processes

MDPI

Article

Personalized FedM^{2}former: An Innovative Approach Towards Federated Multi-Modal 3D Object Detection for Autonomous Driving

Liang Zhao [1,2,*], Xuan Li [1], Xin Jia [1] and Lulu Fu [1]

[1] College of Electrical Engineering, Henan University of Technology, Zhengzhou 450001, China; l19561344399@163.com (X.L.); jx15090349719@163.com (X.J.); fll9681@163.com (L.F.)
[2] Henan Key Laboratory of Grain Storage Information Intelligent Perception and Decision Making, Henan University of Technology, Zhengzhou 450001, China
* Correspondence: zhaoliang_270@163.com

Abstract: With the swift evolution of artificial intelligence in the automotive sector, autonomous driving has ascended as a pivotal research frontier for automotive manufacturers. Environmental perception, as the cornerstone of autonomous driving, necessitates innovative solutions to address the intricate challenges posed by data sensitivity during vehicle operations. To this end, federated learning (FL) emerges as a promising paradigm, offering a balance between data privacy preservation and performance optimization for perception tasks. In this paper, we pioneer the integration of FL into 3D object detection, presenting personalized FedM^{2}former, a novel multi-modal framework tailored for autonomous driving. This framework aims to elevate the accuracy and robustness of 3D object detection while mitigating concerns over data sensitivity. Recognizing the heterogeneity inherent in user data, we introduce a personalization strategy leveraging stochastic gradient descent optimization prior to local training, ensuring the global model's adaptability and generalization across diverse user vehicles. Furthermore, to address the sparsity of point cloud data, we innovate the attention layer within our detection model. Our balanced window attention mechanism innovatively processes both point cloud and image data in parallel within each window, significantly enhancing model efficiency and performance. Extensive experiments on benchmark datasets, including nuScenes, ONCE, and Waymo, demonstrate the efficacy of our approach. Notably, we achieve state-of-the-art results with test mAP and NDS of 71.2% and 73.6% on nuScenes, 67.14% test mAP on ONCE, and 83.9% test mAP and 81.8% test mAPH on Waymo, respectively. These outcomes underscore the feasibility of our method in enhancing object detection performance and speed while safeguarding privacy and data security, positioning Personalized FedM^{2}former as a significant advancement in the autonomous driving landscape.

Keywords: personalized FedM^{2}former; personalized strategy; balanced window attention; 3D object detection; privacy protection

check for updates

Academic Editor: Wanquan Liu

Received: 20 January 2025
Revised: 1 February 2025
Accepted: 6 February 2025
Published: 7 February 2025

Citation: Zhao, L.; Li, X.; Jia, X.; Fu, L. Personalized FedM^{2}former: An Innovative Approach Towards Federated Multi-Modal 3D Object Detection for Autonomous Driving. *Processes* **2025**, *13*, 449. https://doi.org/10.3390/pr13020449

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The autonomous driving technology of intelligent vehicles is advancing at an unprecedented pace, showcasing immense potential in mitigating traffic congestion, enhancing road safety, and optimizing traffic flow efficiency. Perception technology, a pivotal component of autonomous driving systems, primarily hinges on various onboard sensors, including cameras, radars, and Lidar, to identify and comprehend the positions and categories of objects in the surrounding environment [1,2]. Fundamental perception tasks

encompass object classification [3], 2D and 3D semantic segmentation [4], 2D and 3D object detection [5], as well as depth completion. The execution of these tasks notably enhances the autonomous driving system's capability to recognize traffic obstacles and participants, thereby augmenting the system's accuracy, safety, and intelligence level. Notably, in the core task of object detection, the system must accurately identify and track diverse targets in its vicinity, such as other vehicles, pedestrians, and bicycles. This capability is vital for ensuring road safety and enabling the system to excel in various complex traffic scenarios [6]. As real-world driving scenarios continue to evolve over time, autonomous vehicles must continually learn from a broad spectrum of driving situations, especially when confronted with increasingly extreme, emergency, and congested conditions, rendering this capability of utmost importance.

With the advancement of the Internet of Vehicles (IoV), the volume of data generated during vehicle operation has been continuously increasing. While the abundance of data presents substantial opportunities for addressing various tasks, a significant portion of this data remains decentralized and is highly sensitive. As a result, protecting user privacy and ensuring the security of driving data have emerged as new challenges in the realm of autonomous driving. Federated learning (FL) offers a decentralized machine learning approach that enables model training without requiring centralized data storage [7,8]. This approach allows each autonomous vehicle to utilize its local data for training object detection models, meaning that each vehicle trains models using its own sensor data without the need to upload raw data to a central server [9]. Subsequently, by leveraging FL technology, vehicles aggregate the parameters of their locally trained models to create a global model. This global model integrates the knowledge from all vehicles, thereby enhancing the performance and generalization capability of object detection [10]. Therefore, the application of FL to perception tasks in autonomous driving can effectively address data privacy concerns associated with autonomous vehicles. Simultaneously, it enables perception models to learn from a wide range of scenarios, further improving model performance and enabling autonomous systems to adapt to diverse road environments and conditions.

In the realm of autonomous driving, multimodal 3D object detection leverages the synergistic effects of multiple sensors, each offering unique advantages. Multimodal 3D object detection significantly outperforms single-modal methods by integrating data from various sensors and harnessing their complementary strengths in addressing environmental factors such as light, magnetic fields, and temperature [11,12]. Nevertheless, the synchronization and calibration of these data in both temporal and spatial dimensions pose a significant challenge, making the selection of sensors crucial. The prevalent fusion paradigm involves the coupling of cameras and Lidar. Cameras furnish high-resolution images rich in semantic content, while Lidar provides precise depth information, including the distance, height, and shape of objects [13]. This fusion deepens the autonomous system's comprehension of its surroundings [14]. However, the sparsity of Lidar data in three-dimensional space results in inefficient processing by models, which affects the accuracy of 3D object detection tasks. Consequently, it is particularly important for models to process point cloud data efficiently and accurately. The differing representations of raw data pose challenges to the effective fusion of images and point clouds, with the pursuit of efficient and straightforward fusion methods being a primary direction of exploration. Multi-source information fusion methods can be categorized into three primary strategies: point-based, proposal-based, and Bird's Eye View (BEV)-based fusion. Point-based methods [15] augment Lidar points with semantic features, while proposal-based methods enrich 2D image-derived object proposals. BEV-based fusion [16] stands out by unifying camera and Lidar representations in a common BEV space, facilitating subsequent 2D convolutions. This approach

offers a global perspective, simplifies tasks, mitigates occlusions, ensures spatial consistency, and efficiently processes heterogeneous data through a multi-modal encoder. It also enhances representation learning, minimizes inference latency, and bolsters 3D scene understanding [17].

In this paper, we introduce personalized FedM^{2}former, an innovative framework tailored for multimodal 3D object detection within the FL paradigm. This framework incorporates a centralized server that initiates the learning process by disseminating an initialized model to participating connected autonomous vehicles (CAVs). These CAVs leverage their locally collected data to perform individualized training and subsequently share their updated model parameters back with the server. The server aggregates these parameters from all CAVs, refining the global model, which is then redistributed. This process fosters generalization enhanced by user-specific data characteristics, thereby realizing personalization within the FL context. Furthermore, we devise a novel point cloud processing technique to bolster the performance of the local models. Critically, data privacy is upheld through the exclusive exchange of models and parameters between vehicles and the server, mitigating communication overheads and safeguarding the privacy of vehicle data.

The salient contributions of this paper are enumerated as follows:

- We present the personalized FedM^{2}former, a cutting-edge multimodal 3D object detection framework tailored exclusively for CAVs. This framework leverages FL for decentralized training across multiple vehicles, effectively addressing data sensitivity concerns and mitigating communication challenges posed by transmitting voluminous data. By doing so, it enhances the performance of 3D object detection tasks, underscoring its potential to revolutionize CAV perception systems.
- Leveraging the complementary strengths of local and global knowledge, we introduce an innovative adaptation mechanism for stochastic gradient descent (SGD) as a personalization strategy within the FL framework. After updating the global model through federated averaging, a personalized SGD algorithm is seamlessly integrated. This approach enables individual users to fine-tune the global model according to their unique loss functions, ultimately optimizing for local data distributions.
- To enhance the efficiency of processing 3D point cloud data, we introduce a balanced window attention (BWA) mechanism. Building upon the advancements in window attention, BWA optimizes the attention layer's capability to concurrently and effectively handle both 3D point cloud and image data. This innovation fosters a parallel processing paradigm, thereby accelerating overall computational throughput and facilitating more seamless integration of diverse data modalities.

