This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Biased Class disagreement: detection of out of distribution instances by using differently biased semantic segmentation models.

Roberto Alcover-Couso
VPU-Lab UAM
roberto.alcover@uam.es

Juan C. SanMiguel,

Marcos Escudero-Viñolo,

Abstract

Autonomous driving heavily relies on accurate understanding of the surrounding environment, which is facilitated by semantic segmentation models that classify each pixel in an image. However, training these computer vision models using available datasets often fails to capture the diverse conditions and objects that can be encountered during a trip. Adverse weather conditions and the presence of Out-of-Distribution (OOD) instances, such as wild animals and debris, are common challenges in autonomous driving. Unfortunately, current models struggle to perform well in unseen conditions.

To address these limitations, this paper proposes a comprehensive approach that integrates uncertainty quantification and bias reinforcing within the framework of Unsupervised Domain Adaptation (UDA). Our approach leverages multiple models with diverse biases, aiming to assign high-confidence predictions to OOD instances by mapping them to the selected prior semantic category. Extensive evaluations on the MUAD dataset demonstrate the effectiveness of our approach in improving performance and robustness against OOD instances. Notably, our approach achieves outstanding results, securing the first position in the MUAD challenge.

1. Introduction

In the past decade, significant advancements have been made in computer vision systems, largely attributed to the success of deep learning. These breakthroughs have led to substantial community growth and increased industrial investment. However, many current models suffer from a critical limitation—they struggle to operate effectively in unseen scenarios or with unfamiliar objects [9, 22, 24]. This phenomenon is commonly referred to as domain shift [19], which results in the presence of Out-of-Distribution (OOD) instances [29].

In the context of semantic segmentation, OOD instances refer to objects or regions within an image that do not belong to any of the predefined classes during training. In the segmentation of urban scenarios, e.g., for autonomous driving, this limitation poses a significant challenge, especially when faced with diverse adverse weather conditions that can severely hinder the performance of vision systems [19]. Moreover, the presence of OOD instances such as wild animals and debris, which the models have not been trained on, can lead to unpredictable behaviors in the vision system [29].

To address the challenges posed by diverse weather conditions in autonomous driving, domain adaptation methods have emerged as prominent approaches [12, 19, 24]. These methods propose a framework to learn from an extensive source domain and transfer the acquired knowledge to a shallower target domain. In the context of autonomous driving, these frameworks typically involve training models on clear daytime images as the source domain and aim to adapt their performance to handle the more challenging conditions encountered in the target domain.

On the other hand, the challenge of handling OOD instances is inherent to any classification task. Traditional classification frameworks commonly rely on the maximum softmax probability to estimate prediction confidence. Consequently, methods for handling OOD instances have emerged, employing various regularization techniques that can be broadly categorized [29]: Label space redesign [14, 20] or as ensemble models [25]. Notably, ensemble models, which leverage knowledge from all these categories, have shown remarkable results in addressing OOD instances [25].

In this paper, we present a comprehensive framework that leverages Unsupervised Domain Adaptation (UDA) for autonomous driving. Our approach overprioritizes the learning of a specific class through data sampling during model training, i.e., biasing the model. Thereby, it is based on promoting the model to assign the bias class to the OOD instances, as these instances are expected to not be aligned with the learned feature distribution of any other trained class. By employing multiple models with different forced

biases, if an instance is consistently classified as their respective bias by each model, it is prone to be an OOD instance.

Through extensive experiments and evaluations conducted on the MUAD dataset [8], we demonstrate the effectiveness of the proposed method in improving performance and detection of OOD. Our approach achieved first position results in the MUAD challenge by a large margin, showcasing its superiority compared to other approaches. These findings contribute to the growing body of research in uncertainty estimation and adaptation techniques, paving the way for more robust and trustworthy autonomous driving systems.

The remainder of the paper is organized as follows: In Section 2, we provide a background section on related work. Section 3 introduces our proposed method in detail. In Section 4, we present the experimental section, including ablation studies for our method and a study on the performance impact of each component. Finally, we conclude the paper in Section 5, summarizing the key findings and highlighting future research directions.

2. Related Work

In the context of semantic segmentation for autonomous driving, adverse weather conditions and OOD instances compose the main challenges of reliable models [8]. In order to tackle adverse weather conditions, researchers rely on UDA techniques to make robust models which can generalize from labelled clear daytime images to night time, rainy or foggy images [12, 24]. On the other hand, OOD instances are commonly tackled through regularization techniques [14, 29].

