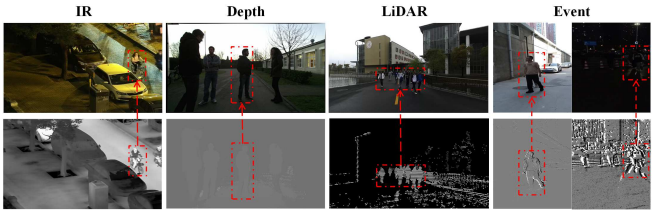

Fig. 1: Characteristics of different multimodal data.

Current multimodal object detection frameworks (including Dual-ViT) often struggle to fuse features from diverse data modalities effectively. Such discrepancies in data distribution and representation would lead to poor detection performance when applied to previously unseen multimodal data [5]. For example, the Dual-ViT (trained based on multimodal data pairs of RGB and event data) may not perform well on multimodal data pairs of RGB and LiDAR data due to spatial misalignment. With the purpose of enhancing the feature fusion ability of the detector, an Expert Feature Selection (EFS) module (which is a novel component developed based on the Mixture of Experts (MoE) [6] structure) is proposed in this paper. By leveraging the dynamic routing capabilities of MoE, the EFS module allows the detector to selectively employ experts for handling data with different modalities, thereby enabling effective feature fusion and improving the overall detection performance [34].

Relying on centralized server training, traditional intelligent transportation systems typically require clients to upload sensitive data to central servers, which results in data silos and privacy risks [38], [41], [47]. Balancing data utilization from all vehicle enterprises while ensuring data security and privacy has become a major challenge in designing a powerful multimodal pedestrian detection framework for autonomous driving. Fortunately, federated learning offers a promising solution by enabling distributed training through parameter sharing, thereby mitigating privacy and security risks without exchanging raw data [18].

Note that the effectiveness of the general federated learning framework (which is designed for homogeneous data) is constrained by the heterogeneity of data modalities collected by different enterprises [17]. The heterogeneity of data modalities limits the applicability of the detector to diverse real-world scenarios. To address the heterogeneity issue, a personalized federated learning strategy has been proposed in [22], which enables each client to perform tasks tailored to its specific requirements while balancing the collaborative training of the global model and the optimization of local models. While current multimodal personalized federated learning frameworks enable model aggregation for designated multimodal data pairs, a unified framework has yet to be thoroughly investigated, which can accommodate model aggregation across clients with diverse multimodal data pairs.

Recent studies in personalized federated learning have demonstrated that hierarchical personalization strategies can effectively improve model performance under modality heterogeneity scenarios [1], [35]. As mentioned in [35], personalized embedding layers have been reported to effectively mitigate both intra-modal and cross-modal discrepancies among clients in multimodal federated learning. Prior studies support embedding-level personalization and demonstrate that the Patch Embedding layer, when properly configured in transformer architectures, is capable of effectively modeling modality-specific properties in multimodal settings [20], [21], [24], [36]. Specifically, the rationality of using the "Patch Embedding" layer close to the input as the personalization layer has been presented in [20], [24]. In addition, the "Patch Embedding" layer, when utilised appropriately in Transformer structures, can successfully capture the properties of multimodal input data [21], [36].

Motivated by the above discussions, this paper identifies three key challenges for developing a pedestrian detection system based on a personalized federated learning framework: 1) How to develop an effective pedestrian detector which is capable of enhancing feature fusion across diverse modalities and modality pairs? 2) How to design a personalized federated learning framework which is adaptable to diverse modality pairs? and 3) How to improve the synergy between the pedestrian detection framework and the personalized federated learning framework? To address the aforementioned challenges, a novel personalised federated learning framework, Kolmogorov-Arnold network-based Dual Expert Transformer Heterogeneous Personalized Federated Learning (KDET-HPFL), is put forward in this paper for pedestrian detection using heterogeneous multimodal data. The main contributions are summarized as follows:

(a) An object detection framework, KDET, is proposed for multimodal pedestrian detection, where the GR-KAN is integrated into the Dual-ViT backbone by replacing the MLP in Dual Block and Merge Block. The learnable edge functions and adaptive spline mapping of the GR-KAN contribute to the improvement of the KDET framework in terms of the feature extraction capability and object detection performance.

(b) A feature fusion module, EFS, is put forward in this paper to facilitate multimodal data analytics. The EFS module utilizes fused multimodal features to guide its gating network in learning dynamic expert weight allocation, and combined with residual connections, enhances the detector's ability to dynamically select and fuse features extracted from different modalities.

(c) An HPFL framework is developed for data privacy protection on heterogeneous multimodal data, which enables cross-client collaborative training while preserving personalized features through a stepwise aggregation mechanism. The cross-client aggregation (CCA) method is proposed based on heterogeneous aggregation and homogeneous aggregation to achieve personalized parameter aggregation of the KDET framework.

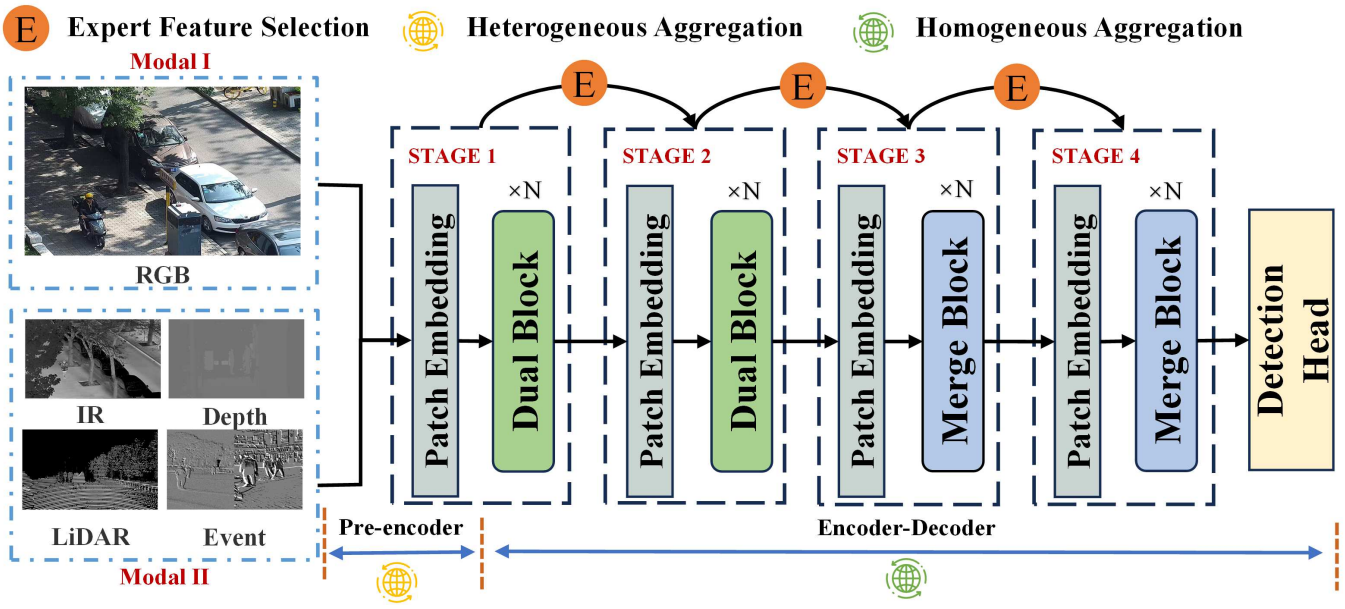

Fig. 2: The overall structure of the proposed KDET-HPFL framework.

