AutoFed: Heterogeneity-Aware Federated Multimodal Learning for Robust Autonomous Driving

Tianyue Zheng[1], Ang Li[2], Zhe Chen[3], Hongbo Wang[1], **Jun Luo**[1]

[1] School of Computer Science and Engineering, Nanyang Technological University, Singapore

[2] Department of Electrical and Computer Engineering, University of Maryland

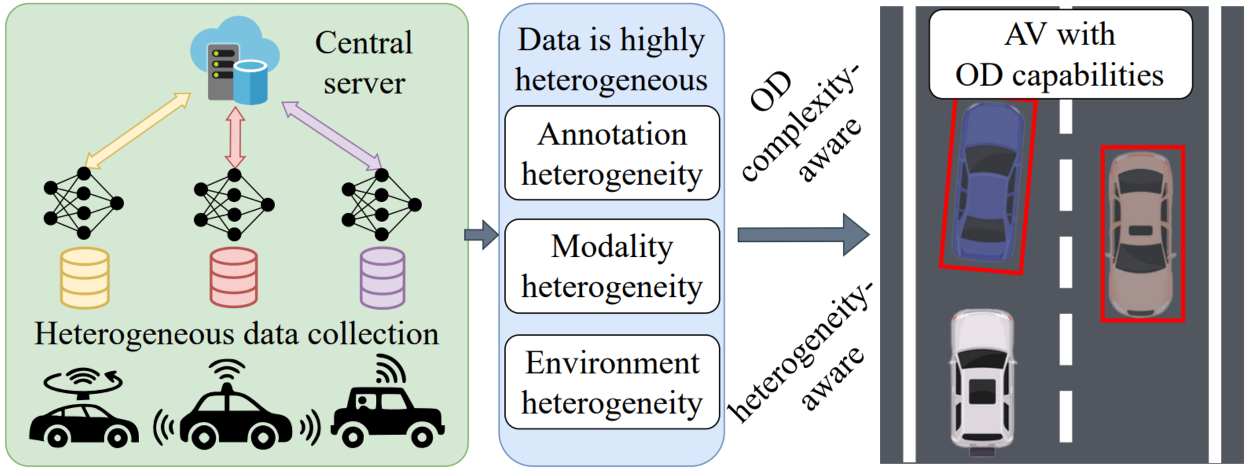
[3] Intelligent Networking and Computing Research Center and School of Computer Science

October, 2023

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

Introduction

- Two-stage object detection (OD).
- Bird's-eye view reconciles view discrepancies among different sensing modalities.
- Exploit crowdsensing to outsource data collection and annotation tasks to autonomous vehicles (AVs).
- Integrating federated learning (FL) into crowdsensing.

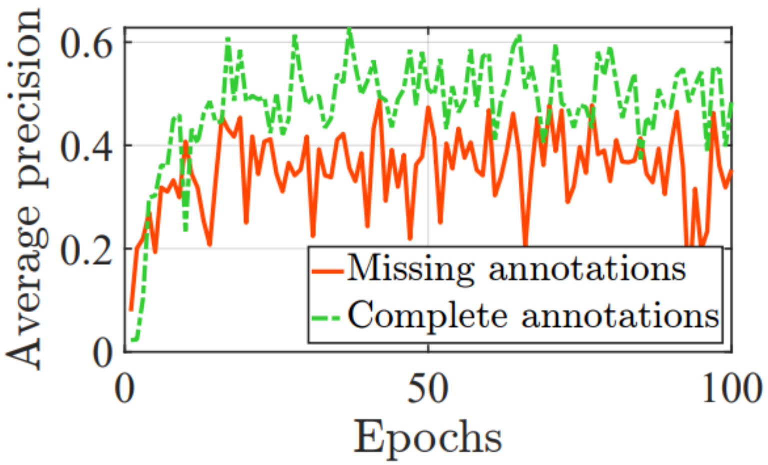


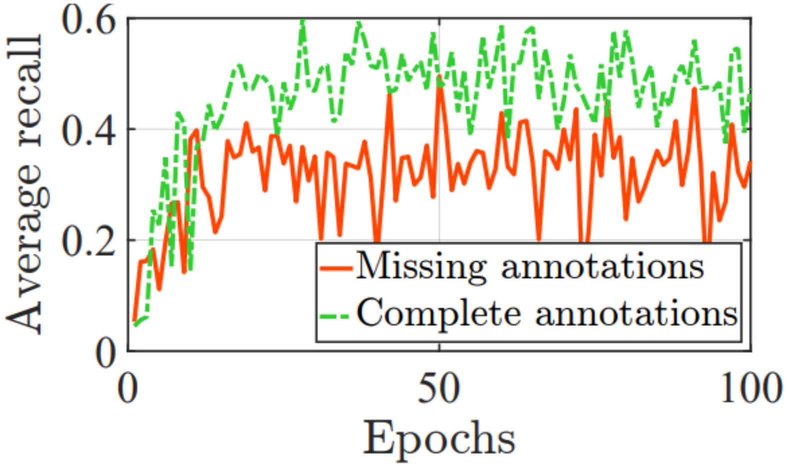
The bird's-eye view FL-OD of AutoFed.

Challenges

- Annotation heterogeneity
- Modality heterogeneity
- Environment heterogeneity

Motivation and Background



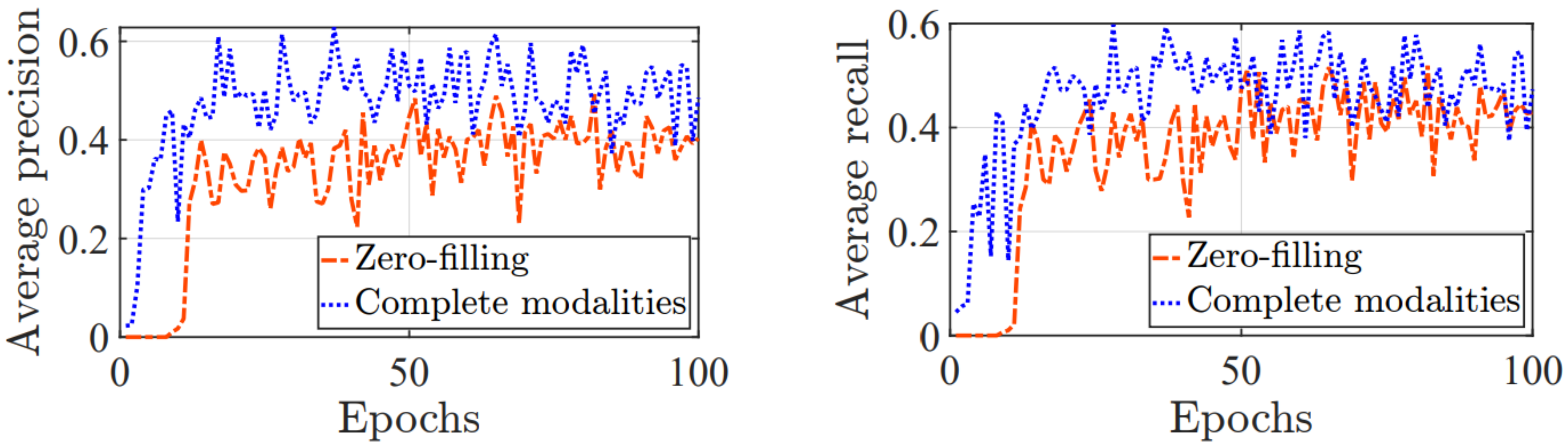


Damaging effects of missing annotations.

