FedDrive v2: an Analysis of the Impact of Label Skewness in Federated Semantic Segmentation for Autonomous Driving

2023 I-RIM Conference
October 20-22, Rome, Italy
ISBN: 978-88-945805-4-9
10.5281/zenodo.10722478

FedDrive v2: an Analysis of the Impact of Label Skewness in Federated Semantic Segmentation for Autonomous Driving

1st Eros Fanì
DAUIN Politecnico di Torino
Turin, Italy
eros.fani@polito.it

2nd Marco Ciccone
DAUIN Politecnico di Torino
Turin, Italy
marco.ciccone@polito.it

3rd Barbara Caputo
DAUIN Politecnico di Torino
Turin, Italy
barbara.caputo@polito.it

***Abstract*—We propose FedDrive v2, an extension of the Federated Learning benchmark for Semantic Segmentation in Autonomous Driving. While the first version aims at studying the effect of domain shift of the visual features across clients, in this work, we focus on the distribution skewness of the labels. We propose six new federated scenarios to investigate how label skewness affects the performance of segmentation models and compare it with the effect of domain shift. Finally, we study the impact of using the domain information during testing.**
Official website: https://feddrive.github.io

***Index Terms*—federated learning, semantic segmentation, autonomous driving, label skewness, domain shift, domain generalization**

I. Introduction and Related Works

An essential challenge for robust decision-making in autonomous vehicles such as self-driving cars is to design systems that can effectively gather and comprehend complex visual cues from their surroundings. A crucial vision task for perception is Semantic Segmentation (SS) [1], which to be effectively trained and deployed requires collecting and annotating large datasets that cover the entire distribution of possible visual events. However, data collected by autonomous vehicles are generally covered by privacy regulations and cannot be freely shared with a central institution to train a centralized model. To tackle this issue, privacy-preserving approaches such as Federated Learning (FL) [2] have been proposed as a possible solution to collaboratively train models across devices or data sources (clients) in a distributed fashion, while maintaining data confinement.

Despite the effectiveness of FL, a significant challenge in training through distributed learning arises from the *statistical heterogeneity* among clients, leading to slower convergence of the global model. This is generally caused by the *domain shift* of features within the same categories across clients or *label distribution skewness*, which refers to an imbalanced or uneven distribution of labeled data across participating clients. Most methods primarily focus on addressing the label distribution skewness [3], [4]. In contrast, only a few consider domain shift as a primary source of statistical heterogeneity [5], [6]. Additionally, these issues have been predominantly studied from a theoretical perspective, with limited emphasis on more structured tasks such as SS, addressing data heterogeneity among clients mainly for the classification task [7]–[9], especially on small datasets. Only a handful of studies have ventured into large-scale visual classification [10]. Still, the research community has shown a growing interest in FL methods for autonomous driving [11]–[14]. However, the SS task has largely been overlooked in this context, with only a few notable exceptions [15]–[17].

In particular, FedDrive [15] was the first benchmark of SS within the context of FL for autonomous vehicles, evaluating model generalization across various real-world conditions on synthetic and real-world federated datasets. While FedDrive mainly focuses on the domain shift problem, another critical source of statistical heterogeneity is label skewness. This arises because certain clients may have access to environments with different sets of categories and observe some of them more frequently than others. Indeed, a fleet of autonomous vehicles deployed in diverse locations may have access to a restricted or imbalanced set of classes. For instance, they might record varying numbers of riders, vehicles, traffic signs, or pedestrians, or might not encounter certain classes at all.

To study this problem, we introduce FedDrive v2, an expansion to the existing FedDrive benchmark [15] to evaluate the impact of label distribution skewness in distributed training. With FedDrive v2, we double the federated datasets introducing a new imbalanced federated split and novel training/test partition intended to investigate the domain shift challenge further. In addition, we analyze how style transfer and domain generalization techniques are affected by the label skewness problem. Finally, at inference time, we explore the impact of using local statistics for SiloBN [5] on client-local test sets, resulting in up to 12 mIoU percentage points improvement.

II. Federated Datasets description

FedDrive v2 introduces a new train/test partition and a novel client distribution, generating six new federated datasets, thus doubling the scenarios already present in FedDrive [15].