(d) A personalized federated-learning-assisted object detection framework, KDET-HPFL, is put forward for multimodal pedestrian detection. Experimental results show that the KDET-HPFL framework outperforms some existing detection frameworks and successfully achieves a proper balance between detection performance and data privacy protection on four public datasets.

The remaining sections of this paper are organized as follows. A novel personalized federated detection framework is introduced in Section II. In Section III, the datasets, evaluation metrics, implementation details, and experimental setting are discussed. In addition, experimental results of the proposed framework and selected methods are presented in Section III. Finally, conclusions and future directions are drawn in Section IV.

II. Methodology

In this section, the proposed KDET-HPFL framework is presented in detail. Firstly, the KDET multimodal pedestrian detection framework is introduced, where the backbone network (i.e., GR-KAN-Based Dual-ViT) of the detection framework and the designed EFS module are discussed. We explain how the GR-KAN is integrated into the Dual-ViT and how the EFS is used to connect each stage within the detection framework. Then, the HPFL framework is introduced with details. The overall structure of the KDET-HPFL framework is depicted in Fig. 2.

A. Multimodal Pedestrian Detection Framework

In this paper, the KDET detection framework is developed for multimodal object detection. To further improve the feature extraction ability of the detection model on multimodal data, the GR-KAN is adopted to replace the MLP layers in the Dual-ViT. To alleviate the problem of insufficient feature fusion caused by the direct connection between Transformer stages, the EFS module is put forward to facilitate a comprehensive integration of multimodal features.

1) GR-KAN-Based Dual-ViT: To achieve a proper balance between the computational cost and the detection performance while adapting to multimodal inputs, the Dual-ViT has been presented in [42] for pedestrian detection. The Dual-ViT receives multimodal data as input and gradually extracts fused features through the dual block and merge block. In the dual block, two modality-specific branches are included. In the merge block, the self-attention [3] module facilitates internal interaction within the feature map. To further improve the feature representation of the detector, we attempt to embed the recently introduced GR-KAN into the Dual-ViT pedestrian detection model in order to improve the feature fusion of multimodal data pairs, as inspired by [43].

In the proposed KDET framework, the GR-KAN is employed to enhance the feature extraction capability of Dual-ViT. To facilitate the embedding of GR-KAN into Dual-ViT, the dual block and the merge block are redesigned. The structures of the new dual block and the new merge block are displayed in Fig. 3.

With the input feature $X \in \mathbb{R}^{H\times W\times C}$, the feature extraction process based on GR-KAN for dual block and merge block can be expressed as follows:

$$\mathrm{z}_0 = \left[\mathrm{M}_{\mathrm{class}}, \mathrm{M}_{\mathrm{p}}^{1}\mathrm{E}, \mathrm{M}_{\mathrm{p}}^{2}\mathrm{E}, \cdots, \mathrm{M}_{\mathrm{p}}^{\mathrm{K}}\mathrm{E}\right] + \mathrm{E}_{\mathrm{pos}}, \tag{1}$$

$$\mathrm{z}'_{\ell} = \mathrm{MSA}(\mathrm{LN}(\mathrm{z}_{\ell-1})) + \mathrm{z}_{\ell-1}, \tag{2}$$

$$\mathrm{z}_{\ell} = \phi(\mathrm{LN}(\mathrm{z}'_{\ell})) + \mathrm{z}'_{\ell}, \tag{3}$$

where M_p denotes the sequences of flattened image patches; M_{class} denotes learnable class token embedding; K is the resulting number of patches; E represents the linear projection layer; E_{pos} denotes the position embedding; z_l is the output feature map of the l-th block; $\mathrm{MSA}(\cdot)$ indicates the Multi-

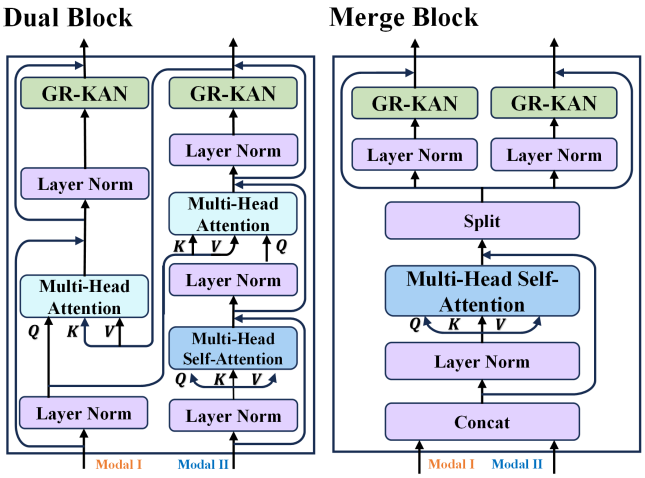

Fig. 3: Design of new Dual block and Merge block structures.

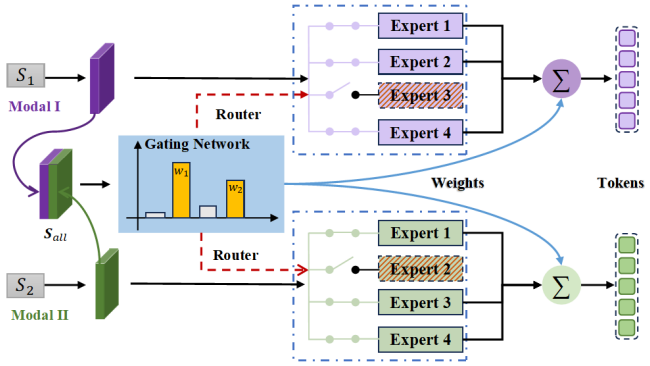

Fig. 4: Structural representation of the EFS module.

head Self-Attention mechanism; $\phi(\cdot)$ represents the GR-KAN block; and $\mathrm{LN}(\cdot)$ denotes the Layer Normalization operation.

Remark 1: In the proposed KDET detection framework, the GR-KAN is employed to replace the MLP in the Dual-ViT model. Comparing with MLP, KAN exhibits superior representation ability and interpretability, which is capable of approximating complex functions with high accuracy while using fewer parameters [23]. Furthermore, with the replacement of the linear weight matrix by a learnable 1D function (i.e., the B-spline), KAN demonstrates better convergence and generalization ability than the MLP. Due to structural limitations, the model training of KAN is slow and unsuitable for distributed training [43]. With the purpose of facilitating parallel computing, the GR-KAN proposed in [43] is adopted in this paper to replace the MLP layers in the Dual-ViT model, leading to a proper balance between flexibility and computational efficiency.