In the following section we provide an introduction to UDA and OOD instance detection.

2.1. Unsupervised Domain Adaptation

The goal of UDA is to train a model presenting a good performance in an unlabeled target domain by training on a labeled source domain. To that end, two main approaches can be differentiated [19]: Input space adaptation [17, 22], which align the color images from both domains so that the domain gap is narrowed. Output space adaptation [1, 23, 24] align output features from the network, forcing the network to present similar activation patterns on both domains. However, most successful approaches employ both alignments [10, 11, 13].

Pseudo-labelling The most prominent output space adaptation is to generate pseudo-labels for the target domain based on the predictions of a tentative model. However, this comes with a big drawback: *concept drifting* [6, 18]. Concept drifting is known as the event where a model trained on pseudo-labels over-fits to false positives, thus, reducing at each time-step the performance on the target domain. The main challenge here is to reduce the number of false positives as much as possible. However, in the context of domain adaptation and uncertainty estimations, DNNs often provide high-confidence mis-classifications. Therefore, to prevent concept drifting, researchers employ a threshold which can be manually defined [22] or dynamically defined [26, 30, 31].

Teacher-Student pseudo-labels Most of the UDA methods pre-compute the pseudo-labels offline, train the model, and repeat the process [32]. Alternatively, pseudo-labels can be calculated online during the training [22]. In this vein, in order to enforce consistency, a teacher-student framework can be employed [10, 11, 13]. By training the student network with pseudo-labels generated from the teacher network, concept drift is avoided to a certain extent due to the teacher student not being updated from those pseudo-labels. However, hard to classify classes which are not captured by the teacher network will never be learnt by the student.

Bias towards popular classes In the context of UDA, DNNs have been observed to exhibit a bias towards the most populated classes in the source set [3]. To alleviate such discrepancy, sampling strategies which over-sample less frequent classes in the source set are employed to improve performance [10]. However, studies have shown that unseen variations of objects in the target set tend to be classified as the most seen semantic category [3, 5, 4]. Additionally, it has been also argued that biases in spatial location also influence the classification of unseen objects [4].

To address these issues, we propose to: First, induce different biases through sampling in order to train models that assign high confidence predictions of the bias class to OOD instances. By employing multiple models with different learned class biases, if an instance is classified as their respective bias by each model, it can be assumed to be an OOD sample. Second, swap the pseudo-labels of easy to classify semantic categories which present low confidence to hard to classify semantic categories.

2.2. Out of Distribution segmentation

Segmenting OOD instances remains a relatively unexplored task, primarily due to the inherent complexity of image segmentation and the prevalent issue of over-fitting in current segmentation frameworks [12, 19, 4, 29]. In the context of autonomous driving, OOD instances such as wild animals and debris are to be expected in any deployed framework. However, their presence can lead to unexpected behaviors since they are not included in the training labels [29].

Label Space Redesign The commonly used one-hot encoding for categorical information in classification introduces a hard transition between semantic categories, resulting in over-fitting to the training images. Moreover, in the context of OOD, this hard encoding restricts the model's output to low entropy predictions, limiting the calibration of semantic segmentation frameworks. To mitigate these constraints imposed by one-hot encoded vectors, label smoothing [16] aims to reduce overconfidence by introducing label noise based on the label distribution. This regularization technique prevents the network from overly relying on ground-truth labels, leading to more robust predictions [16]. Additionally, the use of multi-labeled ground truth has demonstrated remarkable performance in supervised semantic segmentation tasks [2].

Ensemble Methods Ensemble methods involve combining the predictions of multiple individual models to obtain a single prediction that leverages the collective knowledge of the ensemble [15, 27, 28]. Additionally, ensembles allow for the definition of criteria to filter out inconsistent classifications, thereby improving overall performance and robustness [25].

To address OOD segmentation, we propose employing an ensemble method criteria for filtering out inconsistent classifications. Additionally, we utilize label smoothing to alleviate over-fitting in our semantic segmentation framework.

3. Methodology

In this section we provide an overview of our UDA training method and our OOD segmentation inference method. We first formalize UDA and our framework modules. Then, we explain our class disagreement inference method for OOD segmentation. Figure 1 depicts an overview of our framework, illustrating the training procedure where two models are trained as explained in the following section. During inference, these M models are employed together to generate different segmentation maps. By conducting a pixel-wise comparison of the maps, we can identify discrepancies and assign them as OOD instances.