- Some clients may be more motivated to provide annotation with adequate quality.
- Others may be busy and/or less skillful that they miss a large proportion of the proportions.
- The network under complete labeling outperforms that under missing labeling.
- The performance of the DNN under missing labeling experiences a downward trend after the 20-th epoch.

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

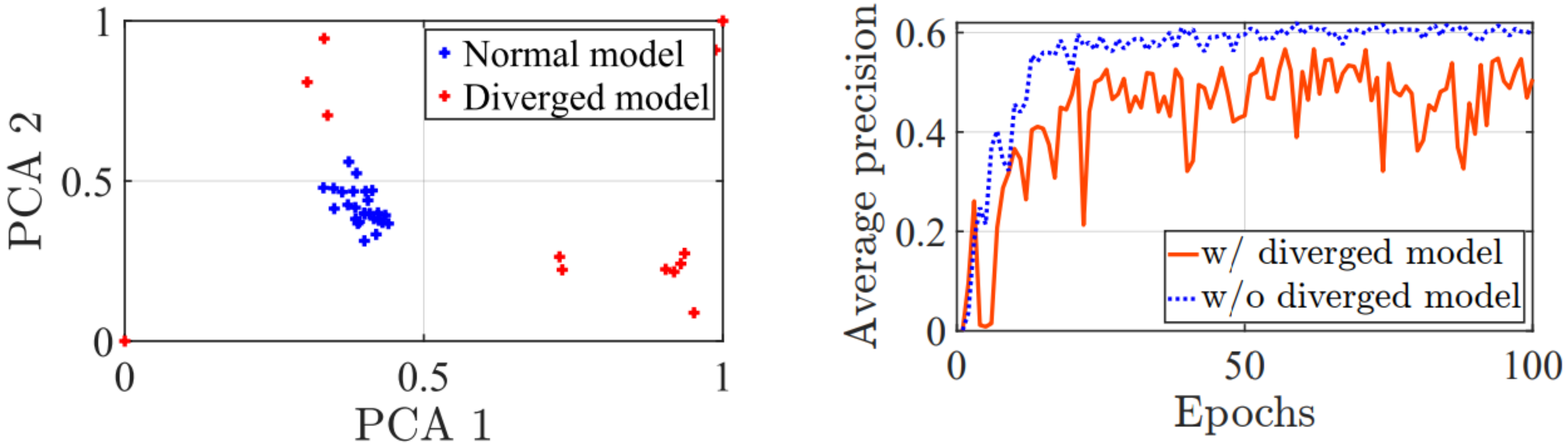
Motivation and Background



Damaging effects of missing modality.

- AVs may be equipped with different types of sensors by their manufacturer.
- Sensor may experience malfunctions.
- Removing data entries or only keeping shared modalities among the clients discard useful information.
- Zero-filling does not fully overcome the challenge (lacking access to global statistics).

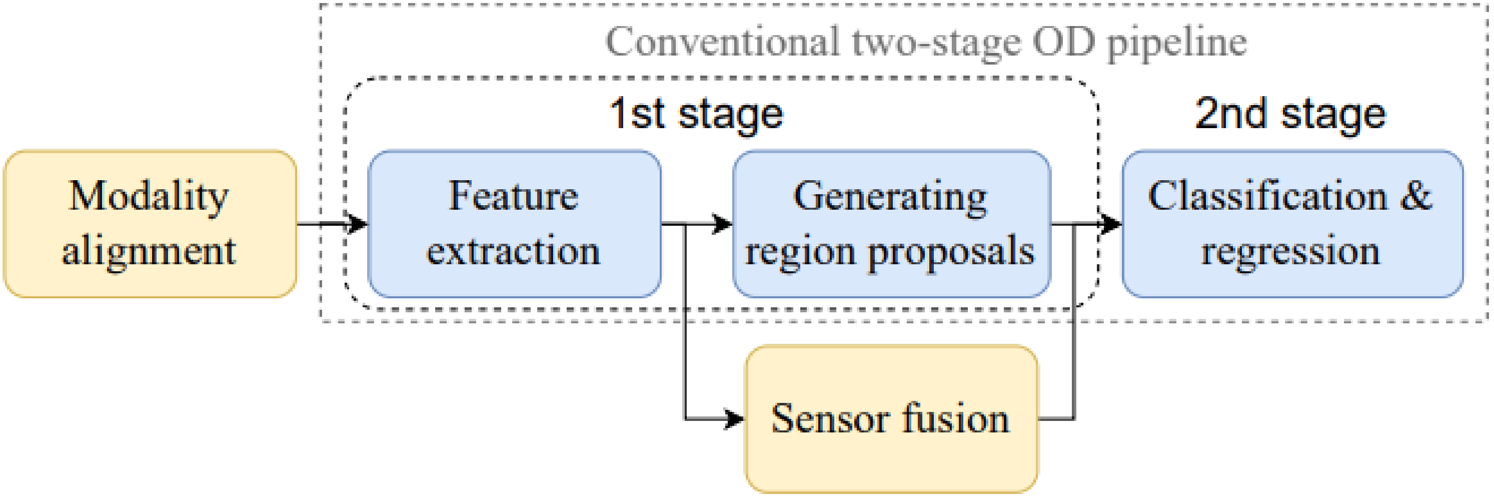
Motivation and Background



Damaging effects of diverging models.

- Heterogeneities introduced by environments (e.g., different weather and road condition).
- Local models on AVs to be diverged.
- The optimization goal can even become contradictory.