2) EFS Module: The MoE mechanism consists of multiple expert networks and a gating network. The gating network dynamically assigns weights to the experts based on the characteristics of the input multimodal features, which allows the model to activate the most relevant combinations of experts for processing different input features. Motivated by MoE, the EFS module is proposed in this paper to dynamically select features between consecutive blocks, which enables dynamic learning of multimodal data features. In the EFS module, the fused features from two data modalities are employed to guide the gating network to produce the weights for each expert network. The obtained weights are utilized to select the most appropriate combination of experts. With the EFS module, the KDET framework can adaptively activate specific experts when dealing with different multimodal data features with promising feature fusion performance. In this case, the KDET framework effectively accommodates unseen multimodal data, which improves the generalisation ability of the pedestrian detector. The structure of the EFS module is shown in Fig. 4.

The inputs S_1 and S_2 are concatenated to obtain S_{all}, which is then fed into the gating network G. The gating network computes S_{all} according to the probability distribution of N experts. The workflow of EFS's computational procedure for one multimodal data is represented as follows:

$$S_{\mathrm{all}} = \mathrm{Concat}(\mathrm{Flatten}(S_1), \mathrm{Flatten}(S_2)), \tag{4}$$

$$G(S_{\mathrm{all}}) = \mathrm{softmax}\big(\mathrm{topK}(\mathrm{S_{all}} \cdot \mathrm{W_{all}})\big), \tag{5}$$

$$y = \sum_{i=1}^{N} G(S_{\mathrm{all}})_i E_i(x_j), \tag{6}$$

where W_{all} is a learnable weight matrix; E_i denotes expert networks; K indicates the number of activation experts; and x_j represents single data modality features.

Remark 2: Leveraging the MoE mechanism, the proposed EFS module enables the detector to quickly adapt to different multimodal data while effectively reducing the risk of overfitting, thus demonstrating satisfactory generalization ability in the face of unseen modal data. In addition, the residual connection mechanism is embedded into the EFS module to bridge different stages in the Transformer to ensure the complete retention and smooth transmission of original feature information, which could effectively alleviate the problems of feature information loss and vanishing gradient that may occur during the training process of the detector.

B. Personalized Federated Learning Framework

To enable effective personalized feature retention for various multimodal data pairs across different tasks, this paper introduces a CCA method. The CCA method consists of two parts: 1) heterogeneous aggregation and 2) homogeneous aggregation. The detailed aggregation process is illustrated in Fig. 5.

1) Heterogeneous Aggregation: The information interaction among the clients is of vital importance in the personalized federated learning process, which means that different clients (with specific multimodal data) are capable of exchanging knowledge and benefiting from complementary information. In this paper, the KDET detector is divided into a pre-encoder and an encoder-decoder. As shown in Fig. 2, the pre-encoder corresponds to the Patch Embedding part of the first stage of the detector, while the encoder-decoder corresponds to the feature extraction part of the detector in addition to the Patch Embedding.

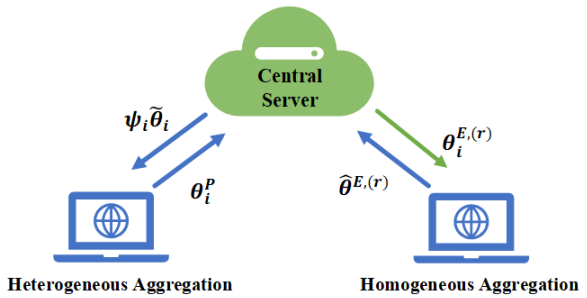

Fig. 5: Illustration of Heterogeneous Aggregation and Homogeneous Aggregation. Here, θ_i^P and $\psi_i\tilde{\theta}_i^P$ respectively represent the parameters transmitted between clients and the central server during the Heterogeneous Aggregation. θ_i^E and $\hat{\theta}^E$ respectively denote the parameters transmitted between clients and the central server during the Homogeneous Aggregation.

Note that multimodal data (i.e., RGB, infrared, depth, LiDAR, and event data) collected from advanced sensors by different vehicle manufacturers exhibit distinct characteristics. In real-world scenarios, the availability of the aforementioned modalities often varies significantly across clients, which brings considerable challenges related to data heterogeneity and fusion [9], [13]. To address the challenge of heterogeneity of multimodal data between different clients, the cross-attention mechanism is deployed to facilitate effective information exchange between clients. Heterogeneous aggregation is proposed in this paper for the pre-encoder, where the central server maintains a dedicated set of weights for each client. Heterogeneous aggregation considers the discrepancy between multimodal data among the local updates of different clients, which makes parameter updates influenced by interactions across multiple clients.

The client's model parameter update process is described as follows:

$$C = [\Delta\theta_1^P, \ldots, \Delta\theta_N^P]^T, \tag{7}$$

$$\tilde{\theta}_i^P = \mathrm{Softmax}\left(\frac{\Delta\theta_{\mathrm{i}}^{\mathrm{P}}\mathrm{C}^{\mathrm{T}}}{\sqrt{\mathrm{d}}}\right)\mathrm{C}, \tag{8}$$

$$\theta_i^{P,(r)} = \theta_i^{P,(r-1)} + \psi_i^{(r)}\tilde{\theta}_i^P, \tag{9}$$

where $\Delta\theta_i^P$ represents the original parameter update; d is the dimension of $\Delta\theta_i^P$; $\tilde{\theta}_i^P$ represents the aggregated update for the i-th pre-encoder; $\theta_i^{P,(r)}$ denotes the pre-encoder parameters of the i-th client at round r; and $\psi_i^{(r)}$ is the parameter of the i-th client, which is learnable. Here, $\psi_i^{(r)}$ is dynamically updated during the training process, which ensures adaptive optimization by integrating personalized weights with the overall optimization objective. The parameter $\psi_i^{(r)}$ in (9) is a learnable parameter that is updated during model training phase as follows:

$$\psi_i^{(r)} = \psi_i^{(r-1)} - \eta\frac{\partial\mathcal{L}_i}{\partial\psi_i}, \tag{10}$$

where η is the learning rate; and $\mathcal{L}_i$ is the local loss function of the i-th client. The parameter ψ_i is constrained to the range $[0, 1]$. According to the characteristics of its client data, ψ_i adaptively learns the optimal degree of personalization.

2) Homogeneous Aggregation: In addition to implementing heterogeneous aggregation for the pre-encoder, homogeneous aggregation is put forward for the encoder-decoder, which allows all participating clients to share information through a parameter-sharing mechanism, thereby obtaining common feature parameters across clients. Homogeneous aggregation effectively leverages the correlations among individual clients, which enhances the generalization of the detection model.

Assume that there are N clients involved in training, and each client has an independent local detector. The parameter aggregation process of the encoder-decoder architecture is expressed by:

$$\hat{\theta}^{E,(r)} = \frac{1}{N}\sum_{i=1}^{N}\theta_i^{E,(r)}, \tag{11}$$

where $\theta_i^{E,(r)}$ denotes the independent encoder-decoder parameters of the i-th client at round r; $\hat{\theta}^{E,(r)}$ denotes the aggregated parameters of each client. The global detection model is updated by summing gradients derived from various multimodal data pairs, thereby learning from multiple data domains.

C. The overall KDET-HPFL Framework

This paper proposes the KDET-HPFL framework for pedestrian detection, which enables the training of a personalised pedestrian detection framework adapted to the specific data characteristics of each client while preserving data privacy. The detector is divided into a pre-encoder and an encoder-decoder, which are adaptively trained using heterogeneous aggregation and homogeneous aggregation strategies, respectively. The main steps of the proposed KDET-HPFL framework are presented in Algorithm 1.

