The Hong Kong University of Science and Technology
Research Proposal

Collaborative Intelligent Perception for Internet of Vehicles System

Peng Haosong

School of Automation
Beijing Institute of Technology, Beijing

October 11, 2025

Key Words

Internet of vehicles, Scene reconstruction, Federated learning, Continuous learning.

Abstract

This proposal explores collaborative intelligent perception within the Internet of Vehicles (IoV) systems, addressing critical challenges in real-time data acquisition, processing, and analysis. It focuses on three main aspects: **large-scale scene reconstruction using a hierarchical cloud-infrastructure-vehicle framework**; **heterogeneity-aware federated learning to tackle data, model, and modality heterogeneities**; and **efficient model adaptation via serverless functions for continuous learning**. These innovations aim to improve the computational efficiency, privacy, scalability, and adaptability of IoV systems in complex environments. The proposed research contributes to intelligent transportation systems, aligning with the digitalization initiatives outlined in the National Overall Layout Plan for Building Digital China. The expected outcomes will advance autonomous driving technologies and facilitate large-scale deployment of robust perception models for IoV systems.

1 Introduction

With the rapid development of intelligent transportation systems, the demand for advanced intelligent perception in Internet of Vehicles (IoV) systems has grown significantly [1]. In today's traffic scenarios, accurate, real-time information acquisition and perception are critical for enhancing traffic safety, optimizing traffic flow, and enabling autonomous driving. Traditionally, single-vehicle perception limits a vehicle's understanding of its surroundings [2], making it difficult to anticipate risks from long distances or blind spots, thereby falling short in ensuring traffic safety. Thanks to the rapid advancement of edge computing, cloud computing, and IoT technologies, the Internet of Vehicles, also known as V2X, has emerged as a transformative innovation. Specifically, as depicted in Figure. 1, a typical IoV mainly includes vehicle-to-vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-cloud (V2C), etc., which enable each vehicle to function as a powerful mobile terminal. This facilitates information exchange between vehicles, infrastructure (Edge), and the cloud, making previously unattainable tasks possible. For example, (a) Sensor information sharing between vehicles eliminates blind spots in single-vehicle perception [3]; (b) Multi-vehicle distributed city scene reconstruction, with data uploaded to the cloud to obtain high-precision maps [4]; (c) Using data from the infrastructure sensors to assist vehicles in object detection [5]; (d) Sensor networks facilitate vehicle re-identification, enabling the construction of accurate vehicle trajectory tracking [6]. Internationally, the market for the IoV industry is projected to grow from $161.51 billion in 2024 to $344.82 billion in 2029, representing a compound annual growth rate of 16.38% [7], while the scale of China's IoV industry reached 288.72 billion yuan in 2023, which is expected to exceed 320 billion yuan in 2024 [8].

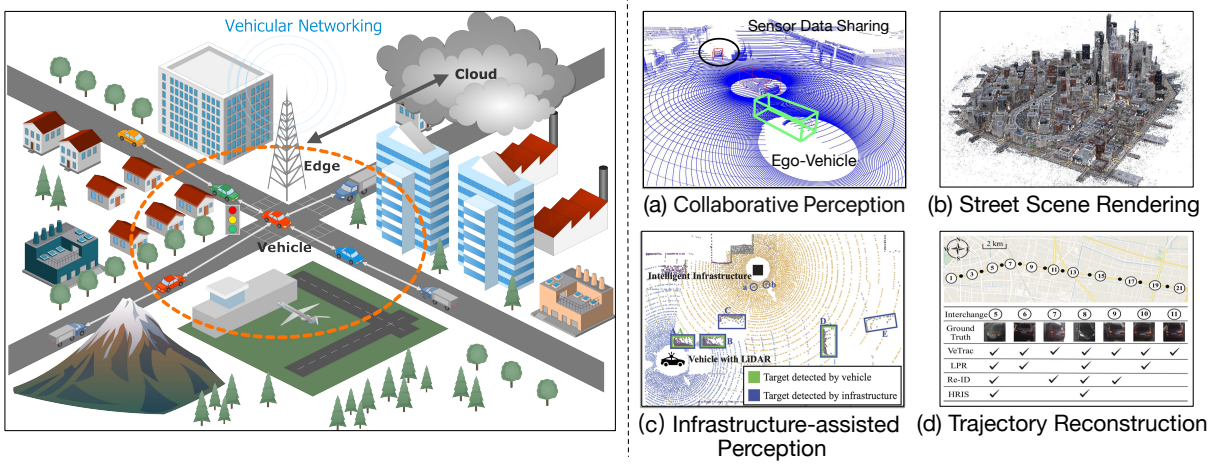

Figure 1: **Left**: A typical IoV system. **Right**: Applications of IoV. (a) Collaborative perception through sensor data sharing among vehicles [3]. (b) Distributed scene reconstruction and rendering [4]. (c) Infrastructure-assisted Perception [5]. (d) Traffic control through large-scale vehicle re-identification and trajectory reconstruction [6].

The above evidence reveals the promising prospects of collaborative perception and its pivotal role in intelligent transportation. However, research on collaborative perception in IoV system is still an open issue, with several key challenges remaining:

1. **The lack of effective large-scale scene reconstruction planning algorithms for heterogeneous IoV systems.** The recently emerged 3D Gaussian Splatting (3DGS) [9] has

increasingly become a popular scene reconstruction algorithm due to its outstanding performance. However, current large-scale scene reconstruction methods [4, 10, 11, 12] using 3DGS remains limited to centralized strategies, where vehicles upload collected images to a cloud server for centralized modeling. This approach poses challenges in terms of **efficiency**, **privacy and security**, and **scalability**. Designing a distributed scene reconstruction and planning algorithm is challenging, especially given the heterogeneity in vehicle computational capabilities. Inefficient task allocation can lead to the Buckets effect, significantly hindering overall task completion efficiency.

2. **Heterogeneous data and modalities make it challenging to train vehicle perception models.** In the current automotive industry landscape, vehicle manufacturers operate independently and emphasize the privacy of their data assets. Against this backdrop, federated learning (FL) [13] offers an innovative and highly promising approach to training vehicle perception models. However, in real-world scenarios, the training data and annotation quality across clients are often inconsistent and heterogeneous, and some may lack reliable modal data [14], resulting in poor performance of the aggregated perception model.

3. **The high retraining cost of vehicle perception models.** Continuous learning of vehicle perception models [2, 15] is an effective approach to address data drift i.e., the live data diverges significantly from the data that was used for training [16], especially in the context of rapidly changing road traffic conditions, varying weather and lighting, and complex road scenarios. However, current continuous learning methods [17, 18, 19, 20] involve updating all model parameters, requiring multiple iterative training cycles on new data. This process is highly resource-intensive and incurs significant latency, making it unacceptable for real-time autonomous driving scenarios.