We build on the same federated setting and *centralized datasets* of FedDrive: Cityscapes [18], and IDDA [19]. Cityscapes is naturally divided into 2975 annotated images for

training and 500 for testing, all gathered from similar cities from Central Europe in optimal weather conditions, while IDDA is a synthetic dataset with 105 different *domains* of 60 images each, uniquely characterized by the triad (*weather*, *viewpoint*, *town*), where *weather*, *viewpoint* and *town* are one among three weather conditions, five different points of view for the camera recording the scenes simulating various vehicles, and six cities and one bucolic country, respectively.

New Bus training/test partition. For IDDA, the original benchmark already provided two different training/test partitions, both made of one training set and two test sets, the *seen* and *unseen*. Specifically, FedDrive introduced the *Country* and the *Rainy* partitions, where the unseen test set consists of all the images from the bucolic country domains for the former and rainy domains for the latter. The rainy weather condition and the bucolic country choices were made to exacerbate the semantic and appearance shift between the training and test images. However, the viewpoint axis, which may constitute another possible source of mismatch between training and test images, has been ignored. Here, we introduce a new unseen test set, *Bus*, where the unseen test set has all the images from the bus domains. We chose the viewpoint of the bus since it provides a point of view set from above that is sensibly different from one of the other available vehicles, to maximize the discrepancy between the training and test images. In all these three partitions, the seen test set has 12 randomly sampled images from each domain left outside the unseen test set. Having two test sets of this nature is a simple yet effective choice to estimate the generalization capabilities of the methods on both seen and unseen environmental conditions. The remaining images constitute the training set and are divided across clients following different distributions (as explained below) to create the FL environment.

New Class Imbalance client distribution. FedDrive proposes two client distributions, the *Uniform* and *Heterogeneous*, to study the effects of the statistical heterogeneity caused by domain shift. In the Uniform distribution, each client possesses random images for Cityscapes or one random image from each domain of the training set for IDDA. On the contrary, in the Heterogeneous, each client has access to images from a single city (for Cityscapes) or domain (for IDDA). The heterogeneity of this distribution mainly derives from the domain shift, based on the agreeable assumption that different cities or domains have various visual features. Potentially, it derives from the *quantity skewness* too for the Cityscapes setting only, where the dimensions of the datasets of the clients are diverse. However, another realistic source of statistical heterogeneity not deemed by FedDrive is the label skewness since some clients may access environments where specific categories are more frequent than others or some clients are prevented from accessing these categories at all. Therefore, we introduce a novel client distribution, the *Class Imbalance*, to analyze the behavior of the methods already studied in FedDrive in the presence of label skewness. We generate the distribution by proposing a general algorithm that maximizes the label skewness across clients. This is done iteratively by allocating images with the least frequent classes to clients, continuously keeping track of the images where each class appears at least once, and then moving to the new least frequent class, until all the images have been allocated to the clients. Our algorithm allows generating clients of any dimension to simultaneously introduce quantity skewness. We formally describe the procedure in Algorithm I.

ALGORITHM I: CLASS IMBALANCE CLIENT DISTRIBUTION GENERATION

Initialize:
$\mathcal{D}$ = set of all the training images
Ordered set of empty client datasets C
Ordered set $S : |S| = |C|$, $\sum_{s \in S} s = |\mathcal{D}|$,
s_i = desired # of samples (x, y) for client i
$\mathcal{D}_c \subseteq \mathcal{D}$: set of all $(x, y) \in \mathcal{D}$ such that class c appears in y
for ***each*** $s_i \in S$ **do**
  $\mathcal{X} = \arg\min_{\mathcal{D}_c} |\mathcal{D}_c|$
  while $|C_i| < s_i$ **do**
    Extract a subset $\mathcal{E}$ of $min(s_i - |C_i|, |\mathcal{X}|)$ uniformly sampled image and ground truth pairs (x, y) from $\mathcal{X}$
    $C_i = C_i \cup \mathcal{E}$
    $\mathcal{D}_c = \mathcal{D}_c \setminus \mathcal{E}\ \forall c$
return C

TABLE I
SUMMARY OF THE FEDDRIVE V2 FEDERATRED DATASETS.