III. Experiments and Analysis

This section begins by introducing the utilized benchmark datasets and experimental settings. Then, the evaluation of the KDET pedestrian detection model is conducted by comparing it with existing multimodal object detection methods. Next, the proposed KDET-HPFL framework is compared with several selected personalized federated learning frameworks to demonstrate its effectiveness in handling heterogeneous multimodal data. The ablation experiments are carried out to validate the effectiveness of the proposed modules and frameworks. Comprehensive discussions are presented to analyse the experimental results.

A. Datasets and Evaluation metrics

To validate the effectiveness of the proposed KDET-HPFL framework, Average Precision (AP) and Log-average Miss Rate (MB^{-2}), are used for comprehensive performance evaluation [32]. The proposed method is validated on four public datasets, including LLVIP [15], STCrowd [2], InOutDoor [28], and EventPed [44]. The LLVIP public dataset is a visible-infrared paired dataset designed for low-light vision tasks, which consists of 33,672 images (16,836 pairs). The STCrowd public dataset comprises 219K annotated pedestrian instances,

Algorithm 1 Main Steps of the KDET-HPFL Framework.

Input: N clients $\{C_1, \ldots, C_N\}$ with private multimodal datasets $\{D_1, \ldots, D_N\}$ for pedestrian detection, personalized federated learning iteration number R, local iteration number E, client learning rate η, aggregation interval t

Output: Trained models $\Theta(R) = \{\theta_1^{(R)}, \ldots, \theta_N^{(R)}\}$

Clients initialize KDET frameworks $\Theta(0) = \{\theta_1^{(0)}, \ldots, \theta_N^{(0)}\}$, each model θ_i consists of a shared encoder-decoder $\theta_{i,E}$ and pre-encoder $\theta_{i,P}$
Initialize local iteration counter $k \leftarrow 0$
procedure FEDERATED TRAINING
  for each global round $r \in \{1, \ldots, R\}$ **do**
    for each client C_i in parallel **do**
      $k \leftarrow k + 1$
      $\Delta\theta_i^{(r)} \leftarrow \mathrm{LOCAL\_TRAINING}(\theta_i^{(r-1)})$
      if $k = t$ **then**
        Server disassembles $\Delta\theta_i^{(r)}$ into:
          - Pre-encoder updates $\Delta\theta_{i,P}$
          - Encoder-decoder updates $\Delta\theta_{i,E}$
        Homogeneous Aggregation to $\{\Delta\theta_{i,E}\}$
        Heterogeneous Aggregation to $\{\Delta\theta_{i,P}\}$
        Broadcast aggregated model $\Theta^{(r)}$
        $k \leftarrow 0$
      end if
    end for
  end for
end procedure
procedure LOCAL_TRAINING($\theta_i^{(r-1)}$)
  $\theta_i \leftarrow \theta_i^{(r-1)}$
  for each local epoch $e \in \{1, \ldots, E\}$ **do**
    for batch $B_{i,e} \subseteq D_i$ **do**
      Compute model losses L_i
      Update local model $\theta_i \leftarrow \theta_i - \eta \nabla_{\theta_i} L_i$
    end for
  end for
  return $\Delta\theta_i^{(r)} = \theta_i - \theta_i^{(r-1)}$
end procedure

averaging 20 individuals per frame, and encompasses varying degrees of occlusion. The EventPed public dataset is a recently collected RGB-event paired dataset focusing on pedestrian detection in outdoor scenarios such as parks and sidewalks. The dataset includes 7,195 image pairs in the training set and 2,435 image pairs in the test set. The InOutDoor public dataset contains 8605 annotated RGB-D frames collected by the robot at a frame rate of 30 Hz.

B. Implementation Details

The proposed KDET-HPFL framework uses Dual-ViT as the backbone network. In this paper, we pre-train the Dual-ViT model for 12 epochs on a combined RGB-based dataset. The pre-trained model is then trained on multimodal datasets. The KDET-HPFL framework in this paper is implemented using the MMDetection framework, known for its flexibility and modularity. The experimental platform is CUDA 11.1 and GTX 3090 GPU $\times 8$ with data parallelism, and the code is implemented using Pytorch 1.8.1.

C. Experimental Setting

The comparison study in this paper can be divided into two parts, the pedestrian detection experiment and the personalized federated learning experiment. Furthermore, an ablation study is conducted on four public datasets (in which each modification rule is implemented separately) to verify the effectiveness of the proposed modules.

Pedestrian Detection Experiment: In this experiment, we focus on the pedestrian detection task using Dual-ViT as the backbone network. The model incorporates multi-head attention mechanisms and hierarchical embeddings for feature extraction. For optimization, we employ the AdamW optimizer with an initial learning rate of 1e-4 and a weight decay coefficient of 1e-4. To achieve differentiated optimization, a learning rate reduction factor of 0.1 is applied to the backbone network. The total loss function of the detector comprises four key components: the RPN head loss ($\mathrm{Loss_{RPN}}$), the Query head loss ($\mathrm{Loss_{Query}}$), the RoI head loss ($\mathrm{Loss_{RoI}}$), and the ATSS head loss ($\mathrm{Loss_{ATSS}}$). In such heads, Focal loss, CrossEntropy loss, and Quality Focal loss are used as categorical loss. L_1 loss and GIoU Loss are used as regression loss [44]. The total loss function is obtained by computing the weighted sum of the aforementioned individual loss components:

$$\mathrm{Loss} = \mathrm{Loss_{RPN}} + \mathrm{Loss_{Query}} + \mathrm{Loss_{RoI}} + \mathrm{Loss_{ATSS}}. \quad (12)$$

Personalized Federated Learning Experiment: This experiment is designed based on the KDET pedestrian detection framework to validate the data privacy protection performance. The public datasets (i.e., LLVIP, STCrowd, InOutDoor, and EventPed) are adopted separately as multimodal training data for each individual client. To ensure the data balance, random sampling is employed to establish the training sets with 4,000 samples each, and the validation sets with 1,000 samples each. Within the federated learning framework, the training configurations for individual clients remain consistent with the configurations in the pedestrian detection experiment. For model aggregation, we implement a strategy where model parameters are aggregated after every 2,000 iterations per client, with a total of 30 aggregation rounds throughout the training process.