Based on this, the proposal aims to explore collaborative intelligent perception for the IoV System by integrating edge computing, federated learning, and continuous learning. The research focuses on three main aspects: **1) large-scale scene reconstruction planning based on cloud-infrastructure-vehicle (hierarchical) framework; 2) heterogeneity-aware federated learning for robust perception; 3) serverless-assisted efficient model adaption for vehicle perception model.** During my PhD, I will focus on these aspects to advance the research field and address practical challenges in real-world applications. The expected outcomes will accelerate the deployment of the vehicle perception model for autonomous driving. This research aligns with the National Overall Layout Plan for Building Digital China and holds significant scientific and practical value.

2 Literature Review

In this section, I first provide a brief overview of the collaborative perception issues in the IoV system, followed by a detailed discussion of the three previously mentioned challenges and the related work. Finally, the research gaps are presented.

Due to the increasing importance, analysis of perception issues in IoV system has recently attracted much attention from researchers in the industry (e.g., NVIDIA [21, 22], Baidu [23, 24], Waymo [25, 26], Wayve [27],) and academic research labs (e.g., The Chinese University of Hong Kong [28, 5, 29, 30], Tsinghua University [31, 32]). Some recent works release new datasets [33, 34], and propose new settings and baselines [35] in the face of potential challenges in collaborative vehicle perception applications, giving momentum to developments in

the field. On the one hand, to deal with the **1) Insufficient environmental perception information problem**, Some approaches aim to enhance single-vehicle perception by leveraging multi-vehicle [3, 36, 37, 38] or infrastructure [5, 28] information fusion. On the other hand, **2) Limited computational capacity of vehicles** poses a challenge as onboard sensors generate massive and complex data. Moreover, the diversity in data formats from multi-source heterogeneous sensors further increases the difficulty of data processing. As a result, existing methods leverage cloud [39] or edge [40, 41, 42] to offload workload to accelerate information processing. However, despite the attention this field has received and its recent progress, there are still many challenges that need to be investigated.

2.1 3D Scene Reconstruction

Traditional approaches [43, 44, 45] follow a structure-from-motion pipeline that estimates camera poses and generates sparse point clouds. However, such methods often contain artifacts or holes in areas with limited texture or speculate reflections as they are challenging to triangulate across images. Recently, NeRF [46] and 3DGS [9] variants have become a worldwide 3D representation system thanks to their photo-realistic characteristics and the ability of novel-view synthesis, which inspires many works [47, 48, 49, 50, 11, 51, 52] to extend it into large-scale scene reconstructions. The above methods can be categorized into centralized and distributed frameworks. Centralized methods [51, 48] adopt the integration of NeRF-based and grid-based methods to model city-scale reconstruction. Distributed methods [11, 49] apply scene decomposition for multiple NeRF / Gaussian models optimization. However, with the growing scene size, both centralized and distributed variants limit their scalability due to the central server's limited data storage and unacceptable computation costs. Nearly proposed Fed3DGS [52] introduces federated learning to leverage the computational resources of clients and emerge clients' 3DGS models into the central server by a distillation update scheme. Nevertheless, this method ignores the scales and the bundle adjustment between different edge models, resulting in a performance drop. Meanwhile, it only focuses on over-fitting the photo-realistic rendering but ignores the geometry performance.

2.2 Large-Scale Federated Learning Framework

The hierarchical cloud-edge-device framework [53, 54, 55] has been proposed to address large-scale distributed edge computing problems and has been widely applied in various distributed domains in recent years. Wang et al. [56] identified the optimal clustering configuration and implemented hybrid federated learning using a blend of synchronous and asynchronous methods. Cui et al. in [54] introduced a heuristic user selection strategy that takes into account heterogeneity to choose devices and set their operating frequencies. Li et al. introduced FedGS [57] for the IIoT environment, which utilizes the gradient-based binary permutation algorithm (GBP-CS) to select subsets. Deng et al. in [55] proposed modifications to the topology of the HFL framework, taking into account data distribution and communication costs. Mhaisen et al. [58] developed an optimization for the pairing of devices and edges to reduce the distance in class distribution. Zhong et al. [59] proposed the FLEE algorithm, which leverages edges and devices to achieve dynamic model updates. Deng et al. [60] proposed FedHKT, a hierarchical framework that utilizes a hybrid knowledge transfer mechanism to enhance learning efficiency and performance. Yang et al. proposed HierMo[61], which applies momentum to accelerate distributed training. Chen et al. [62] introduced SD-GT, which enhances federated learning on fog networks by optimizing device updates and improving model performance.

Abdellatif et al. [63] proposed a training algorithm for user allocation and resource allocation on heterogeneous edge nodes.

2.3 Continuous Learning for Vision Models

Perception models [64, 15] in autonomous systems must adapt to changing road conditions and traffic patterns. Continuous learning [65] focuses on enabling models to adapt dynamically as new data becomes available. A widely adopted technique in this domain is transfer learning [66, 67]. Recently, online model distillation [68] specializes in low-cost semantic segmentation models for specific video streams, achieving significant runtime and computational cost reductions while maintaining high accuracy and temporal stability. AMS [17] utilizes edge servers to continuously train and fine-tune lightweight models deployed on edge devices, improving the performance of semantic segmentation on mobile cameras. Ekya [18] addresses the data drift problem in video perception by proposing a resource allocation method for inference and retraining on edge servers, reducing resource requirements while improving task accuracy. RECL [19] is a continuous learning system that integrates model reuse and online retraining to address data drift efficiently. Kong et al. [20] proposed an edge-assisted framework to improve real-time video analytics in adverse environments on resource-constrained cameras through model updates.

2.4 Research Gaps

Several crucial challenges and motivations requiring further investigation are identified and described below.

2.4.1 Efficient Large-scale Scene Reconstruction

Although existing visual-based reconstruction technology - 3DGS [9] demonstrates excellent performance for individual objects and small-scale scene reconstruction, it still faces several challenges in the field of multi-vehicle collaborative reconstruction. **1) Efficiency.** When dealing with large-scale scene data (i.e., $>$ 1,000 raw images), centralized frameworks [11, 4, 10, 12] (as shown in Figure 3 **Left**) requires the support of distributed training and workload partitioning due to its substantial GPU memory consumption. Additionally, for vehicles with heterogeneous computational resources, imbalanced workload partitioning can cause a bucket effect, further reducing efficiency. Such inefficiency is unacceptable for time-critical perception tasks. **2) Privacy and Security.** To obtain the initial model for the scene, all raw images captured by the vehicles must be uploaded to the cloud for initialization processing [11, 4, 10, 12]. From the perspective of vehicles, they may be unwilling to upload real images or share location information associated with the photos to the cloud, as it could expose their privacy. From the perspective of attackers, uploaded images could be maliciously tampered with to slow down 3DGS training on the cloud and increase memory overhead [69]. Therefore, centralized frameworks cannot mitigate the aforementioned issues. **3) Scalability.** The centralized framework requires applying different strategies [11, 4] to distribute the training workload after initialization, which often takes considerable time. When the scene to be reconstructed expands or a new region is added, newly captured images must be uploaded to the cloud server, restarting the overall process. As a result, the centralized framework lacks scalability. In conclusion, a novel distributed multi-vehicle reconstruction system that jointly considers high efficiency,

privacy and security, and scalability will significantly enhance the practical value of modern high-precision map models.

2.4.2 Federated Learning for Vehicle Perception Model

Federated learning is widely adopted for training vehicle perception models due to its ability to preserve the privacy of data from individual vehicles. However, heterogeneity [70, 71] poses a critical challenge in the practical deployment of FL. **1) Label Skew:** The distributions across participating clients vary significantly. For example, some clients employ meticulous manual annotation, ensuring that both foreground and background information in their collected samples is fully distinguished. Other clients may rely on machine learning models for annotation, which can result in missing positive samples. Conclusively, an imbalanced label distribution can negatively impact the accuracy of the final aggregated model [14]. **2) Model Heterogeneity:** The computational power of onboard chips varies across vehicles, leading to the deployment of different perception models. As a result, varying model sizes [72], or even entirely different models [73], pose significant challenges for transferring knowledge between heterogeneous clients in FL. **3) Heterogeneous Modality:** vehicle perception model often relies on information from multiple modalities (e.g., RGB, LiDAR). However, in cases of sensor failure, training samples may lack data from certain modalities, which can significantly impact model accuracy. A practical FL framework should take all these factors into account simultaneously.

2.4.3 Efficient Continuous Learning for Vehicle Perception Model

Existing continuous learning systems [18, 19, 20] require vehicles to continuously transmit new sample video frames to edge servers, where an expert model (golden model) or human annotators label new classes. The updated perception model is then trained and transmitted back to vehicles. With the increasingly complex structures and growing parameters of vision perception models [74, 75], the time required for model inference and training has also significantly increased. Meanwhile, fine-tuning paradigms [76, 77] (i.e., LoRA adapter) have become a common adaptation method for these large models. Introducing efficient fine-tuning paradigms into the continuous learning process of vehicle perception models could greatly enhance update frequency, leading to improved generalization performance. A well-designed trigger mechanism and model reuse strategy can significantly reduce computational overhead and communication costs.

3 Methods

3.1 Road Map

Addressing the opportunities and challenges of cloud-edge collaborative intelligent perception systems for IoV, this study takes vehicle perception model training and scene reconstruction planning as its entry point. To address the task of efficient scene reconstruction, a large-scale scene reconstruction based on a cloud-infrastructure-vehicle framework is proposed in Sec. 3.2. For the heterogeneous characteristics of vehicle perception models, heterogeneity-aware federated learning for robust perception is introduced in Sec. 3.3. To tackle the challenge of continuous training of perception models, a serverless-assisted efficient model adaptation for

vehicle perception is developed in Sec. 3.4. Finally, a prototype system is implemented for validation and demonstration. The road map of this project is shown in Figure 2.

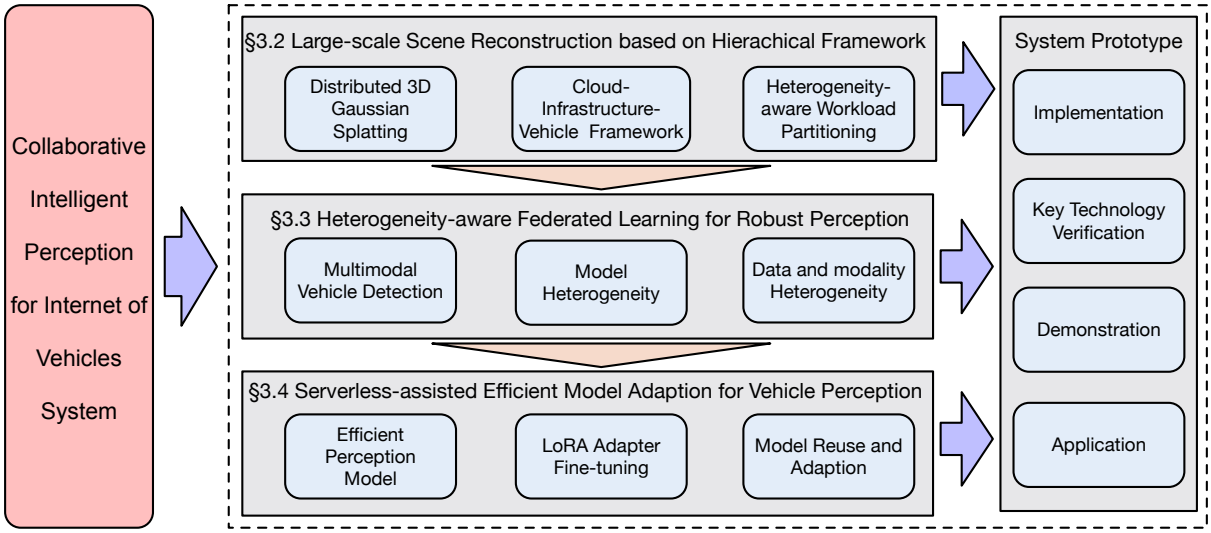

Figure 2: Road map of this project

3.2 Large-scale Scene Reconstruction based on Hierachical Framework

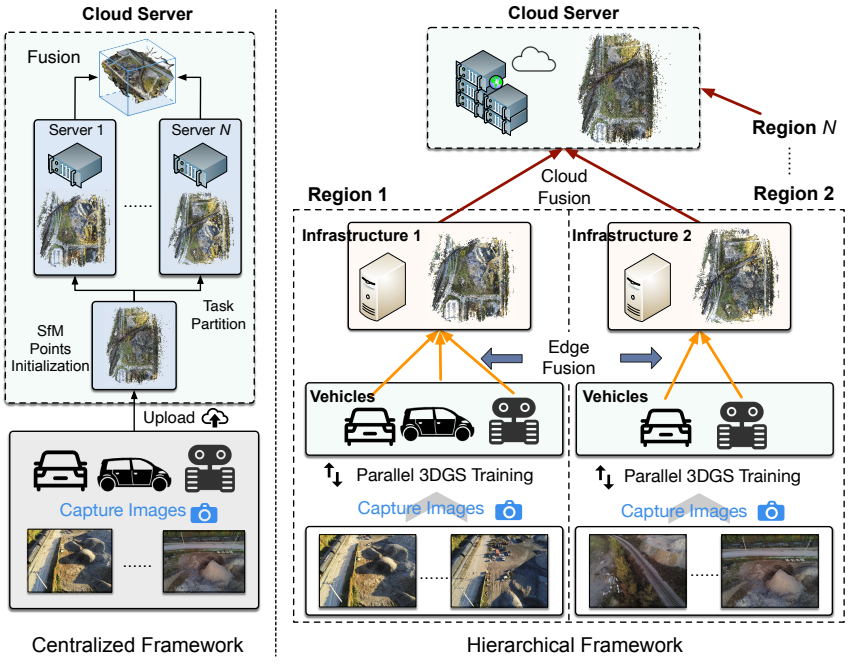

Figure 3: **Left**: Centralized framework. **Right:** Our hierarchical framework for scene reconstruction.

WP1: Develop Hierarchical Framework for Distributed 3DGS Scene Reconstruction.

We employ a cloud-infrastructure-vehicle hierarchical architecture to achieve efficient scene reconstruction. As shown in the right part of Fig. 3, our hierarchical architecture is divided into three layers: vehicle, infrastructure, and cloud. The overall scene is partitioned

into multiple regions, with each region managed by multiple vehicles. Specifically, the vehicles capture their images and independently train the models, which are then uploaded to the infrastructure for preliminary fusion before being further fused in the cloud. In this way, the computational loads are ingeniously reallocated from the cloud to the infrastructure, which can significantly mitigate the burden in the cloud. Besides, the images captured and the model trained by vehicles in different regions do not require sharing or direct uploading to the cloud, thus preserving data and location privacy. Moreover, this framework effectively enhances the scalability of the system. For new scenes, they can be developed as a new area by adding edges and devices to obtain a global model, without impacting the already trained regions. Our design adheres to the following objectives: **1) Efficiency**. We allocate the workload across each vehicle appropriately to minimize completion time while enhancing performance in the large-scale system. **2) Scalability**. Our framework allows new scenes to be added at any time, and once trained, these can be integrated with the models of previous scenes. **3) Privacy and Security**. Images captured by vehicles are not uploaded to the edge or cloud, nor is there any data exchange between vehicles. Additionally, the cloud does not have access to the location information of each vehicle.

WP2: Heterogeneity-Aware Workload Partitioning (HWP).

Although the hierarchical framework can effectively address the problems mentioned in Sec. 2.4.1, significant heterogeneity in real-world systems can lead to a severe straggler effect, ultimately impacting the efficiency of scene reconstruction [56]. For example, the varying number of images on each vehicle leads to differences in model sizes, resulting in heterogeneous communication time for fusion. Even with the same number of images, training time differs across vehicles with varying computational power. On the other hand, the geographical locations of the infrastructure and the camera positions are often constrained in real-world environments. Therefore, it is essential to consider how to partition regions geometrically for each infrastructure and allocate workloads (i.e. camera positions) appropriately based on the heterogeneity of each vehicle to avoid the Bucket effect. The basic idea of the HWP is to achieve load balancing across all regions, which ensures that all infrastructures complete their tasks simultaneously. Suppose that we have the coordinates of all infrastructures, the computational power of each vehicle, and the positions of all images that need to be collected. The output is the final region of each infrastructure and its camera position list. The HWP algorithm contains the following steps (Figure 4 illustrates an example of the HWP algorithm.).

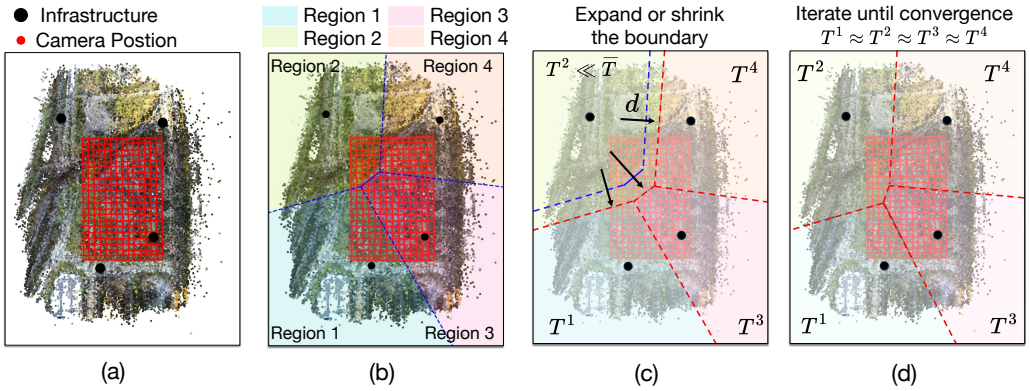

Figure 4: An example of heterogeneity-aware workload partitioning algorithm. (a) Original map of camera positions and infrastructures. (b) Initialize the regions with a Voronoi diagram. (c) Iteratively adjusts the boundaries of the region to balance the load. (d) Converges to equilibrium.

1) Initializing the regions. We generate the initial division of N regions following the rule of the Voronoi diagram [78]. In the Voronoi diagram, the distance from all camera positions within each region to their respective infrastructure center $\mathcal{P}$ is shorter than that of any other infrastructure center. **2) Estimate the completion time.** The camera positions within each region are incorporated into the corresponding infrastructure. We estimate the completion time for all infrastructures using offline profiling, then calculate their average time $\overline{T}$. **3) Move the boundary of the outlier.** We select the infrastructure with the greatest deviation from the average estimated time, determined by: $m^* = \arg\max_m\{|T^m - \overline{T}|\}$. If T^{m^*} of infrastructure m^* exceeds the average $\overline{T}$, the boundary is shrunk by a distance d in the direction of the boundary's normal; if it is below the average, the boundary is expanded by d in the normal direction. **4) Repeat until the condition is satisfied.** At each round, the completion time of each infrastructure for the new regions can be re-estimated, and the process is repeated until the residual completion time of every infrastructure is below a specified threshold.

Using the designs from WP1 and WP2, we will create a large-scale hierarchical framework for scene reconstruction. The performance will be evaluated in data sets such as Mill [49] and UrbanScene3D [79], using latency and reconstruction quality (e.g. PSNR, SSIM [80] and LPIPS [81]) as quantitative metrics.

3.3 Heterogeneity-Aware Federated Learning for Robust Perception

WP3: Modified Loss Function for Label Skew and Data Recovery for Modality deficiency.

According to the challenges mentioned in Sec. 2.4.2, data annotation heterogeneity and modality heterogeneity significantly impact the training accuracy of the global model. To address these issues, we propose modifying the model loss function to mitigate heterogeneity. For missing modalities, we consider the mutual conversion between RGB images and LiDAR data to generate reliable data.

Specifically, we propose a modified cross-entropy (MCE) loss function in the region proposal network (RPN) [82]. When the global model's predicted probability p exceeds a threshold p_{th} and the label is annotated as background (P^*), the MCE loss is set to 0. This adjustment prevents the propagation of erroneous gradients that might misclassify incorrectly labeled foreground as background. When point cloud data is missing, we propose using Dust3r [83] to recover point cloud data from RGB images. Input images are encoded using a shared-weight Vision Transformer (ViT) encoder. The decoder employs a cross-attention mechanism to share information between two branches, with a regression head outputting the point cloud and confidence map. The Dust3r model is pre-trained and does not participate in the training process of FL. When RGB images are missing, point cloud projection is used to recover 2D image based on the camera's projection model.

WP4: Heterogeneity-Aware Federated Learning for Heterogeneous Models.

When vehicle perception model structures differ, we use a public dataset for collaborative learning to facilitate knowledge transfer. Specifically, we consider K clients and one server. We define $\mathcal{C}$ as the collection of all clients, and $|\mathcal{C}| = K$. Therefore, the k-th client $c_k \in \mathcal{C}$ has a private dataset $\mathcal{D}_k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$. In addition, client c_k holds a local model θ_k with a specific neural architecture, and $f(\cdot)$ denotes the network. The logits output of x^k calculated on θ_k is represented as $f(x^k, \theta_k)$. The server cannot access the clients' datasets, but it has a public dataset $\mathcal{D}_0 = \{x_i^0\}_{i=1}^{N_0}$.

In the collaborative learning phase, we use the public dataset $\mathcal{D}_0$ as a bridge for communication between clients. At the $t_c \in T_c$ epoch of collaborative learning, each client c_k uses its local model $\theta_k^{t_c}$ to calculate the logits output on the public dataset $\mathcal{D}_0$. In this way, the knowl-

edge distribution $\mathcal{R}_k^{t_c} = f(\mathcal{D}_0, \theta_k^{t_c})$ of client c_k is obtained. Furthermore, the client uses the Kullback–Leibler (KL) divergence to measure the difference in knowledge distribution from other clients. Given two different clients c_{k_1} and c_{k_2}, the KL divergence is expressed as:

$$\mathrm{KL}(\mathcal{R}_{k_1}^{t_c} \parallel \mathcal{R}_{k_2}^{t_c}) = \sum \mathcal{R}_{k_1}^{t_c} \log \frac{\mathcal{R}_{k_1}^{t_c}}{\mathcal{R}_{k_2}^{t_c}}. \tag{1}$$

Minimizing the KL difference can be considered a process in which c_{k_1} learns knowledge from c_{k_2}. Each client c_k calculates the knowledge distribution difference as:

$$\mathcal{L}_{k,t_c}^{\mathrm{kl}} = \sum_{k' \neq k} \mathrm{KL}(\mathcal{R}_{k'}^{t_c} \parallel \mathcal{R}_k^{t_c}), \tag{2}$$

where k' denotes other clients. In this method, by measuring the knowledge distribution difference of c_k, all other clients can obtain knowledge from c_k without leakage of data privacy or model design details. Clients perform collaborative learning by aligning the knowledge distribution:

$$\theta_k^{t_c} \leftarrow \theta_k^{t_c-1} - \alpha \nabla_\theta \left(\frac{1}{K-1} L_{k,t_c-1}^{\mathrm{kl}} \right), \tag{3}$$

where α represents the learning rate.

Experiments will be conducted on KITTI [84] and Waymo Open [26] Datasets using model precision as metrics.

3.4 Serverless-Assisted Efficient Model Adaptation for Vehicle Perception

WP5: Design an efficient memory reuse perception model for vehicles.

We observe a substantial amount of redundant information in vehicle camera perception [85, 86]. Therefore, designing an information skip and reuse perception model can help reduce computational overhead and increase throughput. Figure 5 shows the proposed efficient memory reuse perception model. In every K frames, the first frame is designated as the keyframe, and the subsequent $K-1$ frames are considered as non-keyframes. It operates in two distinct phases: keyframe inference and non-keyframe inference, then alternating periodically between them. In the keyframe inference phase, it first performs full frame inference by detector and caches intermediate tokens into two memory token pools. In the non-keyframe inference phase, only the interest regions (denoted as patch-of-interest (PoI)) are sent to the backbone to compute the feature, and it queries memory tokens from the pools and reconstructs complete image-wise features to accommodate unstructured samples with variable-length patch sequences. Moreover, we will introduce a Sampler to identify the PoIs in non-keyframes and provide feedback to the camera. Through this mechanism, the computational load for non-keyframes is significantly reduced, resulting in increased inference speed and throughput without compromising accuracy.

WP6: Serverless-assisted Efficient Model Adaption for Vehicle Perception.

Based on the design of WP5, we plan to develop a serverless-assisted efficient model adaption system based on our proposed efficient memory reuse perception model. The overview of the system is shown in Figure 6. Each vehicle performs real-time inference on camera frames and uploads keyframes to the cloud, where they are processed by a model selector. The selector chooses a LoRA adapter for the vehicle perception model from the adapter zoo and determines

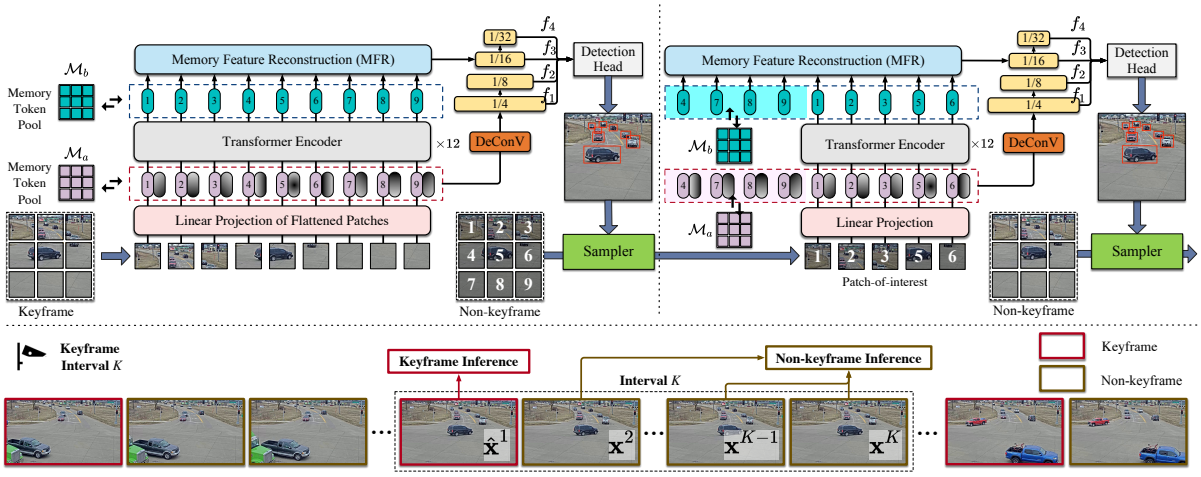

Figure 5: The proposed efficient memory reuse perception model.

whether model retraining is required. If retraining is necessary, a serverless function is invoked to fine-tune the adapter using all keyframes and correspondent labels annotated by the golden model. Finally, the updated adapter is saved back to the adapter zoo. Specifically, let the original transformer layer weight be $W \in \mathbb{R}^{m \times n}$. LoRA decomposes it as $W = W_0 + BA$, where $B \in \mathbb{R}^{m \times r}$, $A \in \mathbb{R}^{r \times n}$, and $r \ll \min(m, n)$, with r being the rank of the low-rank matrices. During forward propagation, for an input x, the original output is $y = Wx$. With LoRA, the output becomes $y = W_0 x + BAx$. In the fine-tuning process, only B and A are updated, significantly reducing the number of trainable parameters while effectively capturing task-specific information. Building on previous work [19], we use Mixture-of-Experts [87, 88] (MoE) as the model Selector, which selects a suitable expert model from a collection of history expert models to quickly adapt vehicles' needs.

For evaluation, we will use datasets such as Cityscapes [89], Waymo Open [26], and BDD100K [90], which were collected using a large number of dash cameras. We plan to implement model retaining on the Alibaba Cloud Function Compute [91]. In addition, we evaluate the proposed system using metrics such as accuracy and latency.

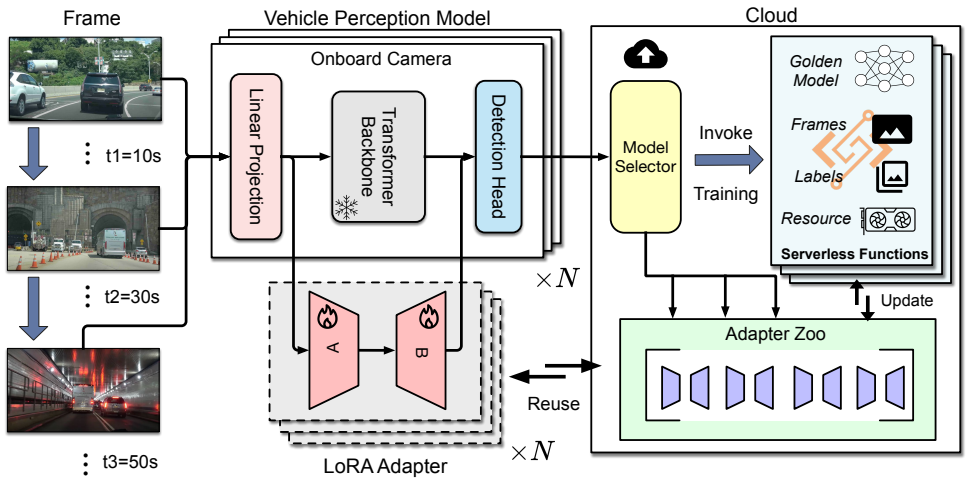

Figure 6: Overview of model adaption pipeline

4 Milestones

(1) WP1 (Month 0-6): Develop the hierarchical framework for distributed 3DGS scene reconstruction.

(2) WP2 (Month 6-12): Design the heterogeneity-aware workload partitioning algorithm. Paper writing and submission.

(3) WP3 (Month 12-18): Design the modified loss function for label skew and data recovery method.

(4) WP4 (Month 18-24): Paper writing and submission.

(5) WP5 (Month 24-30): Design an efficient memory reuse perception model for vehicles.

(6) WP6 (Month 30-36): Develop the serverless-assisted efficient model adaption system for Vehicle Perception. Paper writing and submission.

(7) System Implementation (Month 36-42): Develop and demonstrate the overall collaborative intelligent perception for the IoV system on real multiple unmanned vehicles prototype.

(8) PhD Thesis (Month 42-48): Thesis writing and defense preparation.

5 Candidate Information

Haosong Peng holds a bachelor's degree in automation from the School of Information Science and Technology, Beijing University of Chemical Technology, from 2018 to 2022. With an impressive GPA of 4.07/4.33 and ranking first in his major, he was awarded the National Scholarship twice during his undergraduate studies.

From 2022 to 2025, he studied at the Lab of Intelligent Information Processing and Control, Automation School, Beijing Institute of Technology. His main coursework included Linear System Theory, Linear Algebra in Control, Deep Learning, Image Acquisition and Processing, and Numerical Analysis. His research focused on video analytics, multimedia systems, and edge computing.

During his Master's studies, he deeply participated in several projects, including the National Natural Science Foundation of China's (NSFC) Youth Fund project on *Incentive Mechanisms for Sustainable Training in Federated Learning* and a general project on *Real-time Intelligent Video analysis Methods based on Edge Vision Foundational Models*. The applicant has several years of research experience in the design of computer vision perception systems. His work resulted in publications in journals like IEEE Transactions on Services Computing (TSC) and conferences such as the 2024 International Conference on Distributed Computing Systems (ICDCS), alongside several EI-indexed conferences. With his extensive experience in both deep learning research and practical system implementation, Haosong is fully committed to dedicating himself to this project.

References

[1] B. Ji, X. Zhang, S. Mumtaz, C. Han, C. Li, H. Wen, and D. Wang, "Survey on the internet of vehicles: Network architectures and applications," IEEE Communications Standards Magazine, vol. 4, no. 1, pp. 34–41, 2020.

[2] X. Gao, X. Zhang, Y. Lu, Y. Huang, L. Yang, Y. Xiong, and P. Liu, "A survey of collaborative perception in intelligent vehicles at intersections," IEEE Transactions on Intelligent Vehicles, 2024.

[3] X. Zhang, A. Zhang, J. Sun, X. Zhu, Y. E. Guo, F. Qian, and Z. M. Mao, "Emp: Edge-assisted multi-vehicle perception," in Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, 2021, pp. 545–558.

[4] Y. Liu, C. Luo, L. Fan, N. Wang, J. Peng, and Z. Zhang, "Citygaussian: Real-time high-quality large-scale scene rendering with gaussians," in European Conference on Computer Vision. Springer, 2025, pp. 265–282.

[5] S. Shi, J. Cui, Z. Jiang, Z. Yan, G. Xing, J. Niu, and Z. Ouyang, "Vips: Real-time perception fusion for infrastructure-assisted autonomous driving," in Proceedings of the 28th annual international conference on mobile computing and networking, 2022, pp. 133–146.

[6] P. Tong, M. Li, M. Li, J. Huang, and X. Hua, "Large-scale vehicle trajectory reconstruction with camera sensing network," in Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, 2021, pp. 188–200.

[7] M. Intelligence, "Internet of cars market," https://www.mordorintelligence.com/market-analysis/automotive, 2024, accessed on 2025-01-14.

[8] ASKCI, "Research report on market prospect forecast of china's connected vehicle industry in 2024," https://www.askci.com/news/chanye/20240906/083719272558303902028264_3.shtml, 2024, accessed on 2025-01-14.

[9] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." ACM Trans. Graph., vol. 42, no. 4, pp. 139–1, 2023.

[10] Y. Liu, C. Luo, Z. Mao, J. Peng, and Z. Zhang, "Citygaussianv2: Efficient and geometrically accurate reconstruction for large-scale scenes," arXiv preprint arXiv:2411.00771, 2024.

[11] J. Lin, Z. Li, X. Tang, J. Liu, S. Liu, J. Liu, Y. Lu, X. Wu, S. Xu, Y. Yan et al., "Vastgaussian: Vast 3d gaussians for large scene reconstruction," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 5166–5175.

[12] Y. Chen and G. H. Lee, "Dogaussian: Distributed-oriented gaussian splatting for large-scale 3d reconstruction via gaussian consensus," arXiv preprint arXiv:2405.13943, 2024.

[13] J. Konečnỳ, "Federated learning: Strategies for improving communication efficiency," arXiv preprint arXiv:1610.05492, 2016.

[14] T. Zheng, A. Li, Z. Chen, H. Wang, and J. Luo, "Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving," in Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023, pp. 1–15.

[15] X. Yan, H. Zhang, Y. Cai, J. Guo, W. Qiu, B. Gao, K. Zhou, Y. Zhao, H. Jin, J. Gao et al., "Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities," arXiv preprint arXiv:2401.08045, 2024.

[16] D. Maltoni and V. Lomonaco, "Continuous learning in single-incremental-task scenarios," Neural Networks, vol. 116, pp. 56–73, 2019.

[17] M. Khani, P. Hamadanian, A. Nasr-Esfahany, and M. Alizadeh, "Real-time video inference on edge devices via adaptive model streaming," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4572–4582.

[18] R. Bhardwaj, Z. Xia, G. Ananthanarayanan, J. Jiang, Y. Shu, N. Karianakis, K. Hsieh, P. Bahl, and I. Stoica, "Ekya: Continuous learning of video analytics models on edge compute servers," in 19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22), 2022, pp. 119–135.

[19] M. Khani, G. Ananthanarayanan, K. Hsieh, J. Jiang, R. Netravali, Y. Shu, M. Alizadeh, and V. Bahl, "{RECL}: Responsive {Resource-Efficient} continuous learning for video analytics," in 20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23), 2023, pp. 917–932.

[20] Y. Kong, P. Yang, and Y. Cheng, "Edge-assisted on-device model update for video analytics in adverse environments," in Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 9051–9060.

[21] X. Ren, Y. Lu, H. Liang, Z. Wu, H. Ling, M. Chen, S. Fidler, F. Williams, and J. Huang, "Scube: Instant large-scale scene reconstruction using voxsplats," arXiv preprint arXiv:2410.20030, 2024.

[22] Y. Lu, X. Ren, J. Yang, T. Shen, Z. Wu, J. Gao, Y. Wang, S. Chen, M. Chen, S. Fidler et al., "Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models," arXiv preprint arXiv:2412.03934, 2024.

[23] H. Li, M. Yuan, Y. Zhang, C. Wu, C. Zhao, C. Song, H. Feng, E. Ding, D. Zhang, and J. Wang, "Xld: a cross-lane dataset for benchmarking novel driving view synthesis," arXiv preprint arXiv:2406.18360, 2024.

[24] H. Li, Y. Gao, D. Zhang, C. Wu, Y. Dai, C. Zhao, H. Feng, E. Ding, J. Wang, and J. Han, "Ggrt: Towards generalizable 3d gaussians without pose priors in real-time," arXiv preprint arXiv:2403.10147, 2024.

[25] M. Schwall, T. Daniel, T. Victor, F. Favaro, and H. Hohnhold, "Waymo public road safety performance data," arXiv preprint arXiv:2011.00038, 2020.

[26] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine et al., "Scalability in perception for autonomous driving: Waymo open dataset," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.

[27] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," arXiv preprint arXiv:2309.17080, 2023.

[28] Y. He, L. Ma, Z. Jiang, Y. Tang, and G. Xing, "Vi-eye: semantic-based 3d point cloud registration for infrastructure-assisted autonomous driving," in Proceedings of the 27th Annual International Conference on Mobile Computing and Networking, 2021, pp. 573–586.

[29] J. Cui, S. Shi, Y. He, J. Niu, G. Xing, and Z. Ouyang, "{VILAM}: Infrastructure-assisted 3d visual localization and mapping for autonomous driving," in 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24), 2024, pp. 1831–1845.

[30] S. Shi, N. Ling, Z. Jiang, X. Huang, Y. He, X. Zhao, B. Yang, C. Bian, J. Xia, Z. Yan et al., "Soar: Design and deployment of a smart roadside infrastructure system for autonomous driving," in Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, 2024, pp. 139–154.

[31] X. Gao, X. Zhang, Y. Lu, Y. Huang, L. Yang, Y. Xiong, and P. Liu, "A survey of collaborative perception in intelligent vehicles at intersections," IEEE Transactions on Intelligent Vehicles, pp. 1–20, 2024.

[32] B. Gao, J. Liu, H. Zou, J. Chen, L. He, and K. Li, "Vehicle-road-cloud collaborative perception framework and key technologies: A review," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 12, pp. 19 295–19 318, 2024.

[33] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, and C. Feng, "V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," IEEE Robotics and Automation Letters, vol. 7, no. 4, pp. 10 914–10 921, 2022.

[34] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song et al., "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13 712–13 722.

[35] T. Huang, J. Liu, X. Zhou, D. C. Nguyen, M. R. Azghadi, Y. Xia, Q.-L. Han, and S. Sun, "V2x cooperative perception for autonomous driving: Recent advances and challenges," arXiv preprint arXiv:2310.03525, 2023.

[36] H. Qiu, P. Huang, N. Asavisanu, X. Liu, K. Psounis, and R. Govindan, "Autocast: Scalable infrastructure-less cooperative perception for distributed collaborative driving," arXiv preprint arXiv:2112.14947, 2021.

[37] L. Song, W. Valentine, Q. Yang, H. Wang, H. Fang, and Y. Liu, "Bb-align: A lightweight pose recovery framework for vehicle-to-vehicle cooperative perception," in 2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2024, pp. 1016–1026.

[38] Q. Zhang, X. Zhang, R. Zhu, F. Bai, M. Naserian, and Z. M. Mao, "Robust real-time multi-vehicle collaboration on asynchronous sensors," in Proceedings of the 29th Annual International Conference on Mobile Computing and Networking, 2023, pp. 1–15.

[39] R. Zhang, D. Meng, S. Shen, Z. Zou, H. Li, and H. X. Liu, "Msight: An edge-cloud infrastructure-based perception system for connected automated vehicles," arXiv preprint arXiv:2310.05290, 2023.

[40] R. Yu, D. Yang, and H. Zhang, "Edge-assisted collaborative perception in autonomous driving: A reflection on communication design," in 2021 IEEE/ACM Symposium on Edge Computing (SEC). IEEE, 2021, pp. 371–375.

[41] Y. Cheng, P. Yang, N. Zhang, and J. Hou, "Edge-assisted lightweight region-of-interest extraction and transmission for vehicle perception," in GLOBECOM 2023-2023 IEEE Global Communications Conference. IEEE, 2023, pp. 1054–1059.

[42] M. Hanyao, Y. Jin, Z. Qian, S. Zhang, and S. Lu, "Edge-assisted online on-device object detection for real-time video analytics," in IEEE INFOCOM 2021-IEEE Conference on Computer Communications. IEEE, 2021, pp. 1–10.

[43] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," Communications of the ACM, vol. 54, no. 10, pp. 105–112, 2011.

[44] C. Früh and A. Zakhor, "An automated method for large-scale, ground-based city model acquisition," International Journal of Computer Vision, vol. 60, pp. 5–24, 2004.

[45] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4104–4113.

[46] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[47] M. Zhenxing and D. Xu, "Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields," in The Eleventh International Conference on Learning Representations, 2022.

[48] Y. Zhang, G. Chen, and S. Cui, "Efficient large-scale scene representation with a hybrid of high-resolution grid and plane features," arXiv preprint arXiv:2303.03003, 2023.

[49] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12 922–12 931.

[50] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 8248–8258.

[51] L. Xu, Y. Xiangli, S. Peng, X. Pan, N. Zhao, C. Theobalt, B. Dai, and D. Lin, "Grid-guided neural radiance fields for large urban scenes," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8296–8306.

[52] T. Suzuki, "Fed3dgs: Scalable 3d gaussian splatting with federated learning," arXiv preprint arXiv:2403.11460, 2024.

[53] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Client-edge-cloud hierarchical federated learning," in ICC 2020-2020 IEEE international conference on communications (ICC). IEEE, 2020, pp. 1–6.

[54] Y. Cui, K. Cao, J. Zhou, and T. Wei, "Optimizing training efficiency and cost of hierarchical federated learning in heterogeneous mobile-edge cloud computing," IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, vol. 42, no. 5, pp. 1518–1531, 2022.

[55] Y. Deng, F. Lyu, J. Ren, Y. Zhang, Y. Zhou, Y. Zhang, and Y. Yang, "Share: Shaping data distribution at edge for communication-efficient hierarchical federated learning," in 2021 IEEE 41st International Conference on Distributed Computing Systems (ICDCS). IEEE, 2021, pp. 24–34.

[56] Z. Wang, H. Xu, J. Liu, H. Huang, C. Qiao, and Y. Zhao, "Resource-efficient federated learning with hierarchical aggregation in edge computing," in IEEE INFOCOM 2021-IEEE conference on computer communications. IEEE, 2021, pp. 1–10.

[57] Z. Li, Y. He, H. Yu, J. Kang, X. Li, Z. Xu, and D. Niyato, "Data heterogeneity-robust federated learning via group client selection in industrial iot," IEEE Internet of Things Journal, vol. 9, no. 18, pp. 17 844–17 857, 2022.

[58] N. Mhaisen, A. A. Abdellatif, A. Mohamed, A. Erbad, and M. Guizani, "Optimal user-edge assignment in hierarchical federated learning based on statistical properties and network topology constraints," IEEE Transactions on Network Science and Engineering, vol. 9, no. 1, pp. 55–66, 2021.

[59] Z. Zhong, W. Bao, J. Wang, X. Zhu, and X. Zhang, "Flee: A hierarchical federated learning framework for distributed deep neural network over cloud, edge, and end device," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 13, no. 5, pp. 1–24, 2022.

[60] Y. Deng, J. Ren, C. Tang, F. Lyu, Y. Liu, and Y. Zhang, "A hierarchical knowledge transfer framework for heterogeneous federated learning," in IEEE INFOCOM 2023-IEEE Conference on Computer Communications. IEEE, 2023, pp. 1–10.

[61] Z. Yang, S. Fu, W. Bao, D. Yuan, and A. Y. Zomaya, "Hierarchical federated learning with momentum acceleration in multi-tier networks," IEEE Transactions on Parallel and Distributed Systems, 2023.

[62] E. Chen, S. Wang, and C. G. Brinton, "Taming subnet-drift in d2d-enabled fog learning: A hierarchical gradient tracking approach," in IEEE INFOCOM 2024-IEEE conference on computer communications, 2024, accepted.

[63] A. A. Abdellatif, N. Mhaisen, A. Mohamed, A. Erbad, M. Guizani, Z. Dawy, and W. Nasreddine, "Communication-efficient hierarchical federated learning for iot heterogeneous systems with imbalanced data," Future Generation Computer Systems, vol. 128, pp. 406–419, 2022.

[64] H. Gao, Z. Wang, Y. Li, K. Long, M. Yang, and Y. Shen, "A survey for foundation models in autonomous driving," arXiv preprint arXiv:2402.01105, 2024.

[65] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: theory, method and application," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[66] G. Hinton, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.

[67] K. Shmelkov, C. Schmid, and K. Alahari, "Incremental learning of object detectors without catastrophic forgetting," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 3400–3409.

[68] R. T. Mullapudi, S. Chen, K. Zhang, D. Ramanan, and K. Fatahalian, "Online model distillation for efficient video inference," in Proceedings of the IEEE/CVF International conference on computer vision, 2019, pp. 3573–3582.

[69] J. Lu, Y. Zhang, Q. Shen, X. Wang, and S. Yan, "Poison-splat: Computation cost attack on 3d gaussian splatting," arXiv preprint arXiv:2410.08190, 2024.

[70] D. Gao, X. Yao, and Q. Yang, "A survey on heterogeneous federated learning," arXiv preprint arXiv:2210.04505, 2022.

[71] M. Ye, X. Fang, B. Du, P. C. Yuen, and D. Tao, "Heterogeneous federated learning: State-of-the-art and research challenges," ACM Computing Surveys, vol. 56, no. 3, pp. 1–44, 2023.

[72] L. Shen, Q. Yang, K. Cui, Y. Zheng, X.-Y. Wei, J. Liu, and J. Han, "Fedconv: A learning-on-model paradigm for heterogeneous federated clients," in Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services, 2024, pp. 398–411.

[73] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," Advances in neural information processing systems, vol. 30, 2017.

[74] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[75] 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 international conference on computer vision, 2021, pp. 10 012–10 022.

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

[77] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, and Y. Qiao, "Vision transformer adapter for dense predictions," arXiv preprint arXiv:2205.08534, 2022.

[78] F. Aurenhammer, "Voronoi diagrams—a survey of a fundamental geometric data structure," ACM Computing Surveys (CSUR), vol. 23, no. 3, pp. 345–405, 1991.

[79] L. Lin, Y. Liu, Y. Hu, X. Yan, K. Xie, and H. Huang, "Capturing, reconstructing, and simulating: the urbanscene3d dataset," in European Conference on Computer Vision. Springer, 2022, pp. 93–109.

[80] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE transactions on image processing, vol. 13, no. 4, pp. 600–612, 2004.

[81] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 586–595.

[82] S. Ren, "Faster r-cnn: Towards real-time object detection with region proposal networks," arXiv preprint arXiv:1506.01497, 2015.

[83] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 20 697–20 709.

[84] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," The International Journal of Robotics Research, vol. 32, no. 11, pp. 1231–1237, 2013.

[85] H. Peng, Y. Zhan, P. Li, and Y. Xia, "Tangram: High-resolution video analytics on serverless platform with slo-aware batching," arXiv preprint arXiv:2404.09267, 2024.

[86] Y. Li, A. Padmanabhan, P. Zhao, Y. Wang, G. H. Xu, and R. Netravali, "Reducto: On-camera filtering for resource-efficient real-time video analytics," in Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication, 2020, pp. 359–376.

[87] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," arXiv preprint arXiv:1701.06538, 2017.

[88] C. Riquelme, J. Puigcerver, B. Mustafa, M. Neumann, R. Jenatton, A. Susano Pinto, D. Keysers, and N. Houlsby, "Scaling vision with sparse mixture of experts," Advances in Neural Information Processing Systems, vol. 34, pp. 8583–8595, 2021.

[89] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[90] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, T. Darrell et al., "Bdd100k: A diverse driving video database with scalable annotation tooling," arXiv preprint arXiv:1805.04687, vol. 2, no. 5, p. 6, 2018.

[91] Alibaba, "Alibaba cloud function compute," 2023, https://www.alibabacloud.com/product/function-compute, Last accessed on 2023-6-13.