Dataset	Setting	Distribution	# Clients	# img/cl	Test clients
Cityscapes [18]	-	*Uniform, *Heterogeneous, Class Imbalance	146	10 - 45	unseen cities
IDDA [19]	Country	*Uniform, *Heterogeneous, Class Imbalance	90	48	seen + unseen (country) domains
	Rainy	*Uniform, *Heterogeneous, Class Imbalance	69	48	seen + unseen (rainy) domains
	Bus	Uniform, Heterogeneous, Class Imbalance	83	48	seen + unseen (bus) domains

Updated statistics of the federated datasets, including the Class Imbalance client distribution and the *Bus* setting, are provided in Table I. In all tables, we mark with $(^*)$ records taken from [15]. Moreover, Figure 1 shows a comparison of the class distribution among the clients per each client split for the Rainy training/test partition for IDDA.

III. INFERENCE STRATEGIES

FedDrive examined the use of SiloBN [5] as a method to counter domain shift in FL. They studied its effectiveness in the context of self-driving cars, specifically for the SS task. SiloBN enables the federated training of a model with Batch Normalization (BN) layers, even in cases of statistical heterogeneity, by keeping the BN statistics local to the clients. At inference time, the *Standard strategy* is to compute the BN statistics directly from the test set. However, it is reasonable to assume that each training client has its local test set. We simulated this eventuality for the IDDA Heterogeneous federated datasets with the *By Domain strategy*. We assign each image of the seen test set to the client of its corresponding domain. Then, the model is evaluated over all the local test sets by using the corresponding local statistics.

IV. EXPERIMENTS

We run all the new experiments using a Tesla V100-SXM2-16GB. The chosen lightweight architecture is BiSeNet V2 [20]. The hyper-parameters choice is the same as the original FedDrive paper [15]. In the experiments with the server optimizer, we report only the results with the best server learning rate between 0.1 and 1.0.

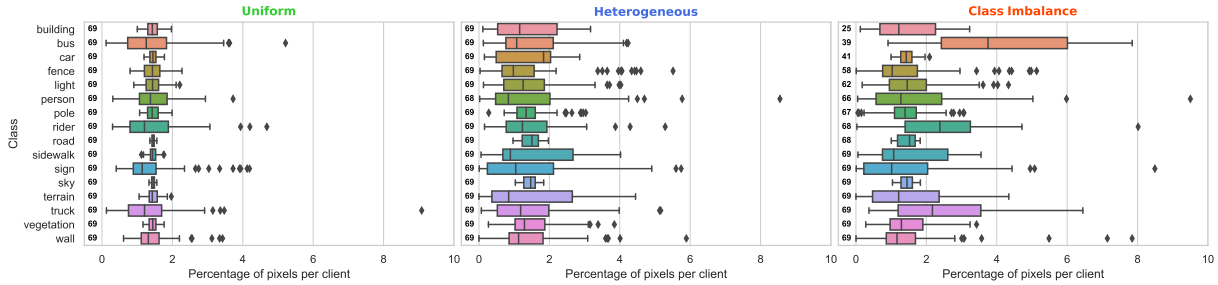

Fig. 1. Class distribution comparison for the Rainy IDDA training/test partition. For each class c, we evaluate the distribution of the number of c pixels in each client that possesses at least one image with a pixel of class c, relative to the total number of pixels of class c among all the clients. The numbers near the boxes are the number of clients having at least one image with that class, over 69 total clients. The Class Imbalance distribution has wider boxes, meaning there is more variation in the quantity of pixels of each class among the clients. Country and Bus provided similar cues on the class distributions.

TABLE II
CITYSCAPES RESULTS, CLASS IMBALANCE SPLIT.

Method	CFSI	LAB	mIoU ± std (%)
FedAvg	✗	✗	44.48 ± 1.70
	✓	✗	**48.28 ± 1.83**
	✗	✓	44.34 ± 1.73
SiloBN	✗	✗	**51.78 ± 1.15**
	✓	✗	50.82 ± 1.08
	✗	✓	51.48 ± 1.22

TABLE III
CITYSCAPES, SERVER OPTIMIZERS COMPARISON.

	SGD	FedAvgM
Uniform	*45.62 ± 1.25	40.04 ± 4.26
Heterogeneous	*43.33 ± 1.66	37.83 ± 4.61
Class Imbalance	44.48 ± 1.70	36.13 ± 4.52
	Adam	**AdaGrad**
Uniform	**45.91 ± 1.28**	44.30 ± 3.66
Heterogeneous	**45.08 ± 1.55**	42.28 ± 3.12
Class Imbalance	**45.21 ± 1.81**	39.78 ± 4.10

TABLE IV
IDDA, FEDAVG EXPERIMENTS.