D. Results and Discussions of Detection Framework

To validate the effectiveness of the proposed pedestrian detection framework, comparative experiments are carried out with some mainstream multimodal object detection models, including YOLOv8 [37], Faster R-CNN [31], YOLOX [12], Co-DETR [46], Swin Transformer [19], ICAFusion [32], Dual-ViT [42] and RSDet [45]. The evaluation process employed three metrics: AP and MB^{-2}. To thoroughly validate the generalization ability and robustness of the KDET pedestrian detection framework across different scenarios, four distinct datasets are employed for model training and testing, which enables a comprehensive assessment of the model's pedestrian

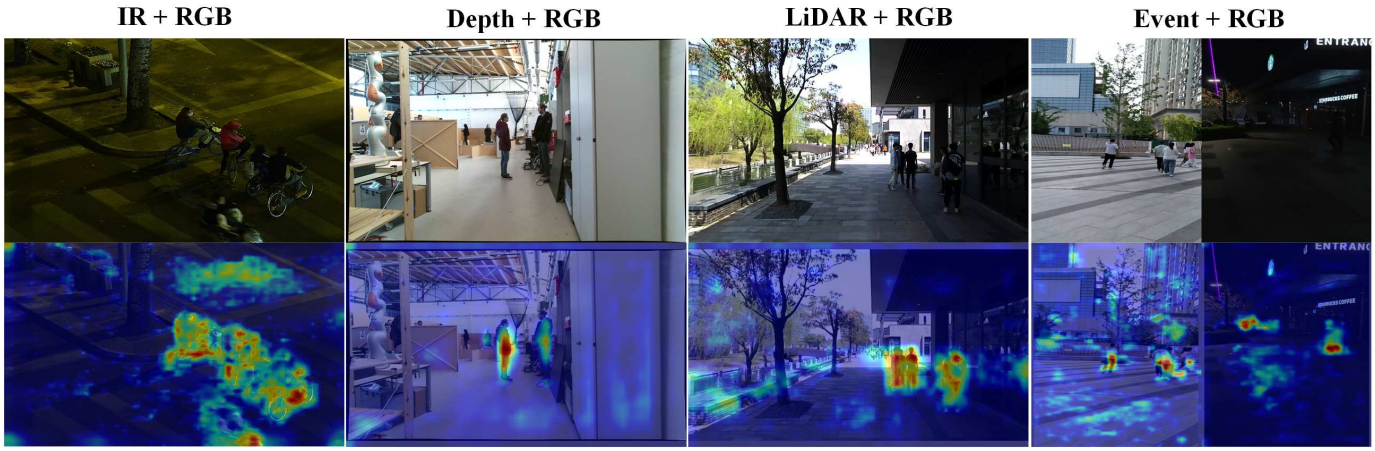

Fig. 6: Visualization of features extracted by the KDET framework.

detection performance on multimodal data. The results are shown in Table I. The pedestrian detection framework KDET proposed in this paper has achieved the best performance on all four test datasets. The KDET's mAP improves by 3.78%, 3.74%, 3.36%, and 4.11% compared to the best-performing comparison method RSDet on the LLVIP, STCrowd, InOutDoor, and EventPed public datasets, respectively.

TABLE I: Comparisons of multimodal-based pedestrian detection methods in recent years. ↑ means higher is better, while ↓ means lower is better.

Method	LLVIP		STCrowd		InOutDoor		EventPed	
	mAP ↑	MB^{-2} ↓	mAP ↑	MB^{-2} ↓	mAP ↑	MB^{-2} ↓	mAP ↑	MB^{-2} ↓
YOLOv8 [37]	51.32	32.65	54.18	34.10	45.20	43.10	53.57	41.80
Faster R-CNN [31]	53.06	31.80	55.41	32.50	46.05	41.97	54.11	39.47
YOLOX [12]	59.13	28.78	62.14	25.34	52.07	37.56	63.86	29.27
Co-DETR [46]	56.74	36.92	59.76	27.05	49.80	38.95	59.98	33.10
Swin Transformer [19]	59.66	28.93	63.49	26.78	50.94	38.40	61.40	31.29
ICAFusion [32]	54.75	30.78	56.04	33.40	47.37	40.59	55.44	40.73
Dual-ViT [42]	60.47	27.91	64.06	26.17	53.78	35.94	62.11	25.51
RSDet [45]	61.30	26.02	64.11	24.68	55.14	35.74	65.69	25.26
Ours	**65.08**	**24.02**	**67.85**	**22.74**	**58.50**	**32.50**	**69.80**	**21.40**

To evaluate the feature extraction capabilities of the proposed KDET framework, visualization analysis of the extracted features is conducted. Results of the visualization analysis are illustrated in Fig. 6. We can see that KDET consistently captures key pedestrian features across diverse scenarios (including daytime, nighttime, indoor, and outdoor environments) and multiple datasets, which indicates that the proposed KDET framework is capable of effectively extracting key features robustly.

The detection results of KDET framework on the benchmark datasets are shown in Fig. 7, which verifies the generalization ability of the KDET framework. The detection results show that the KDET framework is able to accurately identify objects in different multimodal data pairs and take advantage of the complementary strengths of the different modalities to better perform the detection task. For example, in the IR and RGB data pairs, IR data can provide additional information to make up for the deficiencies caused by light and shadow interference, thus enabling the detector to perform the detection task more accurately. Similarly, in Event and RGB data pairs, RGB images are difficult to adequately represent the object features at night, and with the supplement of Event data, the detector can also better realize the detection task.

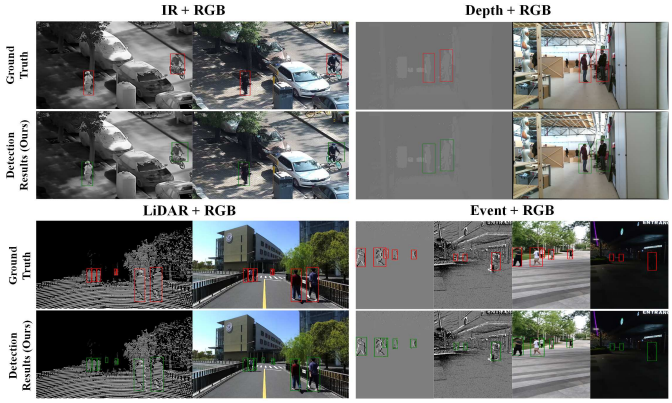

Fig. 7: Visualization of pedestrian detection results of four multimodal data pairs.

To further evaluate the learning ability of the KDET framework on the pedestrian object features, t-SNE visualization is performed on the test set of the LLVIP public dataset, mapping the high-dimensional features of each sample to the two-dimensional plane. The corresponding results are displayed in Fig. 8. It can be seen in Fig. 8 that after adding the EFS module and GR-KAN structure, the feature extraction performance of the baseline is significantly improved, which is reflected by a tighter distribution of features and clearer differentiation. This finding indicates that the KDET framework is able to distinguish the features of the test samples more accurately. In Fig. 8(d), we analyze the distribution of outlier points, the number of outliers is significantly reduced compared with Fig. 8(a), which further shows that most of the features of the test samples can be effectively distinguished by the KDET framework.