System Design: OD Pipeline

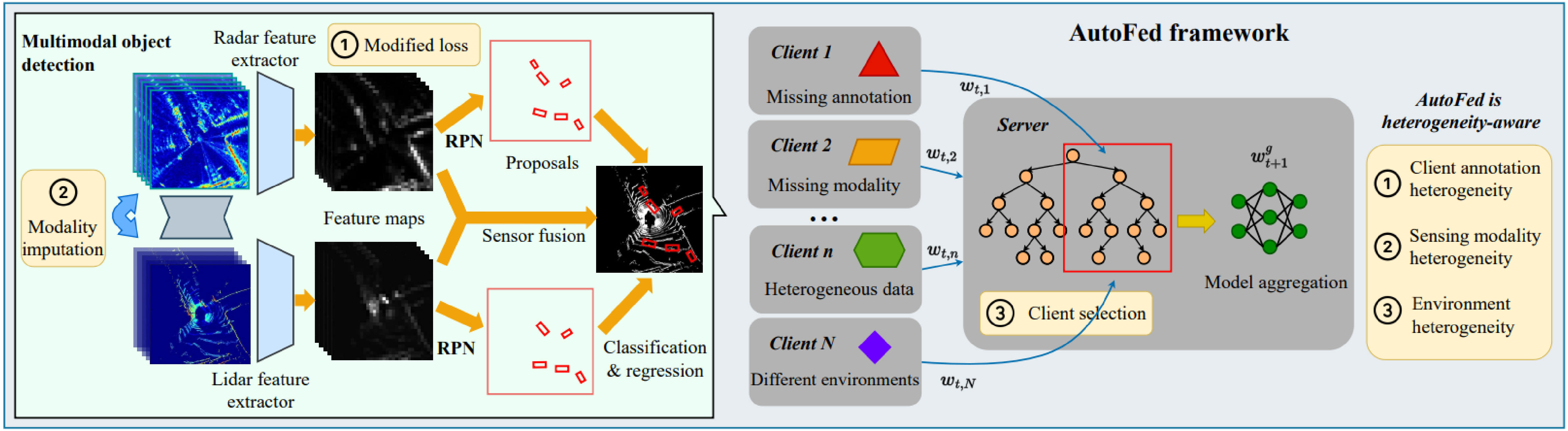

Conventional two-stage OD pipeline
1st stage
2nd stage
Modality alignment
Feature extraction
Generating region proposals
Classification & regression
Sensor fusion

The upgraded OD pipeline of AutoFed's multimodal vehicle detection network.

- A feature map is first extracted using well-accepted feature extractors (e.g., VGGNet or ResNet).
- Region proposal are generated by the region proposal network (RPN). Region proposals filtered by non-maximum suppression (NMS).
- Performs fine-tuning to jointly optimize a classifier and bounding-box regressors.

$$L_{\text{total}} = L^{\text{RPN}} + L_{\text{cls}} + L_{\text{reg}} + L_{\text{dir}}$$

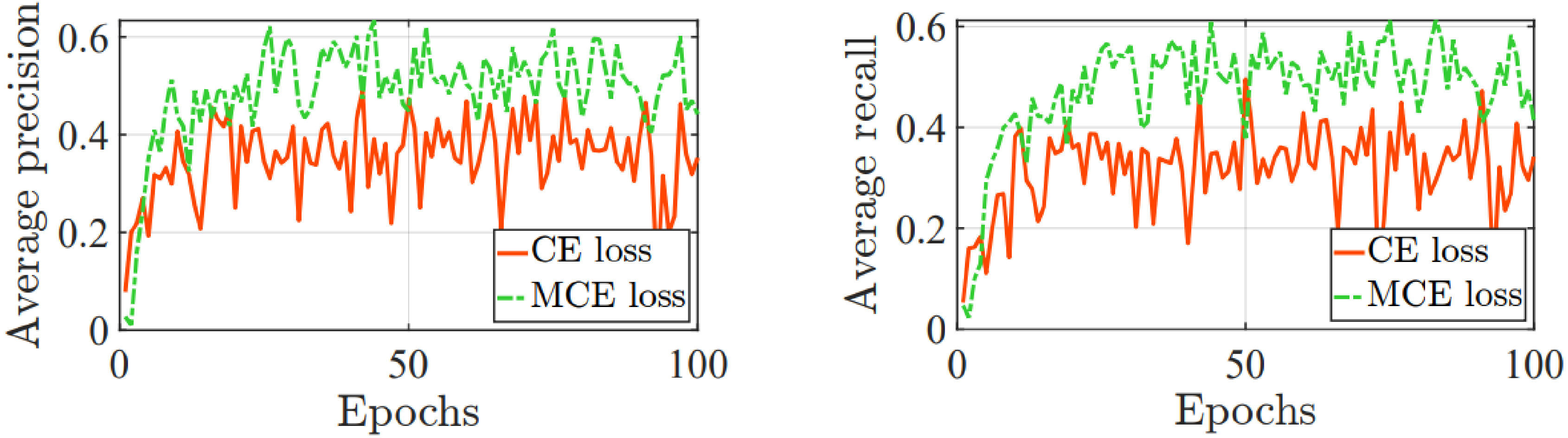
System Design: AutoFed Framework



AutoFed architecture: Federated multimodal learning with heterogeneity-awareness.

System Design: Modified Loss Function

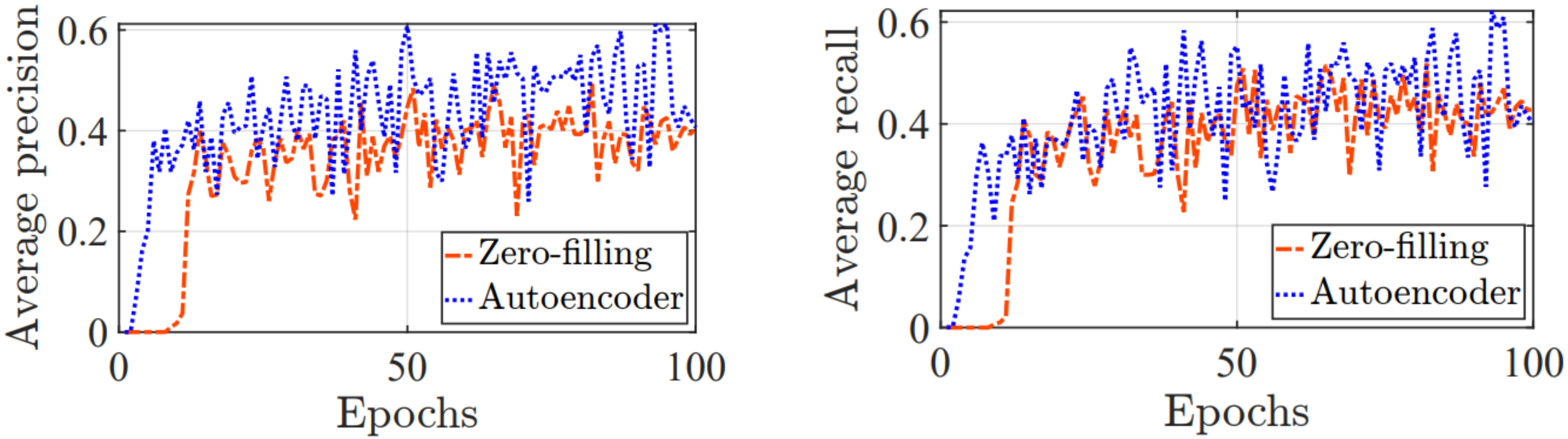
$$\begin{cases} 0, & p > p_{\text{th}} \text{ and } p^* = 0, \\ -p^* \log p - (1 - p^*) \log(1 - p), & \text{otherwise,} \end{cases}$$



Comparison between CE and MCE loss.