3.1. Unsupervised Domain Adaptation

Let $\{X_S, Y_S\}$ and X_T be the labeled source set (S) and the unlabeled target set (T). The goal is to obtain a DNN G_θ which can indistinguishably classify both domains. Extensive research proves that training G_θ on the source set with a pixel-wise cross-entropyloss results in a low performance on the target set. If the prediction of the model for image x_S^i is $\hat{y}_S^i$:

$$\mathcal{L}_{CE}^i = \sum_{j=1}^{H\times W} CE(x_S^{i,j}, y_S^{i,j}) = - \sum_{j=1}^{H\times W} \sum_{c=1}^{C} y_S^{i,j,c} log(\hat{y}_S^{i,j,c}), \tag{1}$$

where $\{x_S^i, y_S^i\} \in \{X_S, Y_S\}$ are random color images and it's corresponding one-hot encoded ground-truth map from the source set of height H and width W. Each of their pixels are identified by: $j \in [1, H \times W]$, C is the number of semantic categories.

Teacher-Student To address the domain gap we follow a teacher-student framework of two DNNs: H_Θ the teacher network and G_θ the student network. The teacher network is updated every time step t following an exponential moving average (EMA) of the student network [21]:

$$\Theta_{t+1} = \alpha\Theta_t + (1-\alpha)\theta_t. \tag{2}$$

This teacher model generates pseudo-labels $p_T^{i,j}$ for the student network:

$$p_T^{i,j} = \arg\max_c H_\Theta(x_T^i)^{j,c}, \tag{3}$$

weighted by the ratio of pixels exceeding a threshold τ of the maximum softmax probability [10, 22]:

$$q_T^i = \frac{\sum_{j=1}^{H\times W} \max_c H_\Theta(x_T^i)^{j,c} > \tau}{H \times W}, \tag{4}$$

to train the student model with a weighted cross-entropy on the target pseudo-labels [10, 22]:

$$\mathcal{L}_T^i = \sum_{j=1}^{H\times W} q_T^i CE(x_T^{i,j}, p_T^{i,j}). \tag{5}$$

In summary, the student network is updated through stochastic gradient descent employing the cross-entropy loss1 on the source set and the weighted cross-entropy on the target pseudo-labels 5.

3.2. Reinforcing infrequent classes

Source Sampling Less frequent classes on the source set tend to present lower performances on the target set regardless of their frequency in the target set [10]. To that end we employ sampling strategies which over-sample less frequent classes based on the class frequency ($\mathbf{f}$) in the source set:

$$f_c = \frac{\sum_{i=1}^{N} \sum_{j=1}^{H\times W} y_S^{i,j,c}}{N \times H \times W}, \tag{6}$$

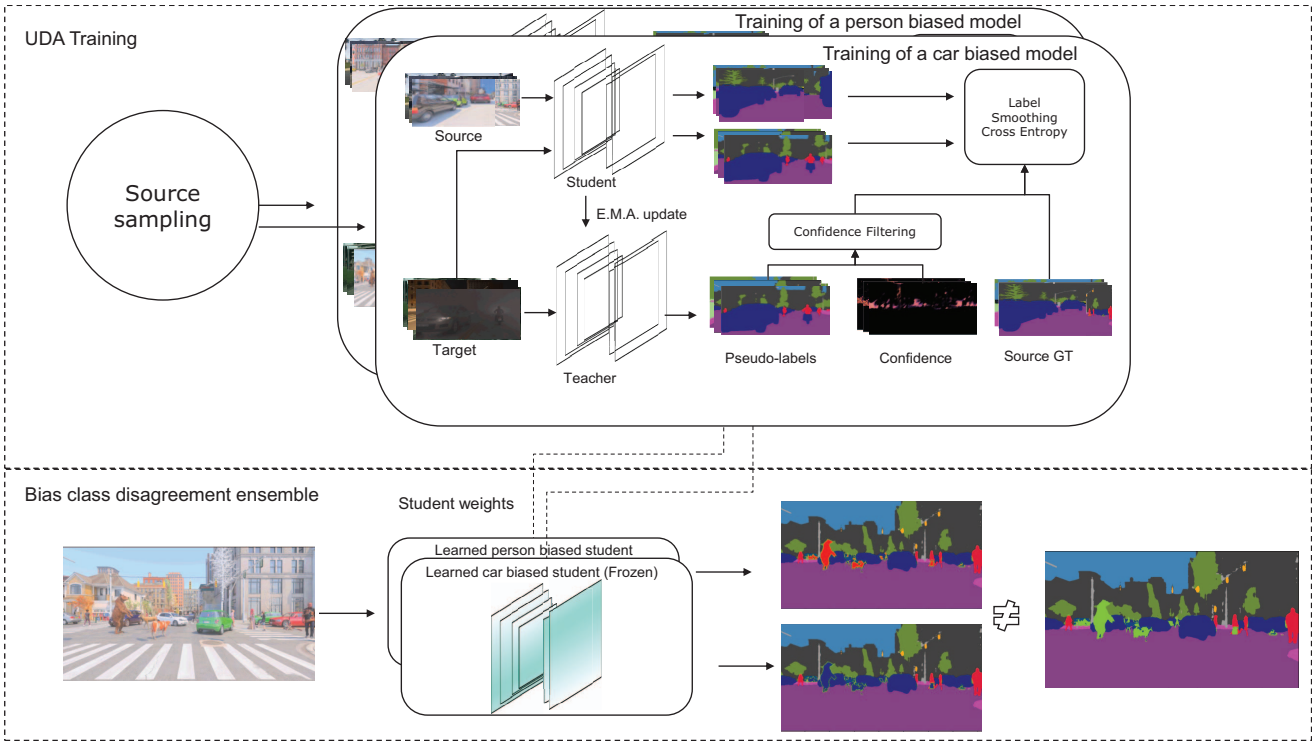

Figure 1: Overview of the proposed framework exemplified with $M = 2$ models, and the bias classes c' represent "Car" and "Person" for each respective model. The training procedure begins with a source sampling procedure selection. Subsequently, M models are trained in a teacher-student manner. During inference, these student models are employed jointly to generate two segmentation maps. A pixel-wise comparison is then performed to identify discrepancies, enabling the detection of OOD instances.

where N is the number of images in the source set. So that the probability of sampling an image crop with a class c is a $softmax$ of the $1 - \mathbf{f}$ frequencies.

We propose to force a class bias during trainining, specifically, if c' is the target biasing class, we propose to sample from a Bernoulli distribution, $B(1, .5)$, whether to select if a sample labelled as class c' is incorporated into the training. In case of *failure*, samples are incorporated with the original probability [10], see Equation 6. This sampling forces a bias towards class c' by training mainly with images containing the selected class.

Label Smoothing Cross Entropy : Aiming to alleviate the natural over-fitting of semantic segmentation models [19, 4], we propose to employ label smoothing on the cross entropy loss, Equation 1, following the standard practices [16]:

$$\mathcal{L}^i_{LSCE} = \mathcal{L}^i_{CE} + \beta \sum_{j=1}^{H \times W} CE(x^{i,j}_{\cdot}, 1 - \mathbf{f}), \qquad (7)$$

where β is the label smoothing weight and $\mathbf{f}$ the source frequencies, see Equation 6. This smoothing can be applied to source or target domain.

Confidence Filtering In order to promote the classification of low represented classes, we propose a method that involves switching the pseudo-labels (Equation 3) of classes exhibiting significant variation in their classification probabilities. We follow the hypothesis that Pseudo-labels with a wide confidence distribution often exhibit a bimodal distribution pattern. In this pattern, high-confidence associations are typically assigned to true positives, indicating correct classifications. On the other hand, false positives tend to be allocated in the tail end of the confidence distribution. This means that the confidence scores for false positives are generally lower and more spread out, indicating uncertainty and potential miss-classifications.

To that aim, we select the 3 classes which have assigned the highest confidence standard deviation throughout the image. Then for those pixels, we filter out of the training as an unknown class the ones with lowest confidence. We employ the mean of the confidences assigned to that class as the threshold for filtering out pixels. This approach aims

to reduce concept drifting by employing only the pseudo-labels associated with high confidence.

3.3. Bias class disagreement ensemble

Finally, to enhance the robustness of our approach, we propose to utilize an output level ensemble of M models. Each of the models is biased towards a different prior class through the proposed sampling. Then, we detect pixels of OOD instances by analyzing label mismatch situations under the assumption that each biased model classifies an OOD pixel as one of their respective biased class. Formally, we condition the output level ensemble by studying the pixels for which the top predicted class is the bias class of each model and consider them pixels of OOD instances.