E. Results and Discussions of HPFL Framework

To validate the performance of the HPFL framework, four public datasets are assigned to each of the four simulated client training nodes, where the 1st node is designated as the central server responsible for parameter aggregation and distribution. All the clients involved in training are distributed

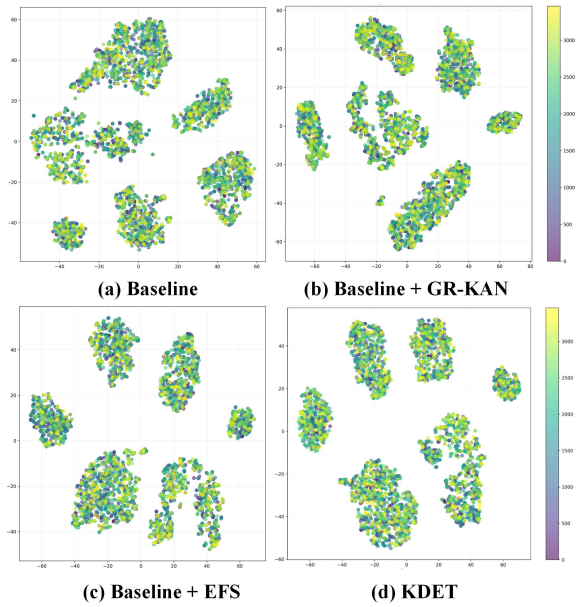

Fig. 8: Visualization of features extracted by different methods based on t-SNE.

according to the aforementioned experimental setup. Table II shows the experimental results of the proposed HPFL framework and chosen federated learning frameworks, including the traditional federated learning frameworks FedAvg [27] and DeceFL [41], and the personalised federated learning frameworks FedProx [25], FedAMP [14], FedRep [4], and Ditto [26].

As demonstrated in Table II, the HPFL framework outperforms the chosen federated learning frameworks on all four datasets. Specifically, our framework achieves the highest detection accuracy to both traditional federated learning frameworks and personalised federated learning frameworks. Compared to the best-performing baseline framework FedAMP, the HPFL framework achieves absolute improvements of 5.13%, 5.35%, 5.19%, and 8.71% in mAP on LLVIP, STCrowd, InOutDoor, and EventPed public datasets, respectively. Results demonstrate the robustness and generalisation ability of our proposed framework on multimodal data. To evaluate the

TABLE II: Comparison to representative methods using four different datasets each assigned to a client.

Method	Client 1 LLVIP	Client 2 STCrowd	Client 3 InOutDoor	Client 4 EventPed
FedAvg [27]	51.47	42.65	46.18	58.12
DeceFL [41]	49.22	40.19	42.02	55.71
FedProx [25]	55.13	46.78	49.74	62.54
FedAMP [14]	68.61	70.04	60.95	70.86
FedRep [4]	66.73	68.46	59.17	69.39
Ditto [26]	61.53	56.68	55.51	67.10
Ours	**73.74**	**75.39**	**66.14**	**79.57**

effectiveness of the proposed HPFL framework under varying numbers of clients, we add more testing using a range of client counts. In this experiment, the performance is measured by calculating the mAP for each local client. The mAP is computed independently for each client (without parameter aggregation) using the following formula:

$$\Delta_p = \frac{1}{N}\sum_{i=1}^{N}\frac{P_{\mathrm{HP,i}} - P_{\mathrm{Local,i}}}{P_{\mathrm{Local,i}}}, \tag{13}$$

where N is the number of clients; $P_{\mathrm{HP,i}}$ and $P_{\mathrm{Local,i}}$ correspond to the efficiency of client i for federated learning methods and the local model, respectively.

The number of clients involved in training is gradually increased from one to four using LLVIP, STCrowd, InOutDoor, and EventPed datasets. As demonstrated in Fig. 9, the proposed KDET-HPFL framework consistently outperforms the compared methods across all configurations. We find that the increase of number of clients will lead to improved detection performance. This finding demonstrates that federated learning enhances model performance while safeguarding data security. Furthermore, by accounting for the heterogeneity of different multimodal data, the KDET-HPFL framework achieves more significant performance gains as the number of training clients increases.

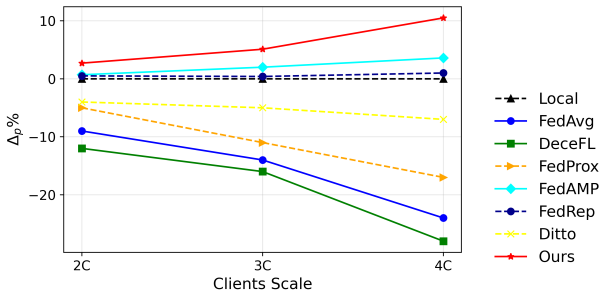

Fig. 9: Comparison of the impact of client additions on the detection performance of frameworks.

To demonstrate the practicality of the HPFL framework proposed in this paper, we have compared the total communication overhead required for different federated learning frameworks to achieve mAP value of 50%, 60%, and 70%. Table III clearly demonstrates the superiority of our method. For client-specific models on the LLVIP public dataset, achieving an mAP value of 50% requires a total communication cost of 7.41 GB using the HPFL framework, which is lower than the communication overhead of other centralized federated learning frameworks. In comparison with total communication overhead of FedRep (17.66 GB), FedAMP (18.54 GB), and Ditto (33.37 GB), our method requires a total communication overhead of 11.12 GB when the detection model for specific clients using the LLVIP public dataset with an mAP value of 60%. Note that in our experiments, only our federated learning framework can achieve mAP value of 70%. In summary, in the scenario of personalization of multimodal data, our HPFL framework achieves the optimal performance with the minimal total communication overhead compared to other centralized federated learning frameworks. This highlights the practical value of HPFL framework in real-world IoT scenarios.

The convergence of three detection models based on KAN, MLP and GR-KAN in federated learning is studied using a

TABLE III: Comparison of total communication overheads for different federated learning frameworks to achieve specified mAP value.

Method	Per-Round Comm. (MB)	Total Comm. to 50% mAP (GB)	Total Comm. to 60% mAP (GB)	Total Comm. to 70% mAP (GB)
FedAvg	474.5	25.95	-	-
FedProx	474.5	22.24	-	-
FedAMP	474.5	11.12	18.54	-
FedRep	205.5	8.03	17.66	-
Ditto	474.5	14.83	33.37	-
Ours	**474.5**	**7.41**	**11.12**	**25.95**

distributed training framework. The trends of the mAP changes on the four clients, and the results are shown in Fig. 10. We can see that the GR-KAN-based KDET-HPFL framework achieves the optimal performance among the KAN-based KDET-HPFL framework and the MLP-based KDET-HPFL framework. Furthermore, our GR-KAN-based KDET-HPFL framework not only speeds up the training process but also maintains better detection performance in terms of mAP across all experimental settings. To summarize, GR-KAN-based KDET-HPFL framework consistently outperforms the KAN-based KDET-HPFL framework and the MLP-based KDET-HPFL framework while exhibiting stable post-convergence behaviour.

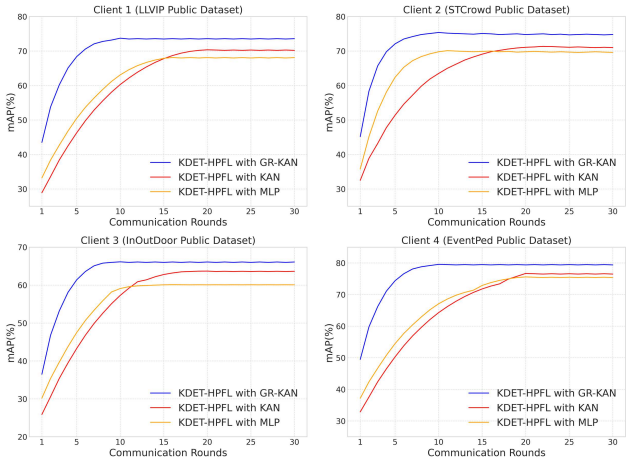

Fig. 10: Comparison of mAP change trends of four clients under different frameworks.