The subsequent sections of this paper are organized as follows: Section 2 comprehensively reviews the relevant literature, establishing a solid foundation for our research. Section 3 elucidates the foundational theories underlying the related technologies. Section 4 introduces a comprehensive framework for our novel methodology, elucidating its essential components and innovative aspects. Section 5 details the experimental design, implementation, and rigorous analysis of results. Lastly, Section 6 offers profound insights derived from the study and envisions promising avenues for future research endeavors in this field.

2. Related Work

In our endeavor, we focus our investigation on 3D object detection within the FL framework for autonomous driving, pursuing two pivotal research trajectories: delving into the core aspects of 3D object detection research, and elucidating the personalized paradigm of FL. The following sections elaborate on each of these aspects.

2.1. Three-Dimensional Object Detection

Three-dimensional object detection stands as a cornerstone research endeavor within the realm of autonomous driving perception tasks, aimed at precisely localizing and identifying objects within a three-dimensional space [18]. Advancing beyond traditional 2D object detection, 3D detection not only conveys the positional and categorical information of objects within images but also imparts precise details pertaining to their spatial location, dimensions, and orientation in the three-dimensional environment. This comprehensive approach significantly enhances the accuracy and completeness of object perception for autonomous systems, fostering safer and more efficient decision-making capabilities [19]. Mathematically, this can be generally formulated as follows:

$$\alpha(a_1, a_2, \cdots, a_n) = f(\beta) \tag{1}$$

where $(a_1, a_2, \cdots, a_n)$ represents N three-dimensional objects, typically containing information such as position, size, category, etc., and potentially additional information based on the model set. $\alpha(a_1, a_2, \cdots, a_n)$ denotes the set of N three-dimensional object states in each frame scene, where $f(\cdot)$ is the three-dimensional detection function, and β represents the input data, usually from sensors.

2.1.1. Single-Modal 3D Object Detection

With the advancement of deep learning technologies, revolutionary changes have taken place in 3D object detection for autonomous driving. As early as 2016, Chen X et al. proposed a camera-based method [20] that utilizes advanced computer vision and deep learning techniques to analyze image data, excelling in inferring the attributes of three-dimensional objects. In 2017, Chen X et al. further introduced MV3D [21], a neural network architecture specifically designed for detecting and locating 3D objects from images captured from multiple viewpoints. MV3D enhances the understanding of 3D scenes by leveraging images from multiple viewpoints. By simultaneously processing and fusing information from these viewpoints, it improves the accuracy and stability of object detection. However, relying solely on cameras for object detection fails to capture depth information, inherently limiting depth precision and thus affecting detection accuracy due to their reliance on image data. In 2017, Charles R. Qi et al. proposed PointNet [22], a deep learning architecture based on Lidar, which is designed to directly process point cloud data without the need for prior conversion into other forms. This enables it to effectively handle irregular and uneven point cloud data. Subsequently, Charles R. Qi et al. introduced PointNet++ [23], a further development and refinement of PointNet, which incorporates a hierarchical structure. This structure allows the network to learn and represent data at various levels of abstraction, thereby enhancing its ability to understand complex structures in point clouds. In 2018, Zhou, Y et al. introduced VoxelNet [24], which was the first to directly convert 3D point cloud data into voxel grids, achieving effective modeling of 3D space through the combination of 3D convolutional networks and dilated convolutions. The Lidar-centric approach overcomes the issue of depth ambiguity by utilizing dense point clouds, but it has limitations such as low spatial resolution, sparse data, and sensitivity to texture variations, which hinder the enhancement of overall performance. As a result, single-modal methods inherently face difficulties in achieving optimal levels of 3D scene understanding and require further development.

2.1.2. Multi-Modal 3D Object Detection

Multimodal 3D object detection primarily leverages cross-modal data to enhance detection accuracy. Compared to single-modal methods, it can exploit the advantages of multiple modalities, such as depth information provided by point clouds and texture

information provided by images. In 2018, Xu D et al. proposed a multimodal object detection framework called PointFusion [25], which effectively fuses image and 3D point cloud information through deep learning methods to achieve more precise 3D object detection. In this fusion network, the input 3D points are used as spatial anchors to predict multiple 3D bounding box hypotheses and their corresponding confidence scores, yielding significant results. In 2020, Vora S et al. proposed PointPainting [15], an end-to-end trainable 3D object detection technique that projects Lidar points onto a semantic segmentation network of images and attaches a score for each category to each point, thereby enhancing the accuracy of 3D object detection by combining the complementary information from images and point clouds. In 2022, Liang T et al. introduced BEVfusion [16], a novel 3D object detection framework whose core lies in utilizing BEV representations to depict and comprehend the environment. A BEV image, which is a top-down view of the scene, offers more intuitive and accurate information regarding spatial layout and object positions, thereby achieving high detection accuracy. BEVfusion converts data from various sensors into BEV representations, allowing all modal features to be expressed in the form of 2D grids, which can then be directly fused through convolutional operations, demonstrating robust feature extraction capabilities. However, current multimodal 3D object detection technologies also have some drawbacks, such as significant information loss, relatively high computational complexity, expensive pre-processing costs, and complex fusion strategy designs. Therefore, the main focus of our research is to develop an object detection method that is simple, efficient, and highly accurate.

2.2. Personalized Strategies

Given the prevalent variations in user local data distributions, traditional FL methods encounter substantial challenges in ensuring convergence and providing personalized solutions tailored to the inherent heterogeneity of data. These obstacles can adversely impact the deployment performance of the global model on individual clients. Therefore, research focused on personalized FL is of crucial importance, as this technology enables personalized adjustments based on users' unique data and requirements, thereby significantly enhancing the adaptability of the global model. In 2020, Mansour, Y., et al. proposed three personalization approaches in their paper titled "Three Approaches for Personalization with Applications to Federated Learning [26]". The user clustering method involves grouping clients with similar data distributions or characteristics and then training a model for each cluster. The data interpolation method generates personalized models by linearly combining the global model with local datasets. The model interpolation method, on the other hand, creates personalized models through the linear combination of the global model and local models. While these three methods enhance model performance, they all face the limitation of restricted personalization degrees. In 2022, Tan, A. et al., in their paper titled "Towards Personalized Federated Learning [27]", proposed two primary strategies for personalizing FL: learning personalized models and personalizing the global model. The former strategy abandons the pursuit of training a single model and instead trains a unique personalized model for each local client. This approach offers greater freedom for each client to customize the model according to its specific needs and data distribution. However, managing and maintaining these decentralized personalized models become more complex and challenging, increasing the overhead of storage and computational resources. The latter strategy involves personalizing the local clients through local adaptation steps based on a globally trained model. This method aims to retain the generalization ability of the global model while improving its performance on specific clients through local adaptation steps. As such, global model personalization emerges as a key enabler, propelling the progression of FL technology towards more customized and efficient learning paradigms.

3. Preliminaries

3.1. Federated Learning in Autonomous Driving

FL presents a collaborative strategy for distributed machine learning, fostering model training collaboration among multiple participants [28]. In the autonomous driving domain, vehicles individually train local models based on their proprietary data, but the scarcity of effective data limits their model efficacy. By participating in FL, CAVs share their training parameters with a central server, enabling the server to aggregate these parameters and iteratively improve the shared global model.

During the local training phase, each participant i trains the model on its local dataset D_i and obtains locally updated model parameters θ_i, with the specific calculation detailed as follows:

$$\theta_i = argmin f_i(\theta, D_i) \tag{2}$$

where $f_i(\theta, D_i)$ represents the loss function for participant i on its dataset. During the model aggregation phase, the central server collects the local model parameters θ_i from all participants and computes the global model parameters θ_{global} using the aggregation algorithm FedAvg. The computation steps are detailed as follows:

$$\theta_{global} = \sum_{i=1}^{N} \frac{n_i}{n} \theta_i \tag{3}$$

where N denotes the number of clients, n_i signifies the sample size of the i-th client, and n denotes the total sample size across all clients.

This methodology empowers CAVs with the capacity to iteratively refine their localized models via collaborative learning, ultimately converging towards a globally robust and precise model. Critically, this process is executed sans direct access to vehicle-specific data, thereby effectively addressing concerns pertaining to data sensitivity and privacy, as exemplified in the framework depiction in Figure 1.

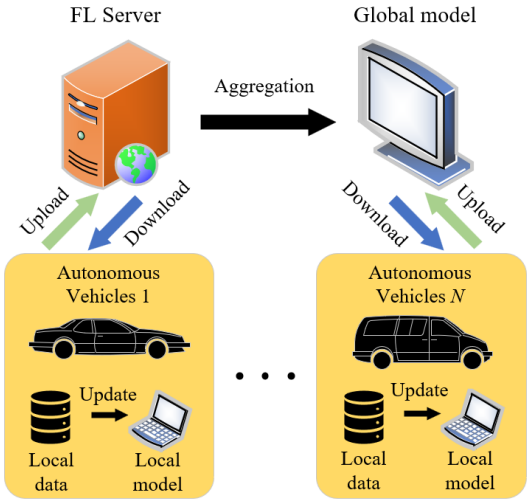


Figure 1. Federated learning framework.