4. Experimental Exploration

In this section, we present the experimental exploration of our framework. Firstly, we introduce the dataset used, the evaluation metrics, and the training hyperparameters. Secondly, we conduct an ablation study to analyze the impact of different framework parameters. We then present the comparative results on the MUAD challenge. Finally, we conclude with remarks on our findings and discuss potential avenues for future work.

4.1. Setup

Dataset We analyze the performance of the proposed approach in the scope of the MUAD dataset [8]. MUAD is created to serve as a benchmark for evaluating multiple uncertainty types and tasks in autonomous driving. The dataset comprises a total of 10,413 images, divided into train, validation, and test sets.

The train set consists of 3,420 images, and the validation set contains 492 images. These sets are used for model training and do not include any OOD instance or adverse weather scenarios.

The test set comprises 6,501 images, and it serves as the evaluation set and the target domain. This set does not have ground-truth labels. It includes 551 train-alike images resembling the training set and 5950 images featuring different degrees of OOD instances and adverse weather [8].

To train our models, we utilize the train and validation sets. Reported performances of our models on the test set are obtained by submitting the results to the challenge's test server.

Evaluation metrics Five metrics are employed for evaluation following the MUAD challenge [8]: mECE, mAUROC, mAUPR, mFPR, mIoU.

mECE (Mean Expected Calibration Error) mECE measures the calibration quality of a model by quantifying the discrepancy between predicted confidence and accuracy. It is calculated as the average difference between predicted confidence and the true accuracy across different confidence intervals or bins. Lower mECE values indicate better calibration, where the model's predicted confidence aligns well with its actual accuracy.

mAUROC (Mean Area Under the Receiver Operating Characteristic Curve) mAUROC evaluates the performance of a model in binary classification tasks by measuring the trade-off between true positive rate (TPR) and false positive rate (FPR) at various classification thresholds. It is calculated as the average area under the Receiver Operating Characteristic (ROC) curve across multiple classes or samples. Higher mAUROC values indicate better discrimination ability of the model between positive and negative instances.

mAUPR (Mean Area Under the Precision-Recall Curve) mAUPR assesses the model's performance in binary classification tasks by measuring the trade-off between precision and recall at different classification thresholds. It is calculated as the average area under the Precision-Recall (PR) curve across multiple classes or samples. Higher mAUPR values indicate better performance in terms of precision and recall trade-off.

mFPR (Mean False Positive Rate) mFPR measures the rate at which false positives are generated by a model in binary classification tasks. It is calculated as the average ratio of false positives to the total number of negatives across multiple classes or samples. Lower mFPR values indicate better performance in terms of minimizing false positives.

mIoU (Mean Intersection over Union) mIoU evaluates the quality of segmentation models by measuring the overlap between predicted and ground truth segmentation masks. It is calculated as the average of the Intersection over Union (IoU) values across different classes or samples. Higher mIoU values indicate better segmentation accuracy and alignment with ground truth masks.

As we are employing pseudo-labels, we focus on the mFPR in our ablation studies. However, the main metric for the challenge [8] is the mAUROC.

Training Architecture and hyperparameters For our experiments, we adopt the same network architecture and set of training parameters as outlined in the paper [11], with one exception: the "min pixels" parameter. In our case, we set "min pixels" to 30. This adjustment is made to account for semantic categories, such as "Bikes", which have a relatively smaller representation and size within the MUAD dataset. By setting a lower threshold, we aim to ensure adequate consideration and classification of such categories during training. Furthermore, for some experiments, we modify the our baseline architecture [11] to accommodate the number of classes in the MUAD dataset (21) [8]. However, since the test server expects fewer classes (19), we

Bias class (c')	mFPR↓	mIoU↑
"Car"	.7499	.5309
"Person"	.5720	.3871
"Bike"	.7039	.0973

Table 1: Ablation study of biased models towards the semantic category: c'.

employ the maximum over the target classes for our submissions, ensuring compatibility with the evaluation system.

4.2. Ablation study

Selection of the biased class Our primary approach to improve performance involves introducing a class bias to each model. However, if we force a bias towards an extremely underrepresented class, the model will primarily be trained only with a small subset of images (i.e., the set containing samples of that class). This can lead to over-fitting, resulting in a model that fails to provide useful knowledge or generalization to any class.

According to the dataset definition [8], there are 21 defined classes, including various semantic categories such as "Road", "Person", "Building", and "Animals". However, in the labeled sets, only three semantic categories of non-static objects are present: "Person", "Car" and "Bike". Considering this, Table 1 presents the mFPR and mIoU metrics of the biased models. For our evaluation, we focus on the "Car" and "Person" models, as the Bike model exhibits notably lower performance.

Label Smoothing As a regularization mechanism, we propose employing label smoothing [16] to both domains and classifying the unknown class in the training set. In Table 2, we present a performance comparison. It is important to note that a high smoothing weight will result in random and highly entropic classification, as indicated by the poor mIoU and low mFPR, as the model doesn't produce high-confidence predictions. Conversely, using a relatively small weight reduces the classification accuracy, leading to a decrease in mIoU. However, this approach significantly reduces the mFPR. For our experiments, we set the weight for label smoothing to 0.1. In Table 3, we present the results of an ablation study where we include the unknown label as a training class. This inclusion significantly improves performance by reducing the mFPR and increasing the mIoU. The allowance of pseudo-labels to include the unknown class is prone to produce less FP during training.

Pseudo-label confidence To filter out pseudo-labels during inference, it is necessary to employ a confidence threshold. In Table 4, we present an ablation study on the impact of different confidence thresholds. The chosen threshold

Label Smoothing	mFPR↓	mIoU↑
0	.7499	.5309
0.1	.6866	.5022
1	.4021	.0026

Table 2: Ablation study of the label smoothing weight.

Unknown	mFPR↓	mIoU↑
	.6866	.5022
✓	.6319	.5325

Table 3: Ablation study of the classification of unknown instances in the training set.

Threshold	mFPR↓	mIoU↑	mAUROC↑
-	.6227	.4887	..7732
.95	.6071	.5592	.7970
.9	.4934	.5795	.7925

Table 4: Ablation study on the pseudo-label threshold.

Method	mAUROC↑	mAUPR↑	mFPR↓	mECE↓	mIoU↑
Best model	.7925	.4029	.4934	.0703	.5795
BCDE	.8510	.4435	.3983	.0603	.6454

Table 5: Performance comparison of the best model obtained ("Car") and the results obtained by using the combination of two biased models ("Car" and "Person"). (KEY. BCDE: bias class disagreement ensemble).

has a significant influence on the mFPR metric. For our subsequent experiments, we employ a confidence threshold of 0.9.

Bias class disagreement ensemble inference To achieve the best possible final model, we leverage pseudo-labels from two models. With the chosen threshold, if both models classify a pixel with their respective selected bias c', we consider it part of an OOD instance. In such cases, we reduce the classification confidence to 1 minus the average confidence of the models predictions. For predictions that are not OOD, we calculate the average between the outputs of the two models.

Table 5 provides a comparison between the best employed model and the prediction generated using our bias class disagreement ensemble. This comparison showcases the performance and effectiveness of our approach in terms of classification accuracy and OOD instance detection.

Framework component analysis Table 6 provides a performance comparison of each module within our proposal and their respective impact on the mFPR metric. These

LSCE	CF	Unkown	BCDE	mAUROC↑	mAUPR↑	mFPR↓	mECE↓	mIoU↑
				.7091	.1907	.7499	.0970	.5309
✓				.7738	.1947	.6866	.1154	.5022
✓	✓			.7733	.2486	.6409	.1009	.4589
✓	✓	✓		.7616	.3591	.6319	.0968	.5325
✓	✓	✓	✓	.8510	.4435	.3983	.0603	.6454

Table 6: Performance comparison of each module of the framework, see Figure 1. Results from the MUAD test set evaluated by the server. The mFPR is highlighted as the most critical metric for our framework. (Key. LSCE: Label Smoothing Cross Entropy loss. CF: Confidence filtering. Unknown: Employing the unknown class for training. BCDE: Proposed bias class disagreement ensemble).

modules are individually evaluated to assess their effectiveness in improving the model's performance. Subsequently, the best-performing modules are combined to obtain our final model, which is optimized across all employed metrics.

Visual comparisons Figure 2 visually compares the predicted segmentation maps of the baseline model, our best individual model, and our bias class disagreement ensemble (see Table 6) for adverse weather samples. Notably, the baseline model (second row) struggles to discriminate infrequent semantic classes like "Bike," as illustrated in the first and second columns. In contrast, our proposed framework exhibits improved performance in accurately segmenting these challenging classes.

Figure 3 focuses on comparing the confidence of prediction for the baseline model, our best individual model, and our bias class disagreement ensemble (see Table 6) for images with OOD instances. It is evident that the baseline model (second row) demonstrates consistently high confidence for every OOD instance, hindering any form of OOD detection [29].

Both figures highlight the substantial improvements our proposed framework offers compared to the baseline model [11]. The bias class disagreement ensemble model, in particular, shows notable advancements in terms of accurate predictions under adverse weather conditions and effectively lowering the predicted confidence of OOD instances.

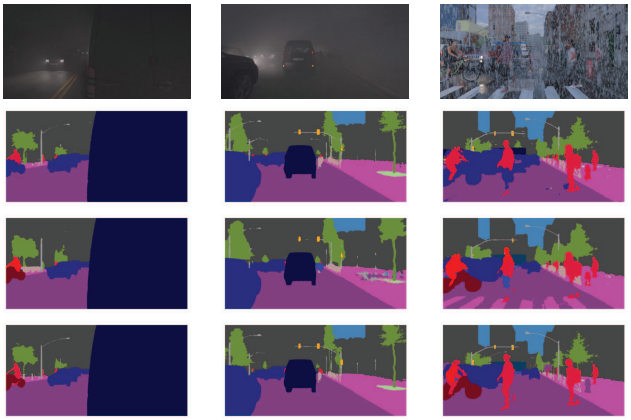

Figure 2: Comparison of predicted segmentation maps. First row illustrate the color images of different adverse weather images. Second to Fourth rows present the baseline model, best individual model and the bias class disagreement ensemble predictions respectively. Each predicted map is codified with the Cityscapes color map [7].

Figure 3: Comparison of prediction confidence maps, yellow indicates highest confidence and blue lowest confidence. First row illustrate the color images of different images containing OOD instances. Second to Fourth rows present the baseline model, best individual model and the bias class disagreement ensemble confidence maps respectively. Note that OOD should be assigned the lowest confidence.

4.3. MUAD challenge

In Table 7, we present the top three results from the MUAD challenge [8]. Our best performing model demonstrates outstanding performance, surpassing the other participants[1] by a significant margin across multiple metrics. Notably, our model achieves nearly three times better results in terms of mECE and mAUPR, indicating substantial improvements in both OOD classification and calibration. Additionally, our model excels in segmentation performance, as demonstrated by the mIoU and mFPR. These results serve as evidence of the efficacy of our approach in addressing various uncertainty challenges in the field of autonomous driving.

[1]Here we include the performance of methods uploaded until 18th July of 2023

Method	mAUROC↑	mAUPR↑	mFPR↓	mECE↓	mIoU↑
Ours	.8510	.4435	.3983	.0603	.6454
2nd	.7673	.1821	.4720	.1855	.3796
3rd	.7429	.1712	.5204	.3116	.3656

Table 7: MUAD Challenge Leaderboard: Top 3 Contestants. Results from 18th July 2023.

5. Conclusions

In this work, we have described an approach for addressing uncertainty in autonomous driving tasks. Through a series of experiments and evaluations on the MUAD dataset, we show evidences of the effectiveness of the proposed method.

By leveraging pseudo-labels and introducing biases through sampling, we successfully improved the robustness of deep neural networks in handling concept drifting and reducing biases towards popular classes. Our approach also incorporated label smoothing to enhance generalization and mitigate overconfidence.

The experimental results showcased the significant impact of our method on various evaluation metrics. We achieved remarkable performance improvements, as evidenced by substantial reductions in mFPR and outstanding results in terms of mAUROC, mAUPR, mECE, and mIoU. Notably, our best-performing model outperformed the other participants in the MUAD Challenge by large margins, attaining nearly three times better performance in mECE and mAUPR. However, there are still areas for further investigation and improvement. Future research can explore the application of our approach in different autonomous driving scenarios and datasets, as well as investigate additional strategies to handle uncertainty, such as training the ensemble or domain adaptation techniques.

Acknowledgement. This work has been funded by the SEGA-CV (TED2021-131643A-I00) and the HVD (PID2021-125051OB-I00) projects of the Ministerio de Ciencia e Innovación of the Spanish Government.

References

[1] Devika A.K., Rakesh Kumar Sanodiya, Babita Roslind Jose, and Jimson Mathew. Visual domain adaptation through locality information. *Engineering Applications of Artificial Intelligence*, 2023. 2

[2] Roberto Alcover-Couso, Marcos Escudero-Vinolo, Juan C. SanMiguel, and Jose M. Martinez. Soft labelling for semantic segmentation: Bringing coherence to label downsampling. In *arXiv:2302.13961*, 2023. 3

[3] Roberto Alcover-Couso, Juan C SanMiguel, Marcos Escudero-Vinolo, and Pablo Carballeira. Per-class curriculum for unsupervised domain adaptation in semantic segmentation. *SSRN preprint https://ssrn.com/abstract=4410425, 2023*. 2

[4] Roberto Alcover-Couso, Juan C SanMiguel, Marcos Escudero-Vinolo, and Alvaro Garcia-Martin. On exploring weakly supervised domain adaptation strategies for semantic segmentation using synthetic data. *Multimedia Tools and Applications*, 2023. 2, 4

[5] Francesco Barbato, Marco Toldo, Umberto Michieli, and Pietro Zanuttigh. Latent space regularization for unsupervised domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, pages 2835–2845, 2021. 2

[6] Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. Curriculum labeling: Revisiting pseudo-labeling for semi-supervised learning. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3212–3223, 2016. 7

[8] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Angel Tena, Rémi Kazmierczak, Severine Dubuisson, Emanuel Aldea, and David Filliat. Muad: Multiple uncertainties for autonomous driving benchmark for multiple uncertainty types and tasks. In *Brit. Mach. Vis. Conf. (BMVC)*, 2022. 2, 5, 6, 7

[9] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv:1612.02649*, 2016. 1

[10] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 9924–9935, 2022. 2, 3, 4

[11] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. HRDA: Context-aware high-resolution domain-adaptive semantic segmentation. In *IEEE Eur. Conf. Comput. Vis. (ECCV)*, page 372–391, 2022. 2, 5, 7

[12] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Domain adaptive and generalizable network architectures and training strategies for semantic image segmentation. *arXiv preprint arXiv:2304.13615*, 2023. 1, 2

[13] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. MIC: Masked image consistency for context-enhanced domain adaptation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023. 2

[14] Y. C. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10948–10957, 2020. 1, 2

[15] Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 17703–17716. Curran Associates, Inc., 2022. 3

[16] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, page 4694–4703, 2019. 3, 4, 6

[17] Kieu Dang Nam, Tu M. Nguyen, Trinh V. Dieu, Muriel Visani, Thi-Oanh Nguyen, and Dinh Viet Sang. A novel unsupervised domain adaption method for depth-guided semantic segmentation using coarse-to-fine alignment. *IEEE Access*, 10:101248–101262, 2022. 2

[18] Claude Sammut and Michael Harries. Concept drift. In *Encyclopedia of Machine Learning*, pages 202–205, 2010. 2

[19] Manuel Schwonberg, Joshua Niemeijer, Jan-Aike Termöhlen, Jörg P. schäfer, Nico M. Schmidt, Hanno Gottschalk, and Tim Fingscheidt. Survey on unsupervised domain adaptation for semantic segmentation for visual perception in automated driving. *IEEE Access*, 11:54296–54336, 2023. 1, 2, 4

[20] Gabi Shalev, Yossi Adi, and Joseph Keshet. Out-of-distribution detection using multiple semantic label representations. In *Conf. Neur. Inf. Proc. Sys. (NIPS)*, page 7386–7396, 2018. 1

[21] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *NIPS*, 2017. 3

[22] Wilhelm Tranheden, Viktor Olsson, Juliano Pinto, and Lennart Svensson. Dacs: Domain adaptation via cross-domain mixed sampling. *IEEE Winter Conf. App. Comp. Vis. (WACV)*, pages 1378–1388, 2020. 1, 2, 3

[23] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 7472–7481, 2018. 2

[24] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2512–2521, 2019. 1, 2

[25] Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L. Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Eur. Conf. Comp. Vis. (ECCV)*, pages 560–574, 2018. 1, 3

[26] Yuxi Wang, Junran Peng, and Zhaoxiang Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 9072–9081, 2021. 2

[27] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Int. Conf. on Machine Learning*, volume 162, pages 23965–23998, 2022. 3

[28] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022. 3

[29] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 1, 2, 7

[30] Zhedong Zheng and Yi Yang. Unsupervised scene adaptation with memory regularization in vivo. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. 2

[31] Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision (IJCV)*, 2021. 2

[32] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *IEEE Eur. Conf. Comput. Vis. (ECCV)*, pages 297–313, 2018. 2