F. Ablation Study

The effectiveness of the designed modules in the proposed KDET framework is demonstrated in the ablation experiments on the LLVIP, STCrowd, InOutDoor, and EventPed public datasets. The ablation experimental results are presented in Table IV. According to the experimental results, the influences of each module in the proposed KDET framework are summarized below.

(a) The influence of the EFS module is presented in the second row of Table IV. By employing the EFS module, the detector's mAP improves by 3.67%, 3.05%, 3.41%, and 4.99% compared to the baseline (Dual-ViT) on the LLVIP, STCrowd, InOutDoor, and EventPed public datasets, respectively.

(b) The influence of the GR-KAN module is shown in the third row of Table IV. Compared with the baseline (Dual-ViT), the GR-KAN module enhances the detector's mAP by 2.01%, 1.91%, 1.45%, and 3.62% on LLVIP, STCrowd, InOutDoor, and EventPed public datasets, respectively.

The KDET framework (which includes the EFS and the GR-KAN modules) achieves the best overall detection performance with significant improvements in mAP of 4.61%, 4.79%, 4.72%, and 7.69% across all datasets, showing the effectiveness of our proposed architecture.

TABLE IV: Results of ablation experiments for KDET detection framework.

Module			Dataset			
EFS	GR-KAN	Method	LLVIP	STCrowd	InOutDoor	EventPed
×	×	Dual-ViT (Baseline)	60.47	64.06	53.78	62.11
✓	×	KDET (without GR-KAN)	64.14	67.11	57.19	67.10
×	✓	KDET (without EFS)	62.48	65.97	55.23	65.73
✓	✓	KDET	**65.08**	**68.85**	**58.50**	**69.80**

Remark 3: Based on the above comparative experimental results, we can conclude that the proposed EFS module for dynamic feature selection can effectively enhance the feature extraction capability of the model. In addition, the introduction of GR-KAN structure instead of MLP in Transformer block not only improves the detection performance of the model, but also meets the demand of distributed training. The HPFL framework proposed in this paper effectively achieves personalized aggregation on each client on the basis of protecting data privacy, and significantly improves the performance of the clients in the object detection task. In summary, the effectiveness of the proposed KDET-HPFL framework is comprehensively demonstrated on the LLVIP, STCrowd, InOutDoor, and EventPed public datasets.

IV. Conclusion

In this paper, the KDET-HPFL framework has been proposed for multimodal pedestrian detection with preserved data privacy in the field of autonomous driving. To be specific, the detection framework, KDET, has been developed for multimodal pedestrian detection, where the GR-KAN has been employed to enhance the feature extraction capability of the detector. The EFS module has been put forward to enhance the multimodal feature extraction and fusion capabilities of the detector using dynamic feature selection and residual connection. The personalised federated learning framework, HPFL, has been proposed to guarantee the data protection of the detection framework and address the challenges of multimodal data processing in intelligent perception systems. The HPFL framework incorporates both heterogeneous and homogeneous aggregation mechanisms to enable cross-client collaborative training while preserving personalized features. The developed KDET-HPFL framework has been evaluated on four public datasets and demonstrated superior performance compared to existing frameworks. Experimental results have shown that our KDET-HPFL framework effectively handles multimodal feature dynamic fusion, while the KDET-HPFL framework enables personalized federated learning across multiple heterogeneous multimodal data clients. We can conclude

that the proposed KDET-HPFL framework further ensures the safety of autonomous driving under complex scenarios. In the future, we aim to: 1) extend the framework to other tasks (e.g., state estimation [16], dual-task learning [29], and representation learning [40]); 2) reduce the communication and computation overhead of the HPFL framework and improve its communication efficiency [8], [30]; and 3) explore advanced personalization strategies by integrating optimization algorithms [11]) and novel feature extraction methods [10], [39].

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grants 62273264 and 61933007, the Engineering and Physical Sciences Research Council (EPSRC) of the UK, the Royal Society of the UK, and the Alexander von Humboldt Foundation of Germany.

REFERENCES

[1] M. Arivazhagan, V. Aggarwal, A. Singh and S. Choudhary, Federated learning with personalization layers, *arXiv preprint arXiv:1912.00818*, 2019.

[2] P. Cong, X. Zhu, F. Qiao, Y. Ren, X. Peng, Y. Hou, L. Xu, R. Yang, D. Manocha and Y. Ma, Stcrowd: A multimodal dataset for pedestrian perception in crowded scenes, In: *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, New Orleans, USA, Jun. 2022, pp. 19608-19617.

[3] H. Chen, P. Wu, W. Wen and N. Zeng, DLA-Net: A dynamically learnable attention network for intelligent surface visual inspection of aero-engine blades, *IEEE Transactions on Instrumentation and Measurement*, vol. 74, art. no. 3532114, 2025.

[4] L. Collins, H. Hassani, A. Mokhtari and S. Shakkottai, Exploiting shared representations for personalized federated learning, In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Online, Jul. 2021, pp. 2089-2099.

[5] Z. Chen, L. Zhang, J. Tang, J. Mao and W. Sheng, Conditional generative adversarial net based feature extraction along with scalable weakly supervised clustering for facial expression classification, *International Journal of Network Dynamics and Intelligence*, vol. 3, no. 4, art. no. 100024, Dec. 2024.

[6] Z. Chen, Y. Shen, M. Ding, Z. Chen, H. Zhao, E. G. Learned-Miller and C. Gan, Mod-squad: Designing mixtures of experts as modular multi-task learners, In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, Jun. 2023, pp. 11828-11837.

[7] P. Dollar, C. Wojek, B. Schiele and P. Perona, Pedestrian detection: An evaluation of the state of the art, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743-761, 2011.

[8] W. Diao, W. He, K. Liang and X. Tan, Adaptive impulsive consensus of nonlinear multi-agent systems via a distributed self-triggered strategy, *International Journal of Systems Science*, vol. 55, no. 11, pp. 2224–2238, 2024.

[9] J. Dou, Y. Song and H. Yu, Hierarchical oversampling based on Cohen's criterion for imbalanced data with missing information, *IEEE Transactions on Computational Social Systems*, in press, DOI: 10.1109/TCSS.2025.3548874.

[10] W. Ehab, L. Huang and Y. Li, UNet and variants for medical image segmentation, *International Journal of Network Dynamics and Intelligence*, vol. 3, no. 2, art. no. 100009, Jun. 2024.

[11] W. Fang, B. Shen, A. Pan, L. Zou and B. Song, A cooperative stochastic configuration network based on differential evolutionary sparrow search algorithm for prediction, *Systems Science & Control Engineering*, vol. 12, no. 1, art. no. 2314481, 2024.

[12] Z. Ge, S. Liu, F. Wang, Z. Li and J. Sun, YOLOX: Exceeding YOLO series in 2021, *arXiv preprint arXiv:2107.08430*, 2021.

[13] Z. Gan, Y. Bai, P. Wu, B. Xiong, N. Zeng, F. Zou, J. Li, F. Guo and D. He, SGRN: SEMG-based gesture recognition network with multi-dimensional feature extraction and multi-branch information fusion, *Expert Systems with Applications*, vol. 259, art. no. 125302, Jan. 2025.