		CFSI	LAB	Unseen	Seen
Uniform	*Country	✗	✗	49.74 ± 0.79	63.57 ± 0.60
	*Rainy	✗	✗	27.61 ± 2.80	62.72 ± 3.65
	Bus	✗	✗	58.51 ± 1.32	64.87 ± 0.65
Heterogeneous	*Country	✗	✗	40.01 ± 1.26	42.43 ± 1.78
		✓	✗	**45.70 ± 1.73**	**54.70 ± 1.12**
		✗	✓	45.68 ± 1.04	56.59 ± 0.90
	*Rainy	✗	✗	26.75 ± 2.32	38.18 ± 1.40
		✓	✗	**31.05 ± 2.68**	55.24 ± 1.65
		✗	✓	26.82 ± 1.78	**58.85 ± 0.89**
	Bus	✗	✗	38.13 ± 1.96	45.71 ± 1.65
		✓	✗	48.88 ± 1.46	56.93 ± 1.39
		✗	✓	**50.48 ± 1.09**	**58.84 ± 0.97**
Class Imbalance	Country	✗	✗	47.58 ± 0.69	58.09 ± 0.78
		✓	✗	48.69 ± 0.82	59.67 ± 0.88
		✗	✓	**48.91 ± 0.78**	**60.07 ± 1.18**
	Rainy	✗	✗	**29.46 ± 1.90**	58.75 ± 1.45
		✓	✗	25.69 ± 2.77	60.37 ± 0.60
		✗	✓	29.11 ± 1.68	**61.22 ± 0.98**
	Bus	✗	✗	53.29 ± 2.05	60.47 ± 1.75
		✓	✗	52.91 ± 1.33	61.24 ± 1.24
		✗	✓	**54.01 ± 1.15**	**62.10 ± 0.55**

TABLE V
IDDA HETEROGENEOUS RESULTS, SILOBN EXPERIMENTS.

	CFSI	LAB	Unseen	Seen	
				Standard	By Domain
Country	✗	✗	*45.32 ± 0.90	54.46 ± 0.72	*58.82 ± 2.93
	✓	✗	*49.17 ± 1.01	63.43 ± 0.58	*61.22 ± 3.88
	✗	✓	***50.43 ± 0.63**	**64.59 ± 0.45**	*64.32 ± 0.76
Rainy	✗	✗	*50.03 ± 0.79	54.36 ± 0.83	*62.48 ± 1.42
	✓	✗	*50.54 ± 0.88	64.85 ± 0.72	*63.04 ± 0.31
	✗	✓	***53.99 ± 0.79**	**65.90 ± 0.55**	*65.85 ± 0.91
Bus	✗	✗	47.37 ± 0.80	57.84 ± 0.89	61.56 ± 1.39
	✓	✗	55.84 ± 0.99	65.78 ± 0.81	64.03 ± 2.68
	✗	✓	**56.23 ± 0.64**	**66.98 ± 0.34**	66.23 ± 0.83

A. Cityscapes results

Table II shows the Cityscapes results with the Class Imbalance distribution. SiloBN still improves the performance, but the gains are less pronounced with respect to the Heterogeneous one since SiloBN mainly tackles the domain shift rather than the label skewness. Analogously, CFSI [21] and LAB [22] seem not to be particularly helpful when combined with SiloBN. However, with this client distribution, CFSI improves the performance by four percentage points with respect to FedAvg alone in the experiments without SiloBN.

Table III compares four server optimizers at different client distributions. Adam consistently performs better than the other server optimizers, and SGD is always the second best. Maybe surprisingly, FedAvgM struggles to achieve good performance.

B. IDDA results

Table IV shows the results for the FedAvg experiments on the IDDA dataset for every proposed scenario, eventually applying CFSI and LAB style translation techniques. First, we observe that style translation techniques are always helpful since they improve the performance or perform the same as FedAvg alone in the worst case, as for the Class Imbalance Rainy experiments. Moreover, we observe LAB outperforms CFSI in all the Class Imbalance experiments. On the contrary, this is not always true for the Heterogeneous experiments, where LAB is superior only for the Bus ones. Finally, the Class Imbalance experiments perform much better than the related Heterogeneous experiments, meaning that the maximum label skewness you can achieve in this SS task only slightly contributes to the statistical heterogeneity, and you can reach higher statistical heterogeneity from the domain shift.

Table V and Table VI report the SiloBN experiments for the IDDA Heterogeneous and Class Imbalance federated datasets, respectively. LAB experiments consistently outperform most of the other experiments in most scenarios, also in this case. The By Domain inference strategy for the Heterogeneous experiments is especially good without style translation techniques. Additionally, if we compare these results with the ones in Table IV, we can observe that SiloBN + LAB is often the best combination of techniques.

Finally, we analyzed the behavior of the server optimizers for IDDA. However, in Tables VII and VIII, we show only the comparison of SGD with FedAvgM for the IDDA dataset because, contrary to the Cityscapes experiments, Adam and AdaGrad failed to achieve good performance in most of the experiments. The only two exceptions were two experiments without SiloBN that showed good performance on the unseen test set using Adam (namely, Uniform Rainy: unseen = $31.72 \pm 1.74\%$, seen = $65.39 \pm 0.52\%$; Class Imbalance Rainy: unseen = $35.78 \pm 1.93\%$ mIoU, seen = $57.50 \pm 1.17\%$ mIoU). In all the other cases, FedAvgM is always the best optimizer. For the SiloBN experiments, the By Domain inference strategy is always the best, improving the performance up to 12 mIoU percentage points for the Rainy FedAvgM experiment.

V. CONCLUSION

In this work, we extend FedDrive [15] by introducing a novel distribution of the clients to study the effect of label

TABLE VI
IDDA CLASS IMBALANCE RESULTS, SILOBN EXPERIMENTS.

	CFSI	LAB	Unseen	Seen
Country	✗	✗	51.19 ± 0.55	66.46 ± 0.35
	✓	✗	51.73 ± 0.69	67.22 ± 0.41
	✗	✓	**53.10 ± 0.55**	**67.33 ± 0.36**
Rainy	✗	✗	**54.68 ± 0.56**	66.41 ± 0.53
	✓	✗	54.35 ± 1.03	67.07 ± 0.39
	✗	✓	52.91 ± 0.84	**67.63 ± 0.32**
Bus	✗	✗	57.63 ± 0.59	67.54 ± 0.41
	✓	✗	57.68 ± 0.66	**67.89 ± 0.59**
	✗	✓	**58.02 ± 0.55**	67.80 ± 0.43

TABLE VII
SGD VS FEDAVGM COMPARISON ON IDDA, FEDAVG EXPERIMENTS.

		SGD		FedAvgM	
		Unseen	Seen	Unseen	Seen
Uniform	*Country	49.74 ± 0.79	63.57 ± 0.60	**55.47 ± 1.07**	**71.27 ± 0.85**
	*Rainy	27.61 ± 2.80	62.72 ± 3.65	29.83 ± 2.03	**70.99 ± 0.71**
	Bus	58.51 ± 1.32	64.87 ± 0.65	**62.32 ± 1.25**	**71.76 ± 0.37**
Heter.	*Country	40.01 ± 1.26	42.43 ± 1.78	**42.42 ± 2.15**	**44.38 ± 1.98**
	*Rainy	26.75 ± 2.32	38.18 ± 1.40	**31.91 ± 3.77**	**41.21 ± 1.98**
	Bus	38.13 ± 1.96	45.71 ± 1.65	**40.39 ± 1.56**	**48.92 ± 1.54**
Cl. Imb.	Country	47.58 ± 0.69	58.09 ± 0.78	**50.22 ± 1.17**	**63.01 ± 1.20**
	Rainy	29.46 ± 1.90	58.75 ± 1.45	32.30 ± 1.93	**63.96 ± 1.11**
	Bus	53.29 ± 2.05	60.47 ± 1.75	**54.71 ± 1.65**	**64.45 ± 1.00**

skewness and a new training/test partition for the IDDA dataset. We do so, by proposing an algorithm for generating class imbalanced splits for federated datasets. We also study the effect of a new inference strategy for SiloBN in the presence of test sets local to the clients. All our studies provide many additional experiments compared to FedDrive. We found that SiloBN [5], CFSI [21], and LAB [22] could still be helpful in the presence of label skewness despite these techniques being designed for domain adaptation and generalization. In addition, results indicate that the domain shift is more challenging than label skewness in SS for autonomous vehicles. Future efforts will focus on designing specific algorithms to address class imbalance and label skewness issues in federated semantic segmentation. Additionally, we aim to enhance FedDrive with large-scale real-world datasets that mirror the long-tail distribution found in autonomous driving scenarios.