- Identify vehicles wrongly labeled as backgrounds according to its own well-established classifier.
- Avoid sending erroneous gradient signals during backpropagation.
- Better guiding the convergence on the OD loss surface.
- Average precision of vehicle detection is 0.57 and 0.4 when CE and MCE loss are used.
- A gap greater than 0.1 in the average recalls when the two losses are used.
- MCE quickly overtakes CE loss, keeps an upward trend and converges faster.

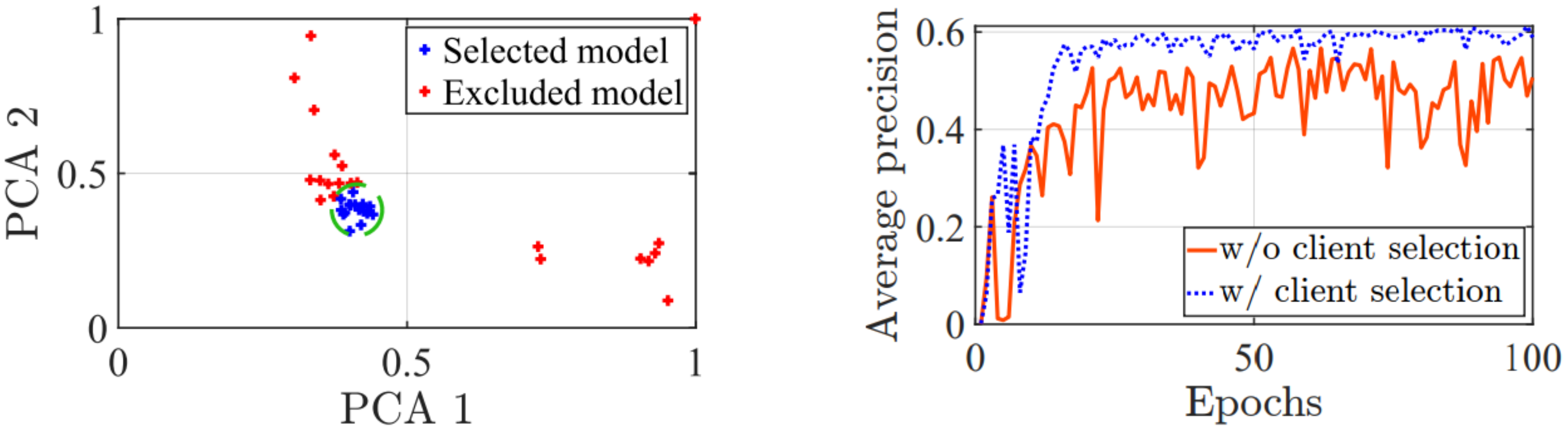
System Design: Modality Imputation



Modality imputation with an autoencoder.

- Employ a convolutional autoencoder with residual connections that facilitates information flow.
- The lightweight architecture of the autoencoder only incurs negligible overhead.
- Zero-filling only achieves an average precision of approximately 0.4, lower than an average precision of about 0.5 achieved by autoencoder imputation.
- Similarly, autoencoder imputation also surpasses zero-filling in terms of average recall by a discernible margin.

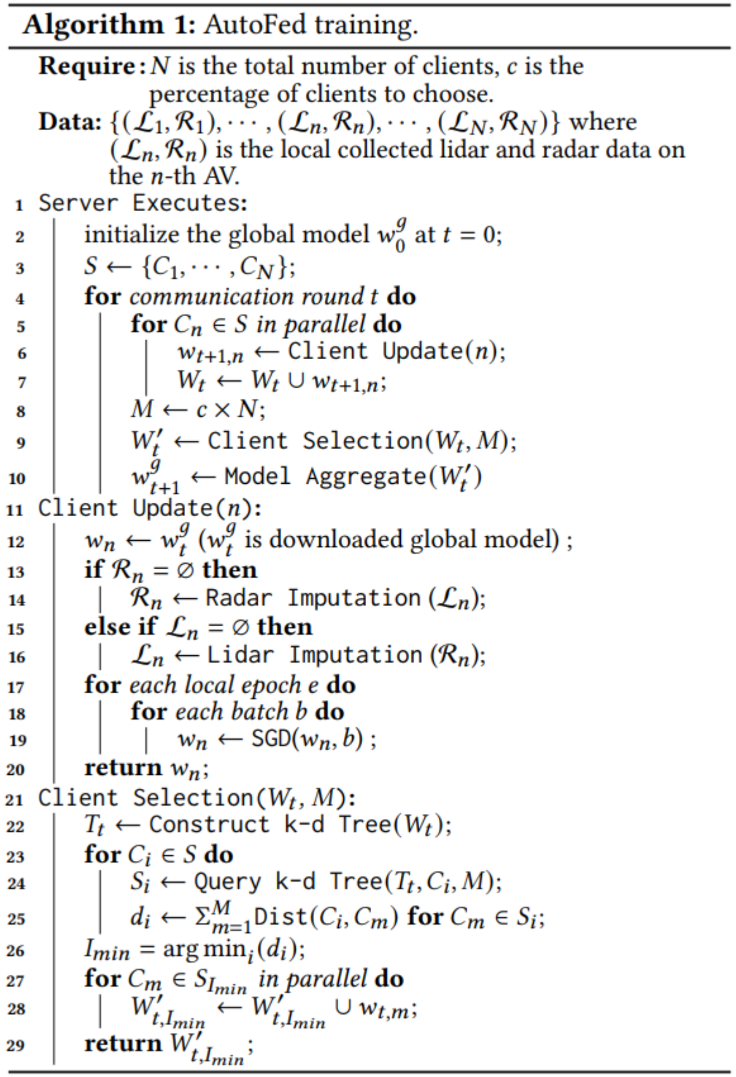
System Design: Client Selection



Client selection mitigates diverged models.

- Environment heterogeneities (weather and road conditions) cause serious model divergence among the clients.
- Chaotic loss surface can disorient the gradient descent algorithm used for training the OD model.
- Devise a novel client selection strategy (*k*-d tree-based) immune to divergence.
- The precision of vehicle detection reaches up to 0.6 when client selection is enabled, and it fluctuates around 0.5 when model weights from all clients are aggregated using the FedAvg algorithm.