[14] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei and Y. Zhang, Personalized cross-silo federated learning on non-iid data, In: *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, Online, Feb. 2021, pp. 7865-7873.

[15] X. Jia, C. Zhu, M. Li, W. Tang and W. Zhou, LLVIP: A visible-infrared paired dataset for low-light vision, In: *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, Online, Jun. 2021, pp. 3496-3504.

[16] F. Jin, L. Ma, C. Zhao and Q. Liu, State estimation in networked control systems with a real-time transport protocol, *Systems Science & Control Engineering*, vol. 12, no. 1, art. no. 2347885, 2024.

[17] B. Li and W. Li, Distillation-based user selection for heterogeneous federated learning, *International Journal of Network Dynamics and Intelligence*, vol. 3, no. 2, art. no. 100007, Jun. 2024.

[18] R. Lan, Y. Zhang, L. Xie, Z. Wu and Y. Liu, BEV feature exchange pyramid networks-based 3D object detection in small and distant situations: A decentralized federated learning framework, *Neurocomputing*, vol. 583, art. no. 127476, 2024.

[19] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, In: *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, Online, Jun. 2021, pp. 10012-10022.

[20] H. Li, Z. Cai, J. Wang, J. Tang, W. Ding, C. Lin and Y. Shi, Fedtp: Federated learning by transformer personalization, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, pp. 13426-13440, 2023.

[21] Z. Li, G. Long, J. Jiang and C. Zhang, Dynamic fusion strategies for federated multimodal recommendations, *arXiv preprint arXiv:2410.08478*, 2024.

[22] Y. Lu, S. Huang, Y. Yang, S. Sirejiding, Y. Ding and H. Lu, FedHCA2: Towards hetero-client federated multi-task learning, In: *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, Seattle, USA, Jun. 2024, pp. 5599-5609.

[23] Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Hou and M. Tegmark, Kan: Kolmogorov-arnold networks, *arXiv preprint arXiv:2404.19756*, 2024.

[24] P. Liang, T. Liu, L. Ziyin, N. Allen, R. Auerbach, D. Brent, R. Salakhutdinov and L. Morency, Think locally, act globally: Federated learning with local and global representations, *arXiv preprint arXiv:2001.01523*, 2020.

[25] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar and V. Smith, Federated optimization in heterogeneous networks, In: *Proceedings of the 3th Machine Learning and Systems (MLSys)*, Austin, USA, Mar. 2020, vol. 2, pp. 429-450.

[26] T. Li, S. Hu, A. Beirami and V. Smith, Ditto: Fair and robust federated learning through personalization, In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Online, Jul. 2021, pp. 6357-6368.

[27] B. McMahan, E. Moore and D. Ramage, Communication-efficient learning of deep networks from decentralized data, In: *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Lauderdale, USA, Apr. 2017, pp. 1273-1282.

[28] O. Mees, A. Eitel and W. Burgard, Choosing smartly: Adaptive multimodal fusion for object detection in changing environments, In: *Proceedings of the 29th IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Daejeon, Korea, Oct. 2016, pp. 151-156.

[29] G. Ma, Z. Wang, Z. Yang, R. Chen, W. Liu, Y. Zhang, and S. Yan, A novel pairwise domain-adaptation-assisted dual-task learning approach to copreduction of robotic machining efficiency and quality in new parameter spaces, *IEEE Transactions on Industrial Informatics*, vol. 21, no. 7, pp. 5150-5159, 2025.

[30] B. Qu, D. Peng, Y. Shen, L. Zou and B. Shen, A survey on recent advances on dynamic state estimation for power systems, *International Journal of Systems Science*, vol. 55, no. 16, pp. 3305–3321, 2024.

[31] S. Ren, K. He, R. Girshick and J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2016.

[32] J. Shen, Y. Chen, Y. Liu, X. Zuo, H. Fan and W. Yang, ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection, *Pattern Recognition*, vol. 145, art. no. 109913, 2024.

[33] L. Shao, Y. Zhang, X. Zheng, R. Yang and W. Zhou, SOH estimation of lithium-ion batteries subject to partly missing data: A Kolmogorov-Arnold-Linformer model, *Neurocomputing*, vol. 638, art. no. 130181, Jul. 2025.

[34] B. Song, J. Chen, W. Liu, J. Fang, Y. Xue and X. Liu, YOLO-ELWNet: A lightweight object detection network, *Neurocomputing*, vol. 636, art. no. 129904, Jul. 2025.

[35] G. Sun, M. Mendieta, A. Dutta, X. Li, and C. Chen, Towards multimodal transformers in federated learning, In: *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, Milan, Italy, Sept. 2024, pp. 229-246.

[36] G. Sun, M. Mendieta, J. Luo, S. Wu and C. Chen, Fedperfix: Towards partial model personalization of vision transformers in federated learning, In: *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 4988-4998.

[37] R. Varghese and M. Sambath, YOLOv8: A novel object detection algorithm with enhanced performance and robustness, In: *Proceedings of the International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, Chennai, India, Apr. 2024, pp. 1-6.

[38] W. Wang, L. Ma, Q. Rui and C. Gao, A survey on privacy-preserving control and filtering of networked control systems, *International Journal of Systems Science*, vol. 55, no. 11, pp. 2269–2288, 2024.

[39] Y. Wang, C. Wen and X. Wu, Fault detection and isolation of floating wind turbine pitch system based on Kalman filter and multi-attention 1DCNN, *Systems Science & Control Engineering*, vol. 12, no. 1, art. no. 2362169, 2024.

[40] H. Wu, Q. Wang, X. Luo and Z. Wang, Learning accurate representation to nonstandard tensors via a mode-aware tucker network, *IEEE Transactions on Knowledge and Data Engineering*, in press, DOI: 10.1109/TKDE.2025.3617894.

[41] Y. Yuan, J. Liu and D. Jin, DeceFL: A principled decentralized federated learning framework, *arXiv preprint arXiv:2107.07171*, 2021.

[42] T. Yao, Y. Li, Y. Pan, Y. Wang, X. Zhang and T. Mei, Dual vision transformer, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10870-10882, 2023.

[43] X. Yang and X. Wang, Kolmogorov-arnold transformer, *arXiv preprint arXiv:2409.10594*, 2024.

[44] Y. Zhang, W. Zeng, S. Jin, C. Qian, P. Luo and W. Liu, When pedestrian detection meets multi-modal learning: Generalist model and benchmark dataset, In: *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, Milan, Italy, Sept. 2024, pp. 430-448.

[45] T. Zhao, M. Yuan, F. Jiang, N. Wang and X. Wei, Removal and selection: Improving RGB-infrared object detection via coarse-to-fine fusion, *arXiv preprint arXiv:2401.10731*, 2024.

[46] Z. Zong, G. Song and Y. Liu, Detrs with collaborative hybrid assignments training, In: *Proceedings of the 19th IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, Oct. 2023, pp. 6748-6758.

[47] Y. Zhang, R. Lan, X. Li, J. Fang, Z. Ping, W. Liu and Z. Wang, Class imbalance wafer defect pattern recognition based on shared-database decentralized federated learning framework, *IEEE Transactions on Instrumentation and Measurement*, vol. 73, pp. 1-17, 2024.