3.2. Federated Transformer

The Transformer, a neural network architecture leveraging the self-attention mechanism, was initially devised for sequence-to-sequence NLP tasks like machine translation and text generation. Its efficacy has transcended these origins, notably in computer vision realms, particularly object detection, showcasing remarkable performance with large datasets. As a nascent training paradigm, FL holds promise to elevate the training of

Transformer models. The Transformer's inherent robustness against distribution shifts underscores its adaptability to diverse data landscapes, which FL can amplify in applications such as object detection. Thus, exploring and developing personalized FL approaches tailored to the Transformer's unique traits constitutes a pivotal research direction. These endeavors aim to mitigate data imbalance and privacy concerns, fostering broader, more efficient model training and deployment.

4. Methodology

This research introduces personalized FedM^{2}former, an innovative framework designed for distributed learning and inference of 3D object detection in autonomous driving. Leveraging distributed Transformer-based model training across vehicles and edge clouds in realistic communication environments, FedM^{2}former enables efficient collaborative learning, accelerating the attainment of high-precision perception and significantly enhancing the safety and reliability of autonomous driving systems.

4.1. Personalized FedM^{2}former

Figure 2 depicts the proposed framework, wherein autonomous vehicles endowed with private data resources serve as clients. Due to the inherent volatility of IoV, our framework aims to alleviate the deterioration of model performance during FL processes by adopting a strategy of selective client participation. These vehicles, functioning as clients, harness their robust network connections to contribute their locally trained parameters, thereby ensuring a consistent and reliable aggregation process across communication rounds. To alleviate communication overhead and enhance communication speed, we employ model compression techniques between the server and clients. This approach not only reduces the transmission burden but also accelerates the overall learning process. Additionally, to enhance the generalization capability of the global model and improve its performance across various use cases, we optimize the FL framework by personalizing the global model. This optimization strategy tailors the global model to each client's unique data landscape, thereby enhancing the accuracy and robustness of 3D object detection within autonomous driving systems.

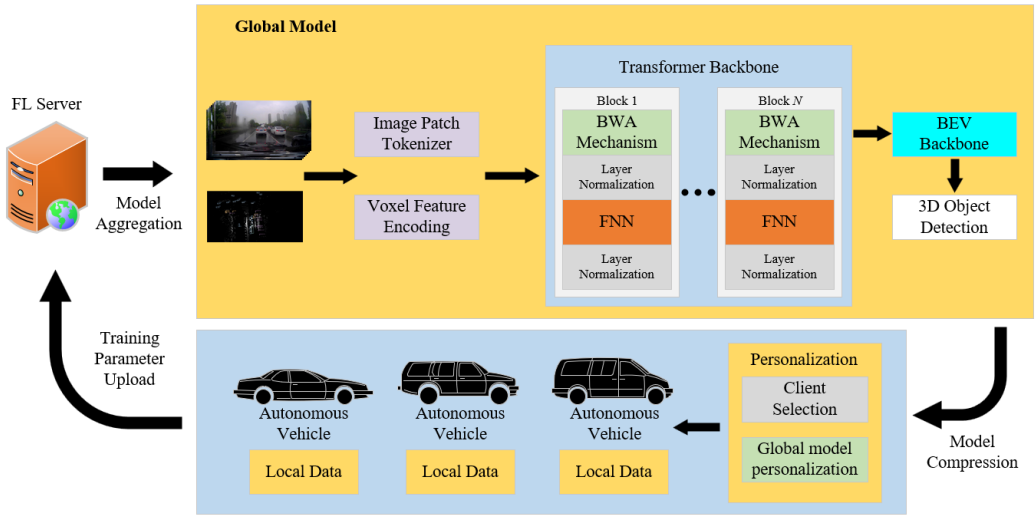


Figure 2. Overall architecture of the personalized FedM^{2}former.

During local training, our model processes image and point cloud data through specific modal tokenizers. For images, we employ an advanced image patch tokenizer, segmenting them into 8×8 pixel patches and converting them into serialized tokens. For point clouds, we utilize a voxel feature encoding tokenizer with a customized grid size of (0.3 m, 0.3 m,

8 m) to achieve effective representation. The processed data then enter the Transformer backbone network, where the attention layers of the encoder can simultaneously handle both modalities, facilitating multimodal feature learning. Notably, we designed a BWA mechanism specifically to enhance the processing of point cloud sequences, efficiently managing sparse point cloud data and promoting 3D feature learning. Finally, these enriched features are passed through a BEV backbone network to facilitate precise 3D object detection.

Upon completion of local training, clients are mandated to share only their training parameters, which are subsequently uploaded to the server for weighted federated averaging to aggregate into an updated global model. This updated global model is then redistributed back to user vehicles. Notably, prior to each local training iteration, users will apply personalized optimization strategies to the global model using stochastic gradient descent (SGD) based on their individual training scenarios. This approach markedly enhances the model's adaptability to local data distributions, thereby improving the generalization capability and overall performance of the global model.

4.2. Personalized Federated Averaging

The conventional federated averaging method involves each user training their local model using their respective local dataset. The local model parameters on user i are denoted as θ_i, where i is the index of the user. After completing local training, users send their local model parameters to a central server. Upon receiving the local model parameters from all users, the central server computes their weighted average to update the global model parameters.

$$\theta_{\text{global}} = \sum_{i=1}^{N} w_i \theta_i \tag{4}$$

where θ_{global} represents the global model parameters, N denotes the number of users, and w_i signifies the weight of device i, typically determined based on factors such as the performance of user devices and the proportion of data they contribute to the overall dataset. In local training, the minimization objective function is usually defined as follows:

$$\min_{\theta \in \mathbb{R}^n} F(\theta) = \frac{1}{N} \sum_{i=1}^{N} f_i(\theta) \tag{5}$$

where $F(\theta)$ denotes the aggregate loss function and $f_i(\theta)$ represents the i-th sub-function.

To ensure the efficiency of FL, this subsection expounds a personalized federated averaging method that incorporates the SGD optimization algorithm into FL, aiming to enhance the performance of the global model and its generalization ability across different users. Specifically, at the beginning of each local training iteration for user vehicles, the global model is initially updated using gradient descent steps based on their respective locally trained loss functions. Thus, the objective solving process is transformed into

$$\min_{\theta \in \mathbb{R}^n} F(\theta) = \frac{1}{N} \sum_{i=1}^{N} f_i(\theta - \alpha \nabla f_i(\theta)) \tag{6}$$

where $\nabla f_i(\theta)$ denotes the gradient of the client's loss function at point θ, and $\alpha \geq 0$ represents the step size. The computation of the gradient of the client's loss function is as follows:

$$\nabla f_i(\theta) = \left(I - \alpha \nabla^2 f_i(\theta)\right) \nabla f_i(\theta - \alpha f_i(\theta)) \tag{7}$$

where I is the identity matrix, $\nabla^2 f_i(\theta)$ represents the Hessian matrix (second-derivative matrix) of the loss function $f_i(\theta)$ at point θ. $\nabla f_i(\theta - \alpha f_i(\theta))$ denotes the gradient of the loss

function $f_i(\theta))$ at the point$(\theta - \alpha f_i(\theta))$. Considering the large amount of data required to compute gradients, we obtain an unbiased estimate $\widetilde{\nabla} f_i(\theta, D^i)$ from the relative sample distribution P_i of the data batch D^i. Similarly, an unbiased estimate of $\nabla^2 f_i(\theta)$ is denoted as $\widetilde{\nabla}^2 f_i(\theta, D^i)$.

$$\widetilde{\nabla} f_i\left(\theta, D^i\right) = \frac{1}{\left|D^i\right|} \sum_{(x,y)\in D^i} \nabla P_i(\theta, (x, y)) \tag{8}$$

In this approach, the server initially transmits the current global model θ_k to a randomly selected subset of users, denoted as $\mathbb{C}_k$, in the k-th round. Each user C then executes σ steps of gradient descent locally, taking into account their individual loss function f_i. This local update process generates a sequence of local models, denoted as $\left[\theta^i_{k+1,t}\right]^{\sigma}_{t=0}$, where $\theta^i_{k+1,0} = \theta_k$. For $\sigma \geq t \geq 1$, we have

$$\theta^i_{k+1,t} = \theta^i_{k+1,t-1} - \beta \widetilde{\nabla} f_i\left(\theta^i_{k+1,t-1}\right) \tag{9}$$

where β denotes the local learning rate, $\widetilde{\nabla} f_i\left(\theta^i_{k+1,t-1}\right)$ represents an unbiased estimate of $\nabla f_i\left(\theta^i_{k+1,t-1}\right)$. The stochastic gradient for local iterations, denoted as $\widetilde{\nabla} F_i\left(\theta^i_{k+1,t-1}\right)$, is independently computed from D^i_t, D'^i_t, D''^i_t, that is

$$\begin{aligned} \widetilde{\nabla} F_i\left(\theta^i_{k+1,t-1}\right) = \Big\{ &\left(I - \alpha \widetilde{\nabla}^2 f_i\left(\theta^i_{k+1,t-1}, D''^i_t\right)\right) \\ &\widetilde{\nabla} f_i\left(\theta^i_{k+1,t-1} - \alpha \widetilde{\nabla} f_i\left(\theta^i_{k+1,t-1}, D^i_t\right), D'^i_t\right)\Big\} \end{aligned} \tag{10}$$

where $\widetilde{\nabla} f_i\left(\theta^i_{k+1,t-1} - \alpha \widetilde{\nabla} f_i(\theta^i_{k+1,t-1}, D^i_t), D'^i_t\right)$ involving one random gradient within another, $\widetilde{\nabla} F_i\left(\theta^i_{k+1,t-1}\right)$ serves as a biased estimate of $\nabla F_i\left(\theta^i_{k+1,t-1}\right)$. Finally, users transmit the locally computed updates to the server, which subsequently aggregates and averages the updates received from all users, as depicted in Equation (4).