System Design: Putting It All Together



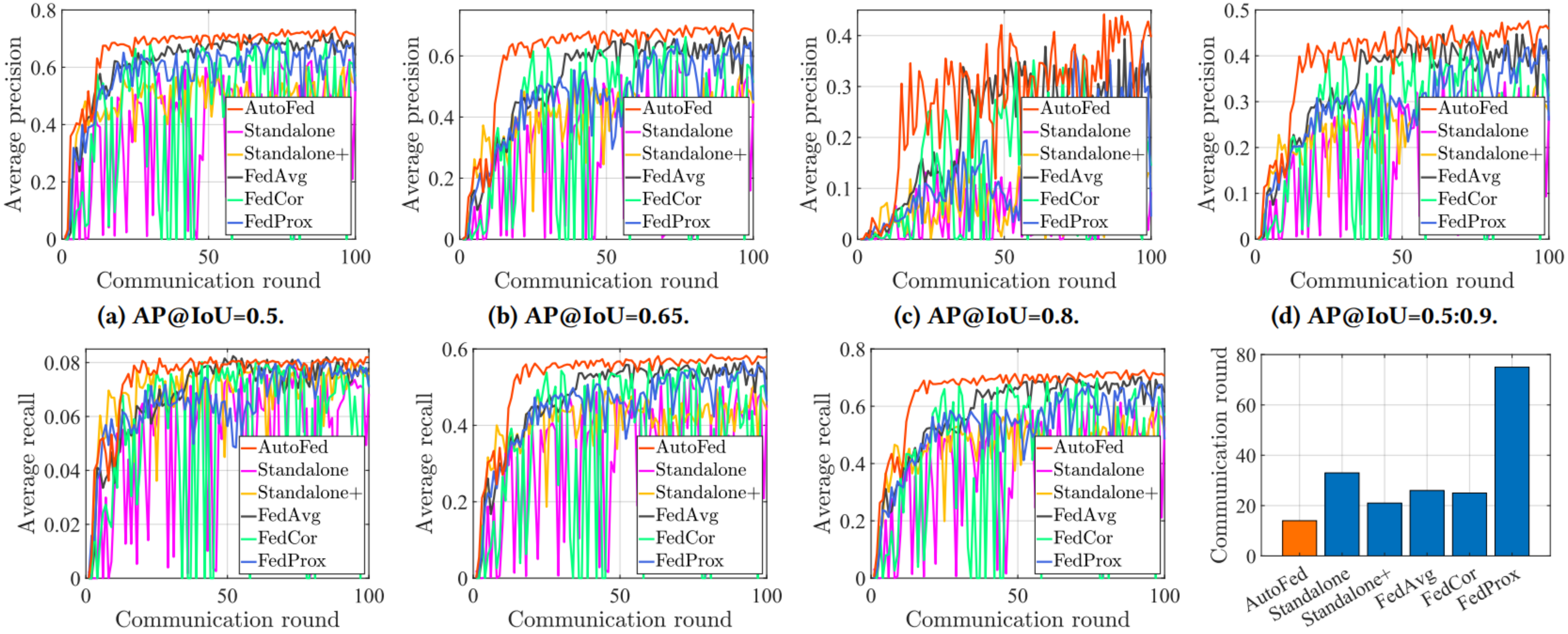
Algorithm 1: AutoFed training.

```
Require: $N$ is the total number of clients, $c$ is the percentage of clients to choose.
Data: $\{(\mathcal{L}_1, \mathcal{R}_1), \cdots, (\mathcal{L}_n, \mathcal{R}_n), \cdots, (\mathcal{L}_N, \mathcal{R}_N)\}$ where $(\mathcal{L}_n, \mathcal{R}_n)$ is the local collected lidar and radar data on the $n$-th AV.
Server Executes:
    initialize the global model $w_0^g$ at $t = 0$;
    $S \leftarrow \{C_1, \cdots, C_N\}$;
    for communication round $t$ do
        for $C_n \in S$ in parallel do
            $w_{t+1,n} \leftarrow$ Client Update($n$);
            $W_t \leftarrow W_t \cup w_{t+1,n}$;
        $M \leftarrow c \times N$;
        $W_t' \leftarrow$ Client Selection($W_t, M$);
        $w_{t+1}^g \leftarrow$ Model Aggregate($W_t'$)
Client Update($n$):
    $w_n \leftarrow w_t^g$ ($w_t^g$ is downloaded global model) ;
    if $\mathcal{R}_n = \emptyset$ then
        $\mathcal{R}_n \leftarrow$ Radar Imputation ($\mathcal{L}_n$);
    else if $\mathcal{L}_n = \emptyset$ then
        $\mathcal{L}_n \leftarrow$ Lidar Imputation ($\mathcal{R}_n$);
    for each local epoch $e$ do
        for each batch $b$ do
            $w_n \leftarrow$ SGD($w_n, b$) ;
    return $w_n$;
Client Selection($W_t, M$):
    $T_t \leftarrow$ Construct k-d Tree($W_t$);
    for $C_i \in S$ do
        $S_i \leftarrow$ Query k-d Tree($T_t, C_i, M$);
        $d_i \leftarrow \Sigma_{m=1}^{M}$Dist($C_i, C_m$) for $C_m \in S_i$;
    $I_{min} = \arg\min_i(d_i)$;
    for $C_m \in S_{I_{min}}$ in parallel do
        $W_{t,I_{min}}' \leftarrow W_{t,I_{min}}' \cup w_{t,m}$;
    return $W_{t,I_{min}}'$;
```

- `Client Update` is the local training process for each client.
- `Radar Imputation` and `Lidar Imputation` are imputation functions.
- `Client Selection` includes `Construct k-d Tree` and `Query k-d Tree` as processes of constructing and querying *k*-d tree.
- `Model Aggregate` is the standard process of averaging the selected local models.

Evaluation - I

- Standalone
- Standalone+
- FedAvg
- FedCor
- FedProx



(a) **AP@IoU=0.5.** (b) **AP@IoU=0.65.** (c) **AP@IoU=0.8.** (d) **AP@IoU=0.5:0.9.**

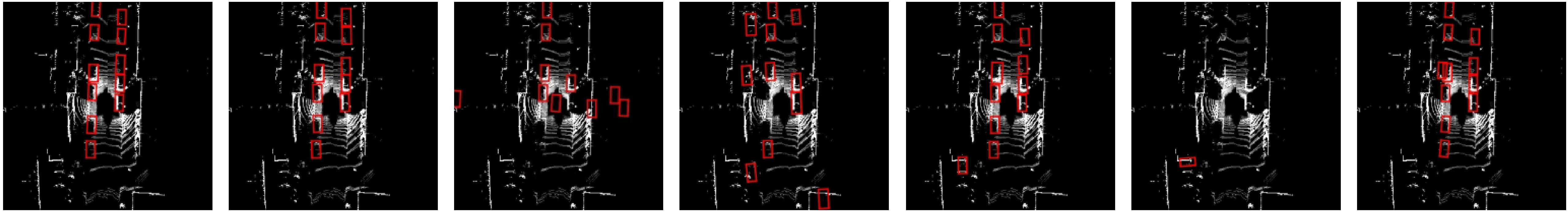
(e) **AR, maxDets = 1.** (f) **AR, maxDets = 10.** (g) **AR, maxDets = 100.** (h) **Convergence time.**

Comparing AutoFed with several baselines, in terms of FL convergence and communication overhead.

- Higher AP and AR
- Faster convergence
- Better stability

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

Evaluation - II



(a) Ground truth. (b) AutoFed. (c) Standalone. (d) Standalone+. (e) FedAvg. (f) FedCor. (g) FedProx.