ACKNOWLEDGEMENTS

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

REFERENCES

[1] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *International journal of multimedia information retrieval*, vol. 7, pp. 87–93, 2018.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] J. Zhang, Z. Li, B. Li, J. Xu, S. Wu, S. Ding, and C. Wu, "Federated learning with label distribution skew via logits calibration," in *International Conference on Machine Learning*. PMLR, 2022, pp. 26 311–26 329.

[4] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022, pp. 965–978.

TABLE VIII
IDDA, HETEROGENEOUS, SERVER OPTIMIZERS COMPARISON USING SILOBN.

Partition	Optimizer	*Unseen	Seen	
			Standard	*By Domain
Country	SGD	45.32 ± 0.90	54.46 ± 0.72	58.82 ± 2.93
	FedAvgM	**46.20 ± 1.20**	54.56 ± 1.29	**61.99 ± 1.51**
Rainy	SGD	**50.03 ± 0.79**	54.36 ± 0.83	62.48 ± 1.42
	FedAvgM	48.49 ± 1.04	51.71 ± 0.73	**63.69 ± 1.25**
Bus	SGD	**47.37 ± 0.80**	57.84 ± 0.89	61.56 ± 1.39
	FedAvgM	46.91 ± 1.04	55.98 ± 0.76	**63.41 ± 1.43**

[5] M. Andreux, J. O. du Terrail, C. Beguier, and E. W. Tramel, "Siloed federated learning for multi-centric histopathology datasets," in *Springer*, 2020.

[6] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," *arXiv preprint arXiv:2102.07623*, 2021.

[7] S. P. Karimireddy, M. Jaggi, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Mime: Mimicking centralized stochastic algorithms in federated learning," *arXiv preprint arXiv:2008.03606*, 2020.

[8] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.

[9] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," *arXiv preprint arXiv:2111.04263*, 2021.

[10] T.-M. H. Hsu, H. Qi, and M. Brown, "Federated visual classification with real-world data distribution," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*. Springer, 2020, pp. 76–92.

[11] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 8423–8434, 2021.

[12] Z. Du, C. Wu, T. Yoshinaga, K.-L. A. Yau, Y. Ji, and J. Li, "Federated learning for vehicular internet of things: Recent advances and open issues," *IEEE Open Journal of the Computer Society*, vol. 1, pp. 45–61, 2020.

[13] D. Jallepalli, N. C. Ravikumar, P. V. Badarinath, S. Uchil, and M. A. Suresh, "Federated learning for object detection in autonomous vehicles," in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*. IEEE, 2021, pp. 107–114.

[14] Y. Tian, J. Wang, Y. Wang, C. Zhao, F. Yao, and X. Wang, "Federated vehicular transformers and their federations: Privacy-preserving computing and cooperation for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, 2022.

[15] L. Fantauzzo, E. Fanì, D. Caldarola, A. Tavera, F. Cermelli, M. Ciccone, and B. Caputo, "Feddrive: Generalizing federated learning to semantic segmentation in autonomous driving," in *Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2022.

[16] C.-H. Yao, B. Gong, H. Qi, Y. Cui, Y. Zhu, and M.-H. Yang, "Federated multi-target domain adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1424–1433.

[17] D. Shenaj, E. Fanì, M. Toldo, D. Caldarola, A. Tavera, U. Michieli, M. Ciccone, P. Zanuttigh, and B. Caputo, "Learning across domains and devices: Style-driven source-free domain adaptation in clustered federated learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 444–454.

[18] 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 *CVPR*, 2016.

[19] E. Alberti, A. Tavera, C. Masone, and B. Caputo, "Idda: A large-scale multi-domain dataset for autonomous driving," *IEEE RAL*, vol. 5, 2020.

[20] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet V2: bilateral network with guided aggregation for real-time semantic segmentation," *CoRR*, vol. abs/2004.02147, 2020.

[21] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," *CVPR*, 2021.

[22] J. He, X. Jia, S. Chen, and J. Liu, "Multi-source domain adaptation with collaborative learning for semantic segmentation," *CoRR*, 2021.