In essence, the proposed methodology seamlessly fuses the robust features of FL with the ability to encapsulate individual user variability, showcasing an elegant integration. In the FL paradigm, each participant initially adopts the global model as its baseline and subsequently fine-tunes it using their unique local training parameters to obtain a personalized version suited to their needs. This strategic approach not only fortifies the global model's generalizability across a heterogeneous user base but also expedites the convergence rate during local training iterations. Consequently, it leads to a comprehensive enhancement in both the efficiency and effectiveness of the FL framework. The precise algorithmic nuances are outlined in Algorithm 1.

4.3. Transformer with Balanced Window Attention Mechanism

In this study, we leverage the Transformer model for the purpose of 3D object detection tasks, capitalizing on its unique design that is fundamentally rooted in the self-attention mechanism. This paradigm shift enables the model to proficiently discern and encode the intrinsic dependencies within the input multimodal data, devoid of any predispositions stemming from its own inherent characteristics. Consequently, this approach augments the efficiency of both the training and inference phases, fostering a more streamlined and effective process for 3D object detection.

The characteristic of the Transformer lies in the stacking of self-attention mechanisms and point-wise fully connected layers within its framework. It comprises N replicated blocks, each structured with a multi-head self-attention layer, succeeded by a position-wise fully connected feedforward network. Both layers are enhanced with residual connections and layer normalization, as specifically illustrated in Figure 3a. As the Transformer architec-

ture gains prominence in object detection, this study applies it to environmental perception tasks. Traditional Transformers struggle with Lidar point clouds due to sparsity and redundant voxels in 3D space. To overcome this, refs. [29,30] introduced window-based attention, reducing complexity while capturing long-range dependencies. However, these methods still exhibit redundant computations and performance instability. This work innovates by proposing the BWA mechanism, facilitating efficient parallel voxel processing. This not only boosts computational efficiency and inference speed but also enhances detection accuracy and model adaptability to diverse data distributions. The specific implementation process is illustrated in Figure 3b.

Algorithm 1 Personalized FedM^{2}former for 3D Object Detection

input: Number of selected users C, FL learning rate η, SGD learning rate β, global communication rounds T, local iteration numbers t
output: Global Model M_{global}
Initialize the global model M_o
for Each global communication round $T = 1, 2, 3, \ldots$ **do**
  Randomly select a federated participant C_i
  for Each client C_i in C **do**
    Receive the global model M_o from the central server
    $M_o \longrightarrow \theta_o$
    Update global model parameters according to the loss function:
    SGD $(\theta_o, \beta) \longrightarrow \theta'_o$
    Local training t epochs to update model parameters:
    Local training $(\theta'_o, \eta) \longrightarrow \theta_i$
    Send θ_i back to the central server
  end for
  Central server uses federated averaging to aggregate locally updated parameters θ_i from users.
  $(\theta_{\text{global}} = \sum_{i=1}^{C} w_i \theta_i) \longrightarrow M_{\text{global}}$
end for
return M_{global}

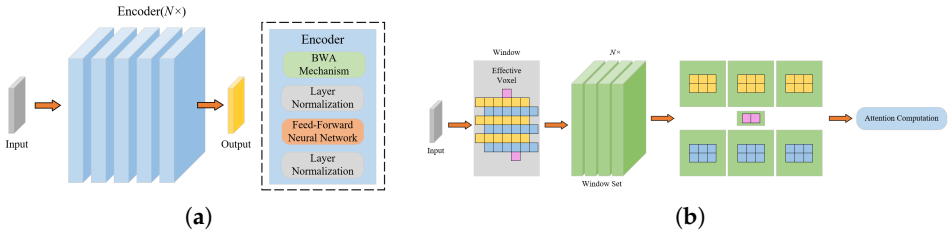


Figure 3. Transformer with BWA mechanism. (**a**) Encoder concrete structure. (**b**) Implementation process of the BWA mechanism.

To efficiently handle point cloud data, we encapsulate them within a series of well-defined, bounded windows and equal-sized sets, allowing for parallel execution of standard self-attention computations. Following traditional window attention practices, we transform the point cloud into voxels and partition them into non-overlapping 3D windows of a predetermined size. Each window encapsulates N valid voxels, i.e.,

$$Voxel = \{p_i | p_i = [(x_i, y_i, z_i); f_i; d_i]\}_{i=1}^{N} \tag{11}$$

where $(x_i, y_i, z_i) \in \mathbb{R}^3$ represent the voxel coordinates, $f_i \in \mathbb{R}^C$ denote the voxel features, and $d_i \in [1, N]$ is the index of each voxel within its corresponding window. The initial step towards generating non-overlapping and uniformly sized local sets involves determining the necessary number of such sets within the defined window, i.e.,

$$M = \left\lfloor \frac{N}{\mu} \right\rfloor + \xi[N\%\mu > 0] \tag{12}$$

where μ signifies the upper limit on the number of valid voxels assigned per set. The expression $\left\lfloor \frac{N}{\mu} \right\rfloor$ precisely calculates the maximum number of complete sets that can be formed from N voxels, each containing at most μ voxels. The binary function $\xi[N\%\mu > 0]$ identifies the need for an additional, possibly partial set when N is not divisible by μ, ensuring no voxel is left uncovered. This approach optimizes the partitioning of voxels into non-overlapping, uniform sets, minimizing resource consumption while guaranteeing comprehensive data coverage within the specified window.

To uniformly distribute N valid voxels among M sets, denote the voxel partition of the j-th set as $Q_j = \{q_k^j\}_{k=0}^{\mu-1}$, where $\mu = \left\lceil \frac{M}{N} \right\rceil$ ensures equal or nearly equal distribution among sets. The computation of voxel partition for the k-th set is then formulated as

$$q_k^j = \left\lfloor \frac{j \times \mu + k}{M \times \mu} \times N \right\rfloor \tag{13}$$

Following the partitioning of the set, the corresponding voxel features and spatial coordinates are extracted utilizing the indices of voxels residing within the specified window, i.e.,

$$\{T, Z\} = \Psi\left\langle Voxel, Q_{(1 \le i \le M)}, index \right\rangle \tag{14}$$

where $T \in \mathbb{R}^{\mu \times C}$ denotes the voxel features associated with a given set, while $Z \in \mathbb{R}^{\mu \times 3}$ represents the voxel coordinates corresponding to the set. $\Psi\langle\cdot\rangle$ signifies the retrieval procedure, which fetches the features and coordinates based on the voxel, set partitioning, and voxel indices.

Leveraging the proposed BWA mechanism, we are able to partition each window into multiple sets, each containing an equal quantity of valid voxels. Consequently, as the number of voxels within a given window increases, so does the number of corresponding sets generated, with each set maintaining a uniform size. This approach significantly enhances the capability for full parallel processing of these sets, thereby fostering an optimal allocation of computational resources. Ultimately, this leads to a marked improvement in both inference speed and detection accuracy.

4.4. Multi-Modal Information Fusion

To enhance the seamless integration of disparate sensor data in autonomous driving, conventional approaches typically require dedicated deep models for each sensor, unifying the data into a common framework, followed by intricate post-processing for fusion, often at the expense of valuable information fidelity. This study innovates by presenting a novel multimodal fusion paradigm that seamlessly bridges the Lidar and camera modalities. It integrates Lidar features into camera-centric space for preliminary feature learning and then strategically maps these enriched features back to Lidar space for deeper modality-specific exploration. The fusion process is elaborated as follows.