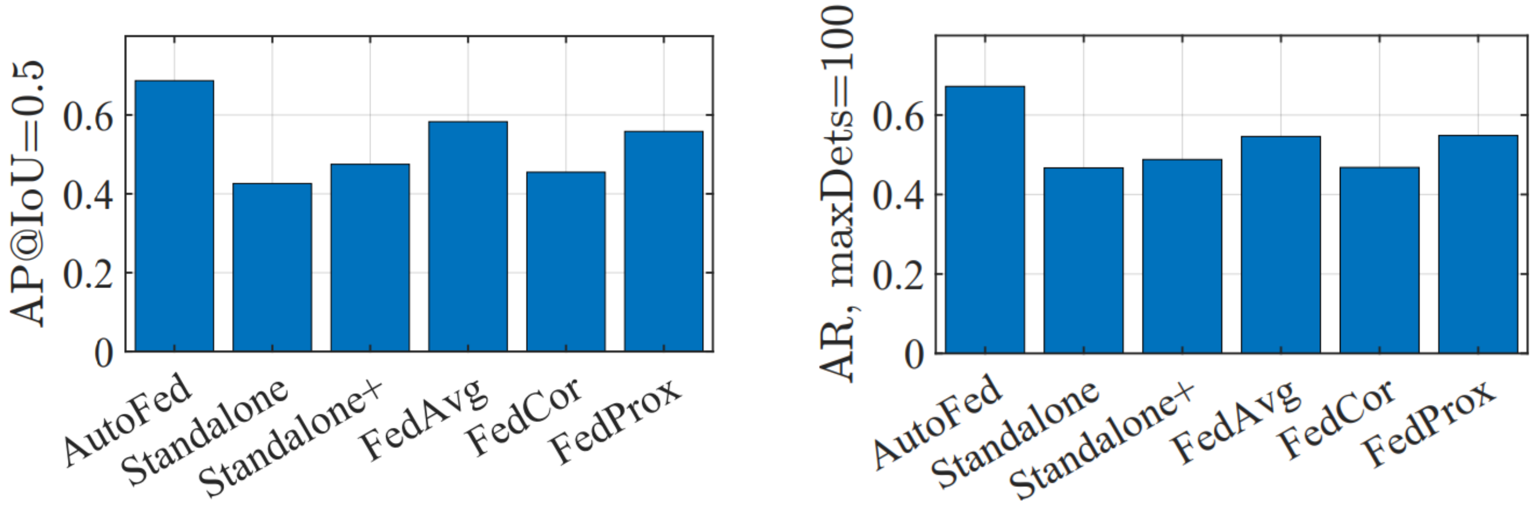
Example detection results of AutoFed and other baseline methods.

- AutoFed generates high-precision vehicle detection results.
- The baseline methods make incorrect prediction outside the road, miss most of the vehicles, and generate inaccurate bounding boxes.

Communication efficiency:

- While centralized training transfers 660000KB of sensor data during each communication round per client, AutoFed only transfers 62246KB of model weights.
- AutoFed reduces up to more than 10× communication cost per client than the centralized training, firmly validating its **communication-efficient**.

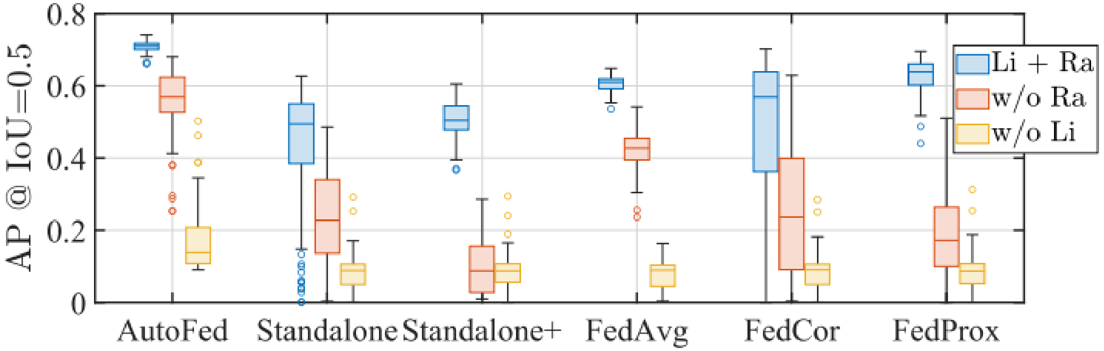
Evaluation - III

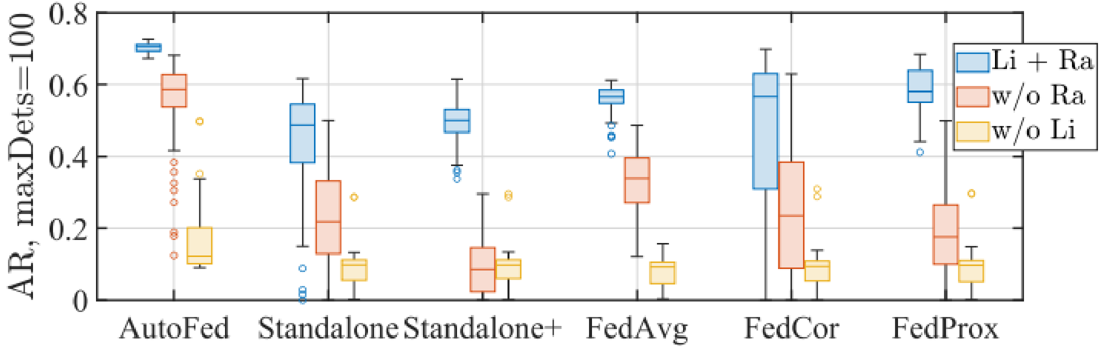


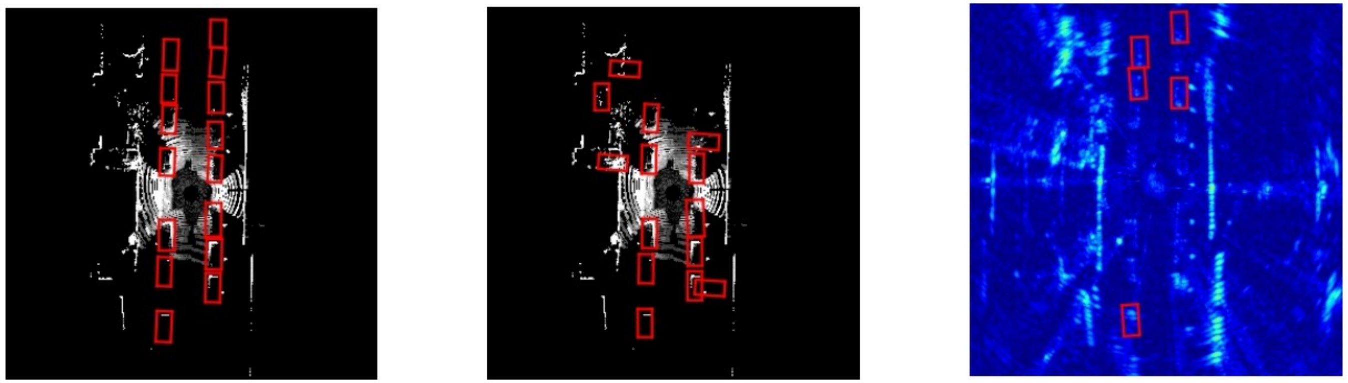
Evaluation on the nuScenes Dataset.

- AutoFed outperforms the baselines on the nuScenes dataset as well.
- Evaluation results are not specific to a single dataset (Oxford Radar RobotCar), but can generalize.
- Overall AP and AR results of AutoFed on this dataset (0.687 and 0.672) are slightly lower than the Oxford Radar RobotCar dataset (caused by the complexity of the scenes and objects, sensor mounting positions, and most importantly, the sparsity and lower quality of the radar point cloud provided by the nuScenes dataset).

Evaluation - IV







Lidar + radar.

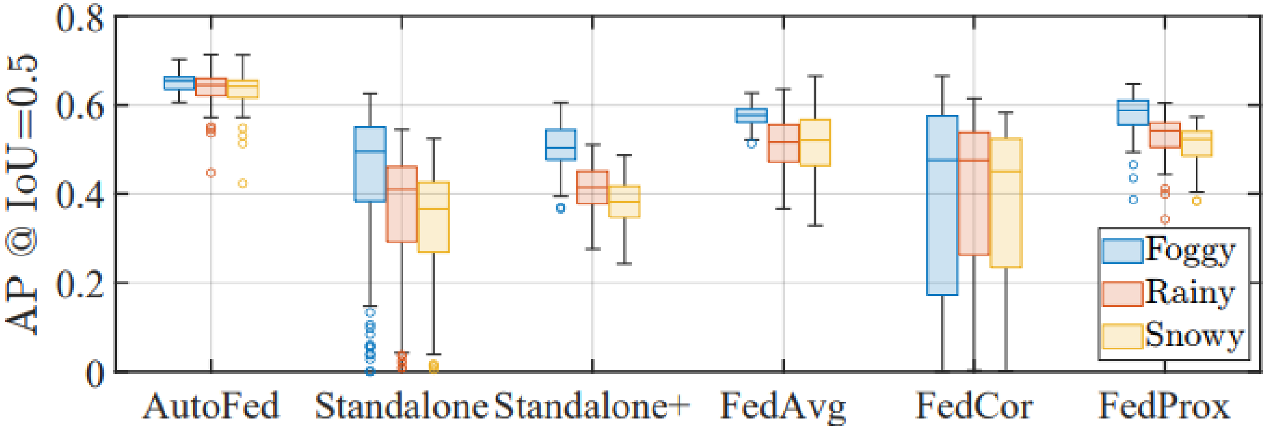
Missing radar.

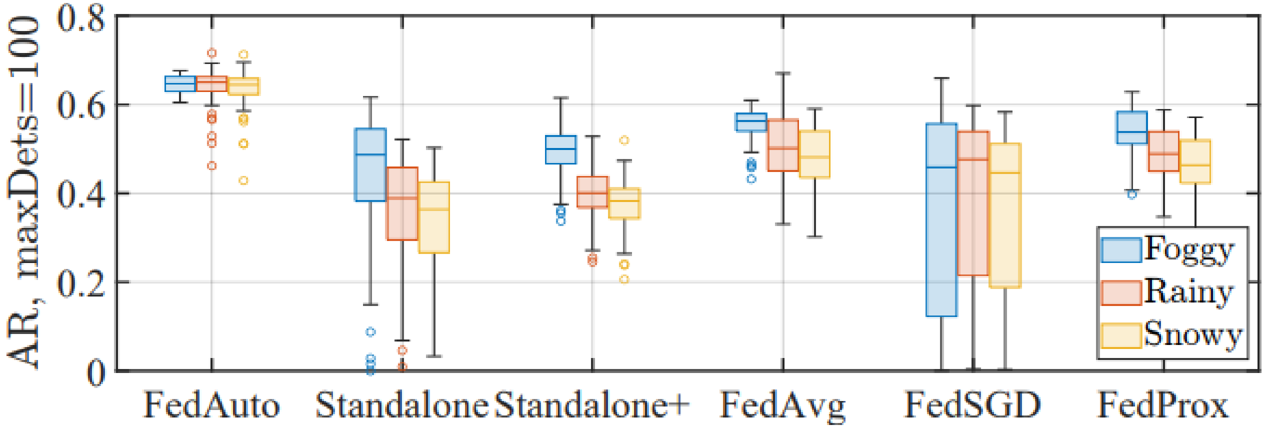
Missing lidar.

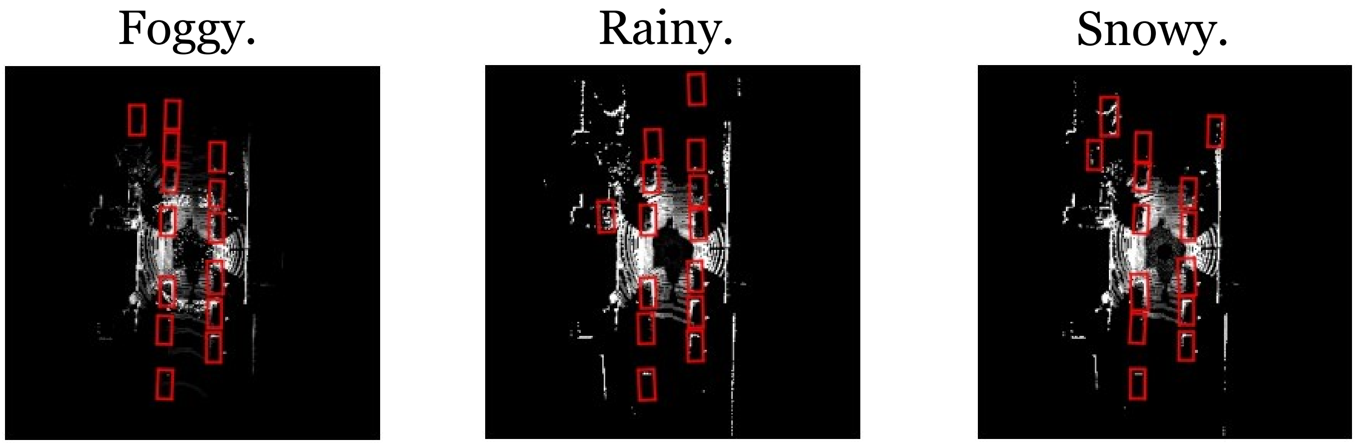
Different missing modalities.