4.4.1. Feature Learning of Lidar in 2D Space

To project 3D point cloud data into a 2D camera space for feature learning, we employ the camera's intrinsic and extrinsic parameters to facilitate a one-to-one projection of Lidar tokens onto a 2D plane. This projection process aims to enable subsequent uniform

perspective partitioning. During this procedure, we select the first view of Lidar tokens and precisely place them at the corresponding locations within the 2D space, i.e.,

$$\Gamma^P \to \Gamma^P_{2D} : \left(x^P, y^P, z^P\right) \to \left(x^P_{2D}, y^P_{2D}, z^P_{2D}\right) \tag{15}$$

where Γ^P_{2D} represents the view index for each Lidar token.

After harmonizing the image and Lidar tokens into their respective 2D perspective spaces, we adopt the methodology articulated in Equation (12) to partition these entities into M cross-modal 2D local sets in a modality-agnostic fashion.

$$\left\{Q^{2D}_m\right\}^M_{m=0} = \left(\left[\Gamma^P_{2D}, \Gamma^I\right], L^I \times W^I \times 1, \Gamma\right) \tag{16}$$

where $\left[\Gamma^P_{2D}, \Gamma^I\right]$ denotes the concatenation operation. Employing a modality-agnostic approach, we systematically partition multimodal tokens into corresponding local sets, guided by their positions within the camera's perspective space. Subsequently, these sets are integrated into the Transformer framework to facilitate 2D cross-modal feature interaction and learning.

4.4.2. Feature Learning of Camera in 3D Space

To unify multimodal representations in 3D, efficient and robust projections are vital for uniquely mapping image patches to 3D space. The image-to-Lidar transformation faces an ill-posed challenge due to inaccurate pixel-based depth. While existing depth estimation methods offer reasonable accuracy, they rely on additional modules and have limited generalization. To address these limitations, we introduce a pre-computable mapping approach that efficiently transforms image patches into Lidar space.

Precisely, we commence by sampling a set of pseudo-Lidar points, denoted as Y^P, and then project them into the corresponding image virtual coordinates, denoted as $\mathrm{Y}^I = \{v_k|(x_k, y_k); b_k; d_k\}^{\mathrm{Y}}_{k=0}$, where b_k and d_k represent the view index and the associated depth information, respectively. Given that we confine our attention to projected points that reside within the view ϕ, the number of valid image points satisfies $\mathrm{Y} \leq |\mathrm{Y}^P| \times \phi$. These pseudo-image points Y^I are utilized to retrieve the depth estimation of each image token from its nearest image virtual neighborhood, i.e.,

$$Lin\left(\mathrm{Y}^I, \Gamma^I\right) = \left\{v^{Lin}_i|(x_i, y_i); b_i; d_i\right\}^M_{i=0} \tag{17}$$

where $Lin(\mathrm{Y}^I, \Gamma^I)$ designates the nearest neighbor function, whereby d_i signifies the depth estimate assigned to the i-th image token, assumed to coincide with that of its nearest neighboring token. Following this, we transform the image tokens Γ^I into the 3D space and compute offset features that are derived from their respective 2D distances, i.e.,

$$\begin{aligned} \Gamma^I \to \Gamma^I_{3D} : \left(x^I, y^I, b^I\right) &\to \left(x^I_{3D}, y^I_{3D}, z^I_{3D}\right) \\ F^I \to F^I + \left(\left\|v^{Lin}(x, y) - \left(x^I, y^I\right)\right\|\right) & \end{aligned} \tag{18}$$

According to Equation (18), image and Lidar tokens are consolidated in Lidar space via projection, subsequently partitioned into 3D cross-modal local sets, i.e.,

$$\left\{Q^{3D}_m\right\}^M_m = \left(\left[\Gamma^P, \Gamma^I_{3D}\right], L^P \times W^P \times H^P, \Gamma\right) \tag{19}$$

Integrating into the Transformer framework for comprehensive processing fosters 3D cross-modal feature interaction learning within the Lidar space. Our holistic approach

meticulously bridges the features from disparate modalities, taking into account the intricate structural correlations between 2D and 3D spaces, resulting in an effective fusion of image and Lidar data.

5. Experiments

In this study, we empirically demonstrate the efficacy of our proposed methodology through rigorous numerical experiments. We commence by outlining the experimental setup, meticulously specifying the hardware and software configurations to ensure reproducibility and rigor. Subsequently, we elaborate on the utilization of public datasets and the selection of evaluation metrics to comprehensively assess performance. Our experiments focus on validating four key aspects: (1) Establishing the practicality of our personalized FL framework for 3D object detection vis-à-vis centralized and standalone training; (2) Quantifying the influence of personalization strategies on training outcomes; (3) Benchmarking our BWA mechanism against other attention mechanisms for multimodal 3D object detection; (4) Investigating the effect of varying user counts on detection performance. Finally, we compare our method's detection results with state-of-the-art approaches using three benchmark datasets, underscoring its superior performance.

5.1. Experiment Setup

Given the substantial demands of this experiment for ample storage capacity and computational resources, the experimental training and testing phases were conducted on an advanced deep learning server, which was furnished with four GPUs. This configuration ensured peak performance and efficiency throughout the experimental process. To ensure the fairness of the experiments, all parameter settings were consistent across all experiments. Considering the magnitude of parameter updates and the need to balance convergence speed with stability, the learning rate η for FL was set to 0.01. For the SGD learning rate β, we will explore and determine its optimal value within the range of 0.1–0.5 in subsequent experiments. The specific settings are presented in Table 1 as follows:

Table 1. Specific experiment setup.

Hardware Setup	CPU	Xeon Platinum 8358P
	GPU	NVIDIA-A100-SXM4
	Video Memory	80 GB
	Internal Storage	120 GB
	Hard Disk	1.5 TB
Software Environment	Operating System	Ubuntu20.04
	Programming language	Python 3.8
	PyTorch	1.10.0
	CUDA	11.3

5.2. Datasets and Evaluation Metrics

In our research, we conducted extensive 3D object detection experiments utilizing the nuScenes, ONCE, and Waymo datasets. To mimic the inherent distributed characteristics of FL, we randomly partitioned each dataset into multiple subsets, where each subset represented the localized data residing on an individual edge device participating in the FL framework. Essentially, these subsets mirrored the local data collections of autonomous vehicles. This diversification of datasets not only encompasses a broader range of driving scenarios but also enhances the generalization capability of our proposed approach. Recognized as standard benchmarks within the autonomous driving perception domain, these three datasets comprehensively encapsulate a myriad of road conditions, diverse time periods, and varying weather conditions, thereby exhibiting remarkable representativeness.

In the subsequent sections, we delve into the specifics of each dataset and the performance evaluation metrics adopted for comprehensive analysis. The specific structure of the dataset is illustrated in Table 2.

Table 2. Specific structure of dataset.

Dataset	Scenes	Hours	Images	Point Clouds	Categories
NUSCENES	400 k	5.5	1.4 M	400 k	23
ONCE	1 M	144	7 M	1 M	5
WAYMO	230 k	6.4	1 M	230 k	4

NuScenes [31] represents a large-scale outdoor driving dataset that offers diverse annotations for a myriad of autonomous driving perception tasks. It comprises a total of 1000 fully annotated driving scenarios sourced from Boston and Singapore, with each scenario's data captured by a 32-beam LiDAR system and six cameras. The primary evaluation metrics for this dataset are mAP (mean Average Precision) and NDS (nuScenes Detection Score). mAP calculates the average AP for different object categories individually across varying IoU (Intersection over Union) thresholds, and then averages these AP values. NDS, on the other hand, is a comprehensive scoring system that integrates multiple True Positive metrics, encompassing not only mAP but also the detection quality based on location, size, orientation, attributes, and velocity measurements.

ONCE [32] represents an ultra-large-scale autonomous driving dataset with 3D object annotations. It was collected over 144 h of driving by vehicles equipped with a 40-beam Lidar on the roof and seven cameras. The dataset comprises a total of 15,000 fully annotated driving scenarios, encompassing diverse conditions such as day and night, sunny and rainy weather, as well as urban and suburban environments. The mAP is the primary evaluation metric for ONCE.

Waymo [33] is a dataset collected by autonomous vehicles under various conditions, utilizing one mid-range Lidar, four short-range Lidars, and five cameras. It encompasses a total of 1150 fully annotated scenarios. The evaluation metrics employed in Waymo are mAP/mAPH (L1, L2), where mAP remains the same as aforementioned, while mAPH is a variant of mAP that considers not only the location and size of the bounding boxes but also their orientation angles. Specifically, L1 represents lower detection difficulty, indicative of larger objects, minimal occlusion, or relatively simple scenarios, whereas L2 signifies higher detection difficulty, including smaller objects, heavy occlusion, complex scenes, or adverse lighting conditions.

5.3. Validation of Our Contributions

5.3.1. Performance Assessment of the Personalized Federated Learning Framework

Herein, we conduct a thorough evaluation of our proposed FL approach for 3D object detection in autonomous driving, benchmarking it against traditional centralized learning (CL) and single-vehicle learning (SVL). During the experiments, we simulated 30 user vehicles, thereby randomly and equally dividing the nuScenes dataset into 30 subsets to ensure heterogeneity and diversity in the training data among vehicles, reflecting real-world traffic scenarios. Here, CL refers to the method of aggregating all data for training, while SVL entails training solely based on the local data of a single vehicle. We report the mean values of 10 experimental runs for each of the three methods, as illustrated in Figure 4, where horizontal axis in the context of FL signifies the number of aggregation rounds, and in CL and SVL, it represents the number of iterations.

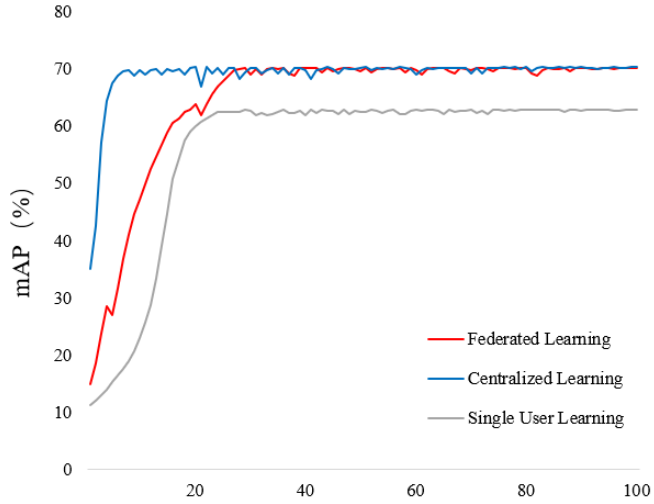
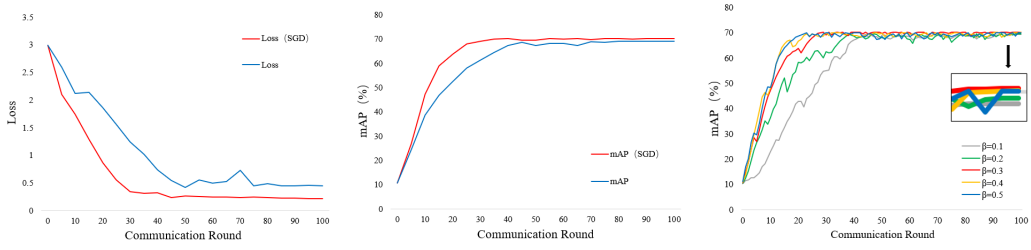


Figure 4. Comparison of training performance among centralized learning, federated learning, and single vehicle learning.

Upon observing Figure 4, it is evident that FL exhibits excellent performance in object detection. After 100 iterations, the mAP of FL closely resembles that of CL, with CL achieving 70.31% and FL achieving 70.23%. Although CL converges significantly faster than FL, both eventually achieve nearly identical accuracy. Moreover, FL's robust privacy protection mechanisms make it more suitable for training autonomous driving object detection models. In contrast, SVL's limited local data from a single vehicle results in slower convergence and lower accuracy, underscoring the importance of data in training autonomous driving detection models.

5.3.2. Validation of Personalized Strategy Utilizing SGD Adaptation Mechanism

Herein, we compare and validate the effectiveness of personalized strategies employing the SGD adaptation mechanism, further delving into the impact of this technique on the training of detection models within FL. In Figure 5, we present the training results of FL on the nuScenes dataset, where the parameter β is set to vary from 0.1 to 0.5.



(**a**) (**b**) (**c**)

Figure 5. The training results of the personalized strategy. (**a**) Training loss function outcomes. (**b**) Training mAP results. (**c**) The influence of parameter β.

Figure 5 presents comprehensive experimental outcomes on the nuScenes dataset, showcasing the merits of our personalized FL approach integrated with SGD adaptation. As evident in Figure 5a, our method attains a faster convergence of the loss function, surpassing conventional FL by approximately 25 communication rounds and achieving a lower loss value of 0.2213 compared to 0.4531 without SGD. Similarly, Figure 5b illustrates a notable 1.06% enhancement in mAP, confirming the efficacy of our personalized strategy in accelerating user adaptation to local data distributions and enhancing model generalization. Moreover, Figure 5c delves into the impact of β, revealing an optimal $\beta = 0.3$ yielding the highest mAP of 70.23%, while lower and higher β values exhibit slower convergence or

instability, respectively. Notably, β = 0.4 and 0.5 demonstrate faster initial convergence but ultimately yield inferior and more volatile results, highlighting the delicate balance between learning rate and stability.

5.3.3. Effectiveness of the Balanced Window Attention Mechanism

Herein, we experimentally compared and studied the training results of the proposed balanced window attention, traditional window attention, and standard attention. In our experiments, the nuScenes dataset was used for centralized training. To eliminate the influence of random factors, we presented the average results of 10 experiments.

As shown in Figure 6, through observation, we found that traditional window attention achieved better results compared to standard attention, with an mAP increase of 2.02% (Attention = 67.88%, Window Attention = 69.9%). Notably, the proposed BWA mechanism demonstrated superior training performance over traditional window attention, achieving an mAP of 70.31%. Additionally, we also observed that the BWA mechanism converged faster than the other two approaches. These results underscore the effectiveness of the BWA mechanism in 3D detection tasks, emphasizing the importance of effectively processing point cloud data due to its unique characteristics in autonomous driving perception tasks.

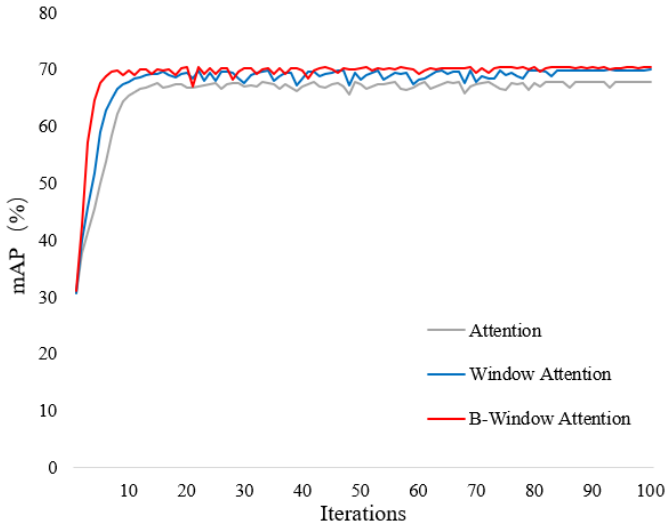


Figure 6. Training results of different attention mechanisms.

5.3.4. Influence of the Varying User Counts on Detection Performance

Given the inherent instability and dynamic evolution characteristics of the wireless networks connecting vehicles, the number of vehicles participating in FL is subject to change at any time. Herein, we investigate the impact of vehicle count on training performance by setting the number of participating vehicles to 10, 20, 30, and 40, with the nuScenes dataset correspondingly and equally divided into 10, 20, 30, and 40 portions. Figure 7 illustrates the training results.

From a meticulous analysis of Figure 7, it is evident that the training performance exhibits a positive correlation with the escalating number of user vehicles, as the mAP consistently improves across different epochs as the number of participating vehicles increases, initially from 10 to 20 and subsequently to 30. This observed trend can be rationalized by the enriched training data pool that arises from the involvement of a greater number of participants. However, the training curves for 30 and 40 vehicles exhibit a near-perfect overlap, and the mAP values are almost identical, with respective values of 70.23% and 70.22%. This finding implies that the training performance attains saturation at a certain

threshold of user vehicles, failing to demonstrate a linear or unbounded enhancement with further increases. Furthermore, curves converge around the 30th epoch, highlighting that the convergence speed of FL remains unaffected by the varying number of users. This minimal sensitivity of training outcomes to the vehicle count underscores the robustness and promising potential of FL frameworks in the realm of autonomous driving perception.

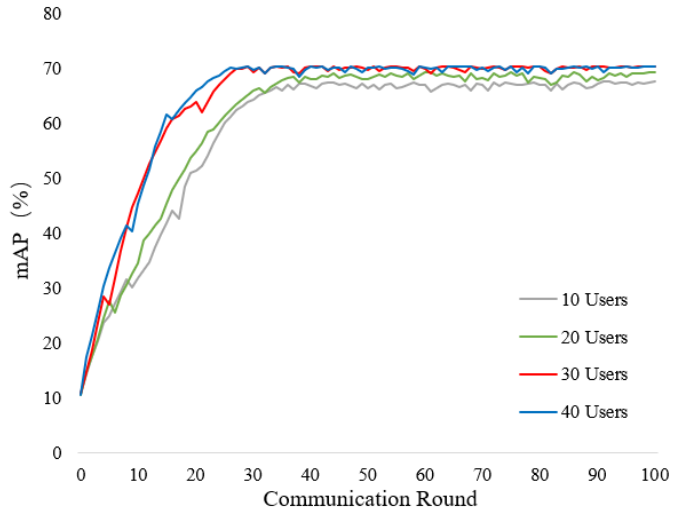


Figure 7. Comparative results of training performance across varying user counts.

5.3.5. Presentation of Visualization Results

To unequivocally exhibit the strengths of our proposed methodology, we present in Figure 8 the forward-facing visual results evaluated using the nuScenes dataset. The blue points accurately align the intricate point cloud data with the corresponding images, showcasing precise spatial correlation. Additionally, the green 3D bounding boxes tightly enclose the various detected objects, emphasizing the model's profound understanding of complex scenes. Notably, the visualized scenes highlight the exceptional capability of our model in accurately identifying even the smallest and most occluded objects, which proves its remarkable accuracy and groundbreaking achievement.

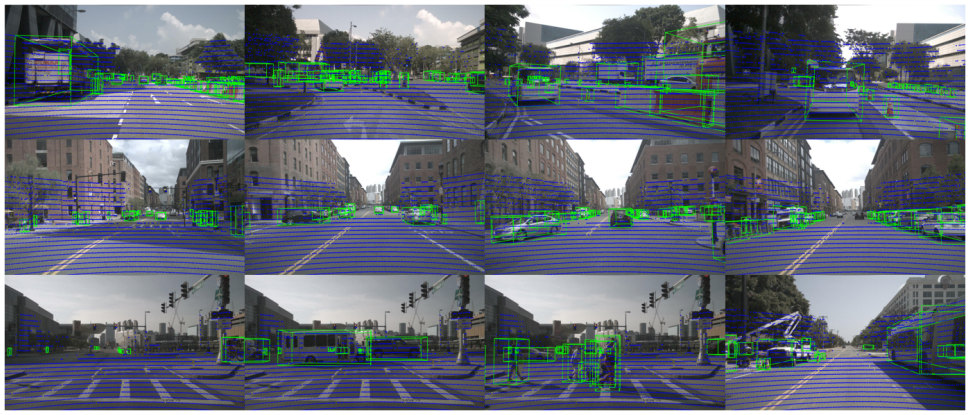


Figure 8. Forward-facing visualization results for 3D object detection in the nuScenes dataset.

In Figure 9, we present the visualization results of 3D object detection within holistic scenes from the nuScenes dataset. This figure meticulously selects and showcases six representative scenarios, wherein the dense white point clouds vividly depict the precise depth information captured by Lidar, rendering a rich and three-dimensional world. Notably, we successfully detected various objects surrounding the vehicles with high accuracy and annotated them using 3D bounding boxes. These bounding boxes are not only morphologically diverse but also colorfully distinct, with differently colored borders ingeniously

representing different types of vehicles and pedestrians, making the objects in the entire scene clearly visible and well-organized. This achievement potently demonstrates the indispensable role of multimodal information in object detection tasks. Simultaneously, it fully showcases the advanced nature and superiority of our proposed method, laying a solid foundation for our research and applications in related fields.

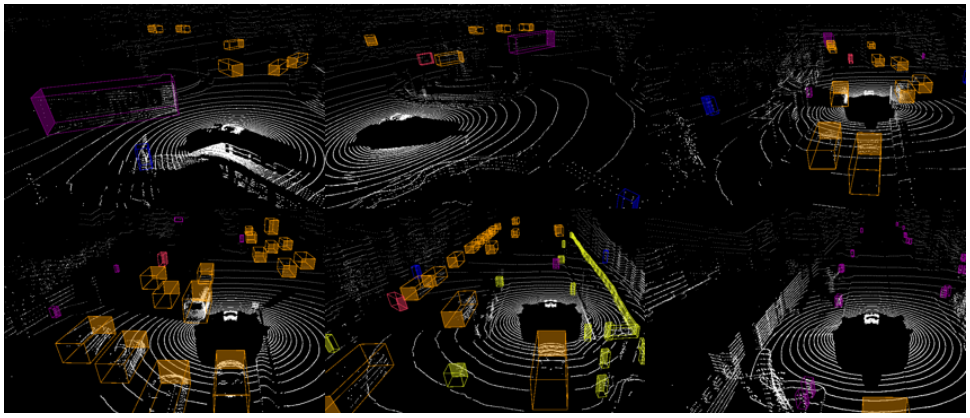


Figure 9. Visualization results for overall scenes of 3D object detection in the nuScenes dataset.

5.4. Comparative Analysis with Other State-of-the-Art Methods

We empirically assess our 3D object detection methodology across three established benchmark datasets, embarking on a meticulous comparison with the current state-of-the-art approaches. This rigorous evaluation encompasses comprehensive performance analyses, designed to systematically validate and underscore the advancements and inherent strengths of our proposed method. To ensure the fairness of the comparative analysis, for each method, we conducted 10 experiments in strict accordance with its original experimental settings on servers with identical configurations, and took the average of these 10 experimental results as a reference.

Table 3 meticulously delineates the comprehensive assessment of our novel approach on the nuScenes dataset, juxtaposing it against nine seminal 3D object detection methodologies to underscore its unparalleled performance. Our methodology stands out as a distinct outlier, achieving an impressive mAP of 71.2% and an NDS of 73.6%, significantly outperforming the current state of the art by a noteworthy margin. Remarkably, our method surpasses benchmark methods like BEVFusion and Transfusion by more than one percentage point in performance. Furthermore, despite having a computational latency of 128.9 ms, which is slightly higher than that of single-modal methods, our method stands at the forefront among multi-modal methods. This emphatically demonstrates the exceptional efficacy of personalized FL in augmenting the perceptual capabilities of autonomous driving systems.

Table 4 meticulously presents a comprehensive comparison of the test results achieved by our innovative method with six prevalent state-of-the-art approaches, all of which have been rigorously evaluated on the challenging ONCE dataset. The values presented for each class here are the overall values. The results indicate that our model attains an mAP value of 67.14%, surpassing the outstanding CenterPoints by a margin of 5.9% and significantly outperforming other competing methods. This notable enhancement underscores the remarkable efficacy and robustness of our proposed personalized FL approach in addressing the intricate and multifaceted challenges inherent within the ONCE dataset.

Table 3. Comparative analysis of methods employing the nuScenes dataset.

Methods	Modality	mAP (%)	NDS (%)	Latency (ms)
SECOND [34]	Lidar	52.8	63.3	53.2
CenterPoints [35]	Lidar	60.3	67.3	62.7
Focals Conv [36]	Lidar	63.8	70.0	66.3
LargeKernel3D [37]	Lidar	65.3	70.5	68.8
UVTR-M [38]	Camera + Lidar	67.1	71.1	262.7
MVP [19]	Camera + Lidar	66.4	70.5	148.1
AutoAlignV2 [39]	Camera + Lidar	68.4	72.4	207.0
Transfusion [6]	Camera + Lidar	68.9	71.6	164.6
BEVFusion [16]	Camera + Lidar	70.2	72.9	130.5
Ours	Camera + Lidar	71.2	73.6	128.9

Table 4. Comparative evaluation of methods utilizing the ONCE dataset.

Methods	Vehicle (%)	Pedestrian (%)	Cyclist (%)	mAP (%)
PointRCNN [40]	52.00	8.73	34.02	31.58
PointPillars [41]	69.52	17.28	49.63	45.47
SECOND [34]	69.71	26.09	59.92	51.9
PV-RCNN [42]	76.98	22.66	61.93	53.85
PointPainting [15]	66.46	47.62	65.27	59.78
CenterPoints [35]	66.35	51.8	65.57	61.24
Ours	78.96	52.41	70.88	67.14

Table 5 presents a detailed comparison of the test results of our novel method on the rigorous Waymo dataset, juxtaposed against five leading 3D object detection techniques. A comprehensive analysis reveals that our method stands out among all six competitors, achieving an mAP of 83.9% and an mAPH of 81.8% at the L1 level, significantly surpassing other advanced methods. Furthermore, at the L2 level, our method attains an mAP of 78.3% and an mAPH of 77.1%, also positioning itself at the forefront. This demonstrates the ability of our method to achieve precise perception and detection even in complex environments. It also highlights the unparalleled effectiveness and advancements of our personalized FL approach and BWA mechanism in addressing the inherent complexities of the Waymo dataset.

Table 5. Comparative assessment of methods using the Waymo dataset.

Methods	L1		L2	
	mAP (%)	mAPH (%)	mAP(%)	mAPH (%)
SECOND [34]	66.2	61.4	58.3	54
SST [43]	76.5	71.7	67.8	63.4
CenterPoints [35]	74.4	71.7	68.2	65.8
AFDetV2 [44]	79.7	78.2	74	73.1
Cascade3D [45]	81.1	79.6	75.8	74.3
Ours	83.9	81.8	78.3	77.1

In the comparative experiments conducted on these three datasets, our proposed method demonstrated superior performance across multiple dimensions, primarily attributed to the deep learning capability of our robust personalized FL framework in diverse traffic conditions and complex scenarios, as well as the efficiency and powerful parallel computing capabilities exhibited by the BWA mechanism in processing point cloud data.

6. Conclusions

This paper presents a pioneering approach for multimodal 3D object detection in autonomous driving, leveraging personalized FL within the perception task. By integrating FL, we effectively address privacy concerns surrounding highly sensitive data while ensuring precise 3D object detection, a cornerstone for enhancing accuracy and safety in autonomous systems. Our framework uniquely tailors local models for individual vehicles, with a centralized server orchestrating global model updates. The personalized strategy significantly enhances the generalization capability of the global model, thereby boosting precision and robustness in 3D object detection. Comprehensive experiments validate the efficacy of our method, demonstrating improved detection performance alongside mitigated privacy risks and optimized communication efficiency. However, in practical applications of the FL framework, users often exhibit imbalanced characteristics in terms of data distribution, device models, and computing capabilities. This imbalance poses a significant challenge to the training effectiveness of the global model, making it difficult for 3D object detection tasks to meet predefined performance benchmarks. To tackle this challenge, future research endeavors should deeply explore ways to optimize the FL framework, such as designing lightweight model architectures to alleviate device burdens and leveraging data synchronization techniques to address device heterogeneity. We envision this reliable and potent 3D detector as a cornerstone for autonomous driving, empowering other tasks and accelerating the widespread adoption of advanced technologies on intelligent vehicles. Ultimately, it paves the way for safer, more efficient, and universally accessible autonomous transportation systems.

Author Contributions: Conceptualization, L.Z. and X.L.; methodology, X.L.; validation, L.Z. and X.L.; formal analysis, L.F.; investigation, X.L.; resources, L.Z.; data curation, X.L.; writing—original draft preparation, X.J.; writing—review and editing, L.Z.; visualization, X.L.; supervision, X.J.; funding acquisition, L.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant No.61473114), the Key Scientific Research Projects of Higher Education Institutions in Henan Province (Grant No.24B520006), and the Open Fund of Key Laboratory of Grain Information Processing and Control (Henan University of Technology), Ministry of Education (Grant No.KFJJ2024013).

Data Availability Statement: The original contributions presented in the study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Song, Z.; Liu, L.; Jia, F.; Luo, Y.; Jia, C.; Zhang, G.; Yang, L.; Wang, L. Robustness-aware 3d object detection in autonomous driving: A review and outlook. *IEEE Trans. Intell. Transp. Syst.* **2024**, *11*, 15407–15436. [CrossRef]
2. Yang, S.; Lu, H.; Li, J. Multifeature fusion-based object detection for intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.* **2022**, *24*, 1126–1133. [CrossRef]
3. Staravoitau, A. Traffic sign classification with a convolutional network. *Pattern Recognit. Image Anal.* **2018**, *28*, 155–162. [CrossRef]
4. Chen, Z.; Zhang, J.; Tao, D. Progressive lidar adaptation for road detection. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 693–702. [CrossRef]
5. Dao, M.Q.; Berrio, J.S.; Frémont, V.; Shan, M.; Héry, E.; Worrall, S. Practical Collaborative Perception: A Framework for Asynchronous and Multi-Agent 3D Object Detection. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 12163–12175. [CrossRef]
6. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.L. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
7. Lin, Y.M.; Gao, Y.; Gong, M.G.; Zhang, S.J.; Zhang, Y.Q.; Li, Z.Y. Federated learning on multimodal data: A comprehensive survey. *Mach. Intell. Res.* **2023**, *20*, 539–553. [CrossRef]
8. Sun, F.; Diao, Z. Federated Learning and Blockchain-Enabled Intelligent Manufacturing for Sustainable Energy Production in Industry 4.0. *Processes* **2023**, *11*, 1482. [CrossRef]

9. Kong, Q.; Yin, F.; Lu, R.; Li, B.; Wang, X.; Cui, S.; Zhang, P. Privacy-preserving aggregation for federated learning-based navigation in vehicular fog. *IEEE Trans. Ind. Inform.* **2021**, *17*, 8453–8463. [CrossRef]
10. Posner, J.; Tseng, L.; Aloqaily, M.; Jararweh, Y. Federated learning in vehicular networks: Opportunities and solutions. *IEEE Network* **2021**, *35*, 152–159. [CrossRef]
11. Feng, D.; Haase-Schütz, C.; Rosenbaum, L.; Hertlein, H.; Glaeser, C.; Timm, F.; Wiesbeck, W.; Dietmayer, K. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 1341–1360. [CrossRef]
12. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3d object detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XV 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 35–52.
13. Roriz, R.; Cabral, J.; Gomes, T. Automotive LiDAR technology: A survey. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 6282–6297. [CrossRef]
14. Wang, X.; Li, K.; Chehri, A. Multi-sensor fusion technology for 3D object detection in autonomous driving: A review. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 1148–1165. [CrossRef]
15. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. Pointpainting: Sequential fusion for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4604–4612.
16. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA), London, UK, 29 May–2 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 2774–2781.
17. Wang, H.; Shi, C.; Shi, S.; Lei, M.; Wang, S.; He, D.; Schiele, B.; Wang, L. Dsvt: Dynamic sparse voxel transformer with rotated sets. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13520–13529.
18. Zhang, C.; Wang, H.; Cai, Y.; Chen, L.; Li, Y. TransFusion: Multi-Modal Robust Fusion for 3D Object Detection in Foggy Weather Based on Spatial Vision Transformer. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 10652–10666. [CrossRef]
19. Yin, T.; Zhou, X.; Krähenbühl, P. Multimodal virtual point 3d detection. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 16494–16507.
20. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3d object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2147–2156.
21. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.
22. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 652–660.
23. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *arXiv* **2017**, arXiv:1706.02413.
24. Zhou, Y.; Tuzel, O. Voxelnet: End-to-end learning for point cloud based 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4490–4499.
25. Xu, D.; Anguelov, D.; Jain, A. Pointfusion: Deep sensor fusion for 3d bounding box estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 244–253.
26. Mansour, Y.; Mohri, M.; Ro, J.; Suresh, A.T. Three approaches for personalization with applications to federated learning. *arXiv* **2020**, arXiv:2002.10619.
27. Tan, A.Z.; Yu, H.; Cui, L.; Yang, Q. Towards personalized federated learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *34*, 9587–9603. [CrossRef] [PubMed]
28. Lim, W.Y.B.; Luong, N.C.; Hoang, D.T.; Jiao, Y.; Liang, Y.C.; Yang, Q.; Niyato, D.; Miao, C. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Commun. Surv. Tutorials* **2020**, *22*, 2031–2063. [CrossRef]
29. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 12124–12134.
30. Sun, P.; Tan, M.; Wang, W.; Liu, C.; Xia, F.; Leng, Z.; Anguelov, D. Swformer: Sparse window transformer for 3d object detection in point clouds. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 426–442.
31. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. Nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11621–11631.
32. Mao, J.; Niu, M.; Jiang, C.; Liang, H.; Chen, J.; Liang, X.; Li, Y.; Ye, C.; Zhang, W.; Li, Z.; et al. One million scenes for autonomous driving: Once dataset. *arXiv* **2021**, arXiv:2106.11037.

33. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2446–2454.
34. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef]
35. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 11784–11793.
36. Chen, Y.; Li, Y.; Zhang, X.; Sun, J.; Jia, J. Focal sparse convolutional networks for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5428–5437.
37. Chen, Y.; Liu, J.; Zhang, X.; Qi, X.; Jia, J. Largekernel3d: Scaling up kernels in 3d sparse cnns. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 13488–13498.
38. Li, Y.; Chen, Y.; Qi, X.; Li, Z.; Sun, J.; Jia, J. Unifying voxel-based representation with transformer for 3d object detection. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 18442–18455.
39. Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F. Deformable feature aggregation for dynamic multi-modal 3D object detection. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Cham, Switzerland, 2022; pp. 628–644.
40. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 770–779.
41. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 12697–12705.
42. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10529–10538.
43. Fan, L.; Pang, Z.; Zhang, T.; Wang, Y.X.; Zhao, H.; Wang, F.; Wang, N.; Zhang, Z. Embracing single stride 3d object detector with sparse transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 8458–8468.
44. Hu, Y.; Ding, Z.; Ge, R.; Shao, W.; Huang, L.; Li, K.; Liu, Q. Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 969–979.
45. Liu, S.; Zhang, J.; Li, T.; Yan, H.; Liu, J. A cascade 3D U-Net for dose prediction in radiotherapy. *Med. Phys.* **2021**, *48*, 5574–5582. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.