- When both lidar and radar are available, AutoFed is able to recognize most of the vehicles on the road.
- Missing radar: detection of vehicles in the further distance is affected, but the nearby vehicles can still be identified.
- Missing lidar: the vehicles in distance can be well detected by the radar.

Evaluation - V





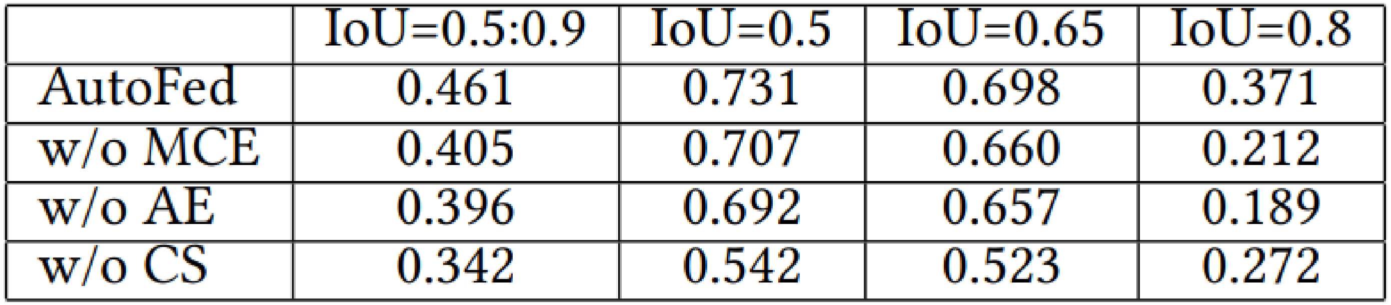


Different weathers.

- Employ physical model (DEF and LISA) to simulate fog, rain, and snow.
- Foggy weather attenuates lidar signal and shrinks the field of view.
- Rainy and snowy weathers mainly affect the lidar signal by inducing scattered reflection near the sensor.
- The three adverse weather conditions degrade the median AP of AutoFed from 0.71 to 0.65, 0.63, and 0.63, respectively.
- Median AR from 0.71 to 0.64, 0.63, and 0.63, respectively.

Evaluation - VI

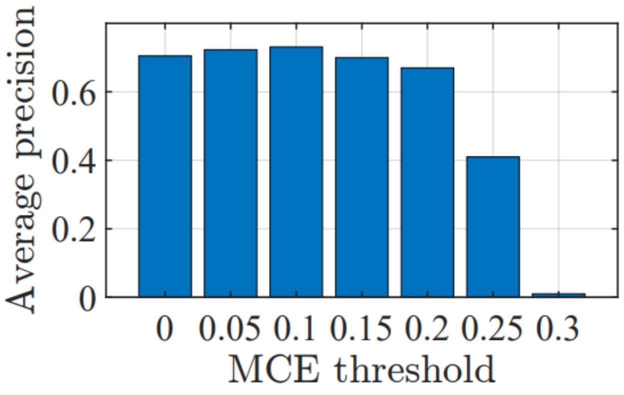
Effects of key AutoFed parts in terms of AP (ablation study).

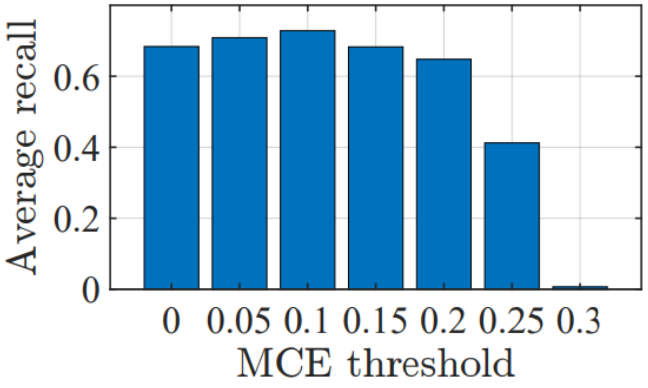


	IoU=0.5:0.9	IoU=0.5	IoU=0.65	IoU=0.8
AutoFed	0.461	0.731	0.698	0.371
w/o MCE	0.405	0.707	0.660	0.212
w/o AE	0.396	0.692	0.657	0.189
w/o CS	0.342	0.542	0.523	0.272

- Take the AP when IoU is above 0.5 as an example, AutoFed achieves an AP of 0.731, while AutoFed without MCE loss, modality imputation with autoencoder, and client selection obtain the AP of 0.707, 0.692, and 0.542, respectively.
- Both MCE loss and modality imputation are indispensable parts: although the lack of the two can be compensated by client selection (which excludes erroneous gradients) to a certain extent, there still are many heterogeneous scenarios that cannot be addressed by client selection alone.
- The integration of MCE loss and modality imputation, together with client selection, can act as “belt and braces” to guarantee the robustness of AutoFed in diversified heterogeneous scenarios.

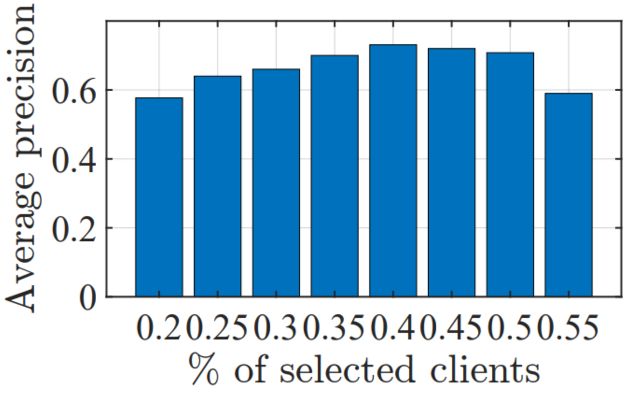
Evaluation - VII

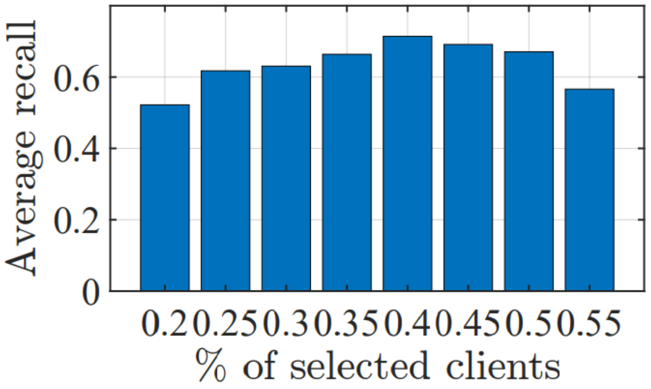




Impact of the MCE thresholds:

- When p_{th} is small, incorrect gradients induced by missing annotation cannot be excluded.
- Many real backgrounds can be mistakenly excluded if p_{th} is set too large.





Impact of the selected client percentage:

- A small percentage of selected clients could not fully utilize the diverse data collected by different clients and introduce bias into the federated model.
- If a very large proportion of the clients are selected, we cannot effectively mitigate the detrimental effect caused by diverged local models.

Thank you!

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE