

The Robust Semantic Segmentation UNCV2023 Challenge Results

Xuanlong Yu, Juan C Sanmiguel, Xiaowen Zhang, Marcos Escudero-Viñolo, Rui Peng, Hanlin Tian, Xinyi Wang, Kenta Matsui, Jiaxuan Zhao, Tianhao Wang, et al.

▶ **To cite this version:**

Xuanlong Yu, Juan C Sanmiguel, Xiaowen Zhang, Marcos Escudero-Viñolo, Rui Peng, et al.. The Robust Semantic Segmentation UNCV2023 Challenge Results. 2023 IEEE/CVF International Conference on Computer Vision (ICCVW) Workshops, Oct 2023, Paris, France. pp.4620-4630, 10.1109/iccvw60793.2023.00496 . hal-04321188

HAL Id: hal-04321188
https://hal.science/hal-04321188v1

Submitted on 4 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

arXiv:2309.15478v1 [cs.CV] 27 Sep 2023

The Robust Semantic Segmentation UNCV2023 Challenge Results

Xuanlong Yu[1,2], Yi Zuo[3], Zitao Wang[3], Xiaowen Zhang[3], Jiaxuan Zhao[3], Yuting Yang[3], Licheng Jiao[3], Rui Peng[3], Xinyi Wang[3], Junpei Zhang[3], Kexin Zhang[3], Fang Liu[3], Roberto Alcover-Couso[4], Juan C. SanMiguel[4], Marcos Escudero-Viñolo[4], Hanlin Tian[5], Kenta Matsui[5], Tianhao Wang[6], Fahmy Adan[5], Zhitong Gao[7], Xuming He[7], Quentin Bouniot[8], Hossein Moghaddam[8], Shyam Nandan Rai[9], Fabio Cermelli[9], Carlo Masone[9], Andrea Pilzer[10], Elisa Ricci[11], Andrei Bursuc[12], Arno Solin[13], Martin Trapp[13], Rui Li[13], Angela Yao[14], Wenlong Chen[5], Ivor Simpson[15], Neill D. F. Campbell[16], and Gianni Franchi[2]

[1]SATIE, Paris-Saclay University, France, [2]U2IS, ENSTA Paris, Institut Polytechnique de Paris, France, [3]Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University, China, [4]VPU-Lab, Autonomous University of Madrid (UAM), Spain, [5]Imperial College London, UK, [6]The University of Texas at Dallas, USA, [7]ShanghaiTech University, China, [8]LTCI, Télécom Paris, Institut Polytechnique de Paris, France, [9]Politecnico di Torino, Italy, [10]NVIDIA AI Technology Center, Italy, [11]University of Trento, Italy, [12]valeo.ai, France, [13]Aalto University, Finland, [14]National University of Singapore, Singapore, [15]University of Sussex, UK, [16]University of Bath, UK

Abstract

This paper outlines the winning solutions employed in addressing the MUAD uncertainty quantification challenge held at ICCV 2023. The challenge was centered around semantic segmentation in urban environments, with a particular focus on natural adversarial scenarios. The report presents the results of 19 submitted entries, with numerous techniques drawing inspiration from cutting-edge uncertainty quantification methodologies presented at prominent conferences in the fields of computer vision and machine learning and journals over the past few years. Within this document, the challenge is introduced, shedding light on its purpose and objectives, which primarily revolved around enhancing the robustness of semantic segmentation in urban scenes under varying natural adversarial conditions. The report then delves into the top-performing solutions. Moreover, the document aims to provide a comprehensive overview of the diverse solutions deployed by all participants. By doing so, it seeks to offer readers a deeper insight into the array of strategies that can be leveraged to effectively handle the inherent uncertainties associated with autonomous driving and semantic segmentation, especially within urban environments.

1. Introduction

As computer vision applications based on Deep Neural Networks (DNN) penetrate various aspects of our lives, ensuring reliable and robust performance becomes paramount. This is particularly evident in domains like autonomous driving, where DNNs play a pivotal role in comprehending the surrounding environment. However, the potential for inaccuracies due to various sources of uncertainty looms large. The quantification of uncertainty inherent in DNN predictions holds significant importance. Inaccurate assessments of uncertainty could lead to erroneous decisions, a scenario that could be catastrophic, as witnessed in cases like autonomous driving [25, 28]. An increasing number of datasets are emerging that could model a broader spectrum of potential uncertain scenarios [43, 46]. Nonetheless, the datasets served for the DNNs may not adequately encapsulate the vast array of conditions and objects prevalent in real-world scenarios. Autonomous driving, for instance, contends with adversities weather conditions like inclement snow, rain, and fog, and may contend the presence of Out-of-Distribution (OOD) instances. By delving into the study and quantification of uncertainty, we attain invaluable insights into areas where the model may be less confident, guiding us to adopt precautionary measures or invoke human intervention as necessary.

Thanks to the release of relative datasets [4, 14, 40, 53], significant progress has been achieved in recent years

on the OOD example detection in semantic segmentation [2, 10, 13, 16, 23, 29, 31, 47]. To advance the field of semantic segmentation robustness and uncertainty quantification, the first ACDC Challenge [43], which was held at Vision for All Seasons workshop in IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR 2022), aims to deal with semantic segmentation under complex weather conditions and changes in the visual description of objects caused by weather. Yet OOD detection task is not covered in the challenge. Meanwhile, compared with the previous anomaly segmentation benchmarks such as SegmentMeIfYouCan [4], MUAD [14], a larger-scale dataset for autonomous driving with multiple uncertainty types, is proposed to evaluate the robustness of the algorithms in both segmentation accuracy and uncertainty quantification quality, especially OOD detection.

Inspired by the previous works, in this paper, we will introduce the MUAD challenge [1], which is held at the ICCV 2023 UNCV workshop. The challenge is launched on the CodaLab [39] platform. The researchers were encouraged to submit both the semantic segmentation class prediction maps and the corresponding prediction confidence maps of a subset selected from MUAD test sets. According to the final ranking with respect to the semantic segmentation and OOD detection performance, we select the best solutions and will go into the details in the later sections. We believe that this challenge is meaningful for the progress of safety AI.

2. The MUAD challenge

2.1. The dataset

MUAD [14], a synthetic dataset pioneering autonomous driving research, introduces various uncertainties in the real world. It offers 10413 instances encompassing day and night scenes: 3420 in the training set, 492 in the validation set, and 6501 in the test set. There are seven subsets in the test set, including normal set, OOD set, low adversity and high adversity sets, etc. Only test sets contain the OOD objects or special weather conditions such as rain, fog, or snow with different severities. We randomly select the images from the test to conduct the challenge test set.

The participants download the training and validation sets (containing the RGB images and the corresponding ground truth maps) as well as the test set (only the RGB images are provided), then design and train the models. Class prediction maps and confidence maps are required to be submitted during evaluation, which aims to provide enough information to help the decision-maker identify the patterns and find out the uncertain prediction like OOD objects. Different levels of weather conditions will also be challenging to the robustness of the models and the OOD example detectors.

[1] https://codalab.lisn.upsaclay.fr/competitions/8007

2.2. The evaluation protocols

We evaluate the overall semantic segmentation performance of the algorithms using the mean Intersection over Union (mIoU). For the confidence maps, we choose the mean Expected Calibration Error (mECE) [37] for the calibration of uncertainties and measure the OOD detection performance using mean Areas Under the operating Curve (mAUROC) and mean Areas Under the Precision/Recall curve (mAUPR), as well as mean False Positive Rate at 95% recall (mFPR), similarly to [20]. All the mean is calculated across all pixels. For mIoU, mAUROC and mAUPR, the higher is better. For mECE and mFPR, the lower is better.

3. The MUAD challenge results

3.1. Notations

For the training of segmentation algorithms, all the process commences with a dataset of training examples denoted as $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{|\mathcal{D}|}$, where $|\mathcal{D}|$ is the total number of image-label pairs. These pairs consist of images $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ and corresponding label maps $\mathbf{y}_i \in [\![0, N_C[\![^{H\times W}$, treated as realizations of a joint distribution $\mathcal{P}(X, Y)$. In this context, C, H, and W denote the number of channel, height, and width of the image, respectively, while N_C represents the number of classes within the dataset.

The neural network $F_{\boldsymbol{\theta}}(\cdot)$ is then employed to process the input data $\mathbf{x}_i$. This network functions as a parametric probabilistic model, denoted by $\hat{\mathbf{y}}_i = F_{\boldsymbol{\theta}}(\mathbf{x}_i) = P(Y = \mathbf{y}_i | X = \mathbf{x}_i; \boldsymbol{\theta})$. In essence, it estimates the conditional probability of the label $\mathbf{y}_i$ given the input image $\mathbf{x}_i$ based on a set of parameters $\boldsymbol{\theta}$.

3.2. Overview of the results

Baseline The baseline model is established using DeepLabV3+ [5], which employs a ResNet101 [19] as its underlying architecture. The training protocol adheres to the official MUAD configuration [14]. Let $\hat{\mathbf{y}}_i^{[\text{pred}]}$ symbolize the class prediction map generated by selecting the class with the highest value along the classes of $\hat{\mathbf{y}}_i$. This can be formally expressed as: $\hat{y}_i^{[\text{pred}]}(h, w) = \arg\max_c \hat{y}_i(c, h, w)$. Subsequently, the confidence map $\hat{\mathbf{y}}_i^{[\text{conf}]}$ is based on the associated maximum probability values [20]: $\hat{y}_i^{[\text{conf}]}(h, w) = \max_c \hat{y}_i(c, h, w)$.

Overview of the results By the end of the challenge, a total of 73 teams registered to participate. Throughout the course of the challenge, more than 700 results were submitted to the CodaLab server, with 19 teams opting to disclose their rankings publicly. The conclusive outcomes are presented in Table 1, where the most successful two methods are distinctly highlighted. The "NA" designation indicates instances where mECE metrics were impacted by unnormalized confidence scores. For the ultimate ranking, we utilize

#	User Name	mAUROC ↑	mAUPR ↑	mFPR ↓	mECE ↓	mIoU ↑
1	zuoyi	**0.9516**	0.6053	0.1722	NA	0.4887
2	arui	0.9515	**0.6379**	**0.1499**	0.0726	0.5123
3	bee	0.9149	0.6072	0.2230	0.0698	0.5340
4	MahouShoujo	0.9094	0.3520	0.2391	NA	0.3837
5	rac	0.8510	0.4435	0.3983	0.0603	**0.6454**
6	zachtian	0.8393	0.2196	0.3850	0.1075	0.3274
7	Tonnia	0.8393	0.2745	0.4196	NA	0.4599
8	drop08	0.8356	0.2502	0.3898	0.0942	0.4558
9	Shyam671	0.8341	0.3038	0.3091	**0.0531**	0.3817
10	xtz	0.8280	0.4767	0.3536	0.1540	0.5049
11	Team_Rocket	0.7673	0.1821	0.4720	0.1855	0.3796
12	h.m	0.7644	0.2031	0.4512	0.1657	0.4134
13	HPeter	0.7429	0.1712	0.5204	0.3116	0.3656
14	Janes_migadel	0.7418	0.1806	0.5880	0.1799	0.2848
15	Yuehua.DING	0.7411	0.1641	0.5020	0.2165	0.3668
16	dens03	0.7399	0.1884	0.4787	0.2521	0.3764
17	Baseline	0.7337	0.1790	0.5253	0.2880	0.3690
18	NathLaMenace	0.7337	0.1790	0.5253	0.2880	0.3690
19	Sp4n	0.7091	0.1907	0.7499	0.0970	0.5309

Table 1. Results for valid submissions.

the mAUROC results, recognizing that the solutions might involve trade-offs across various metrics. It is apparent that the majority of teams have outperformed the baseline, particularly the top 9 teams, which have demonstrated notable advancements across nearly all metrics. Figure 2 shows the visualizations on the confidence maps given by the baseline, the 2nd place solution, and the 1st place solution. When different weather conditions occur in the input images, the 1st place solution can much better highlight the OOD examples than the baseline confidence map.

Overview on the submitted solutions In their pursuit of innovative solutions, a significant majority of participants opted for the Deeplab V3+ architecture coupled with Resnet101. However, several participants proposed to modify the architecture or utilized multiple DNNs, as visualized in Figure 1. Furthermore, competitors employed a diverse array of strategies, such as Energy Score, Biased Class disagreement, and temperature scaling approaches. Most prominently, ensembling strategies were extensively employed, accompanied by appropriate data augmentation methodologies aimed at enhancing DNN robustness. For a comprehensive view of the participants' statistical breakdown, refer to Figure 1. Furthermore, detailed insights into the distinct solutions adopted can be found in Section 3.3.

3.3. Top solutions

3.3.1 Solution 1:

Yi Zuo, Zitao Wang, Xiaowen Zhang, Jiaxuan Zhao, Yuting Yang, Licheng Jiao
yzuo_1@stu.xidian.edu.cn

We used Swin-L [33]+Mask2Former [6], InternImage H [48]+Mask2Former, and Segformer+MIT-B5 [52] as base models to complete the corresponding tasks.

Due to the significant category overlap between the MUAD and Cityscapes datasets, we introduced Cityscapes as a pre-trained dataset. Firstly, we trained 200000 iters on three networks on Cityscapes data to obtain corresponding pre-training weights. Subsequently, we will use the pre-trained weights obtained as initial weights and train them on the MUAD dataset. During this process, due to the lack of severe weather conditions in the MUAD test set, we designed a data augmentation method for harsh environments (AE) to simulate this situation.

In AE, we have designed new data augmentation methods for several scenarios, including rain, snow, and darkness. For example, for train data augmentation methods, we first create a random noise image to simulate raindrops. Then rotate and blur the noisy image to simulate the effect of raindrops falling. Simulate the angle and length of raindrops by rotating a diagonal matrix while applying Gaussian blur to the rotated matrix to simulate the blurred edges of raindrops. Finally, combine the raindrop effect with the original image to generate the final image with the raindrop effect. For data augmentation methods of night, we adjust the brightness, contrast, saturation, and hue of the image to simulate the effect of night. In addition, some conventional data augmentations have also been added to AE, such as Random Flip, Random Rotate, Random Crop, Random CutOut [9], etc.

In the experiment, we found that as training increases, although mIoU gradually increases, mAUROC shows a downward trend. We believe this situation is because the model, although fitted on known categories, did not adapt to OOD categories. In order to achieve good results in mIoU and mAUROC, we use two strategies (categories that have not appeared in the train and validation sets) to address them separately. For mIoU, it relies on high accuracy in predicting category results, corresponding to "pred" in the competition. In order to improve mIoU, we need to obtain a model that can predict accurately. Therefore, we will train the three base models on MUAD with 200000 items each. Afterward, the results of the three models were ensembled. The ensemble method is as follows: (1) Use three models to predict on the test set and obtain three prediction results; (2) Vote for each pixel in the three predicted results and select the category with the highest number of votes as the final category. If it is equal, use the result of Segformer+MIT-B5; (3) Gradually iterate until the resulting voting for all test sets is completed.

For mAUROC, we use another strategy. We only fine-tune the three base networks based on Cityscapes pre-training and set a higher learning rate to prevent them from being insensitive to OOD categories. During this process, we found that the Segformer network does not clearly distinguish the boundaries of instance objects. The confidence image obtained from it presents a result similar to a noise map, but its corresponding mAUROC score is very high. For mask2Former, it corresponds to each region with a clear tendency and often clearly separates the boundary lines of the

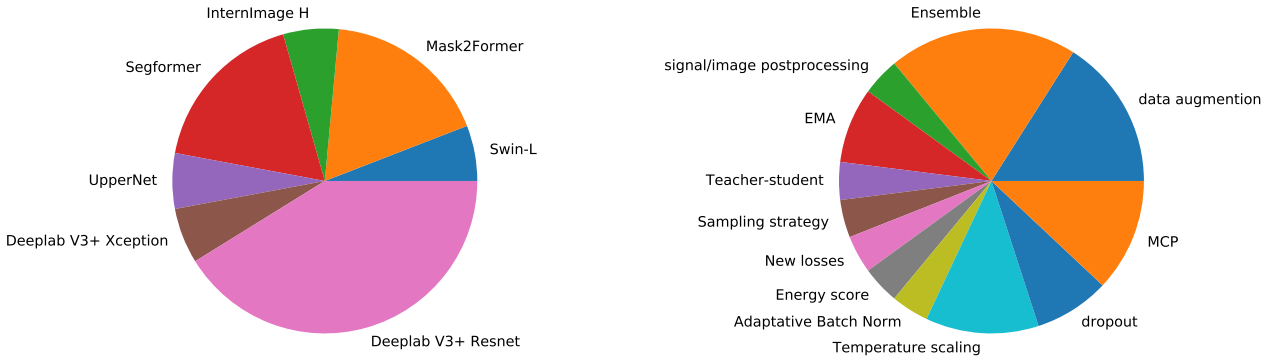


Figure 1. **The statistics of the different solutions.** The pie chart on the left displays a visualization of the diverse architectures employed throughout the challenge, while the chart on the right showcases the array of solution types embraced by the participants.

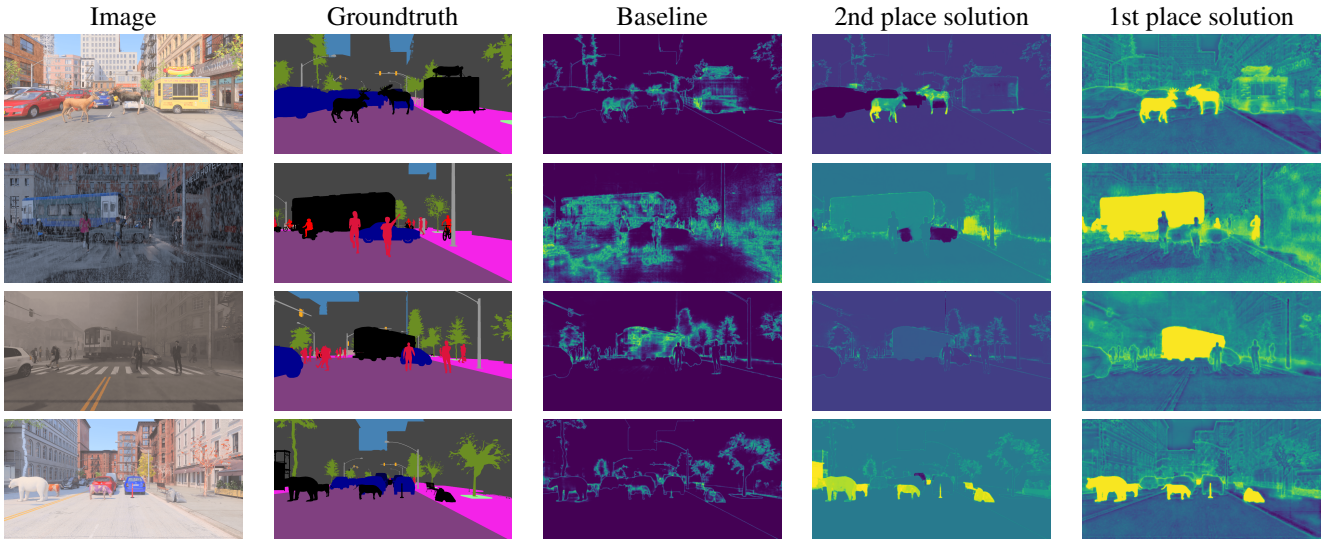


Figure 2. **Illustration of confidence maps on one image of MUAD.** The latter three columns are the confidence maps given by the baseline, the 2nd place solution, and the 1st place solution, respectively. The brighter, the more uncertain the corresponding class prediction is. Note that the class `train`, `bicycle/motorcycle`, `stone`, `stand food`, and the `animals` are OOD examples, which are visualized as black areas on the groundtruth maps.

instance object. In our experiment, its score was lower than Segformer. In order to integrate high-scoring results, we selected the confidence map of Segformer as the background and fused the Mask2former results on top of the Segformer results. Simply put, we first take the reciprocal of the confidence results for Swin-L+Mask2Former and InternImage H+Mask2Former, then add them together, and then take the reciprocal as the fusion result for Mask2Former. Then, select the Mask2Former fusion result with a confidence level lower than 0.6 to directly cover the Segform confidence level result at the corresponding position.

Although this method dramatically improves the score of mAUROC, in order to pursue higher scores, we have designed a region normalization strategy. Specifically, we first use mean filtering for areas with confidence levels below 0.6, then use connected domain algorithms to segment the entire image into multiple mask regions. If more than half of the pixels have a confidence level below 0.4 in a region, we define it as an OOD region. For all pixels in the region, we cover it with the minimum confidence level of the region.

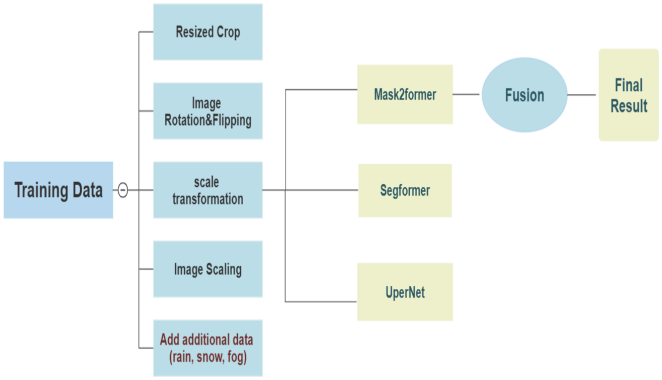


Figure 3. Overall workflow for Solution2.

3.3.2 Solution 2:

Rui Peng, Xinyi Wang, Junpei Zhang, Kexin Zhang, Fang Liu
22171214876@stu.xidian.edu.cn

The overall workflow of our proposed solution is shown in Figure 3. First, we train using the original dataset on three models, Mask2former [6], Segformer [52] and UperNet [51], respectively. Basic data enhancement operations, including image flipping, cropping, rotation, scaling, brightness adjustment, scale transformation, etc., were also performed on the training set. In addition, after analyzing the training, validation, and test sets, we identified data enhancement methods that can effectively improve the performance of the models. We added different degrees of effects of special weather conditions such as rain, snow and fog to the given training set as a way to enhance the robustness and generalization of the models. In the testing phase, we test the results of the three models at different stages of training to produce results. After fusing the results from multiple models, we found that the results were not optimal. Therefore, arithmetic averaging was finally used to fuse different results from the same model to obtain the final class prediction and confidence results to improve the segmentation effect of the model. In the training phase, we use an image size of 512*1024 for training. For each GPU, there are 2 images per batch. In order to reduce overfitting [22], for Mask2former, we set the maximum number of iterations of training to 90000 and the initial learning rate to 6e-5 . For UperNet, we set the maximum number of iterations for training to 26800 with an initial learning rate of 6e-5 . For Segformer, we set the maximum number of iterations for training to 80000 and the initial learning rate to 1.2e-4 . We trained our model on a single NVIDIA RTX 3090 GPU.

3.3.3 Solution 5: Biased Class disagreement

Roberto Alcover-Couso, Juan C. SanMiguel, Marcos Escudero-Viñolo
roberto.alcover@uam.es

Our proposal leverages multiple models with diverse biases, aiming to assign high-confidence predictions to OOD instances by mapping them to the selected prior semantic category. These biases are introduced through the sampling strategy. Additionally, we employ soft cross-entropy loss and confidence filtering to enforce the classification of under-represented classes. We employ an HRDA [21] architecture with the "min pixels" parameter set to 30. Our proposal is described in detail in the following reference [1].

Teacher-student training In order to train each model, we follow a teacher-student protocol, where two DNNs are trained: the teacher network and the student network. The teacher network is updated every time step following an exponential moving average (EMA) of the student network.

The student network is trained by minimizing a soft cross-entropy loss on the labeled images and the pseudo-labels generated by the teacher network of the unlabeled images.

Image Sampling Our main contribution is the employment of sampling strategies that over-sample less frequent classes based on the class frequency ($\mathbf{f}$) in the training set. In order to oversample towards a selected class c', we sample from a Bernoulli distribution, $B(1, .5)$, whether to select if a sample labeled as class c' is incorporated into the training. In the case of *failure*, samples are incorporated with a probability defined by the $softmax$ of the $1 - \mathbf{f}$ frequencies.

Confidence Filtering In order to enhance the classification of low-represented classes, we propose a method that involves switching the pseudo-labels 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. This means that the confidence scores for false positives are generally lower and more spread out, indicating uncertainty and potential misclassifications.

Specifically, 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 the lowest confidence. We employ the mean of the confidences assigned to that class as the threshold for filtering out pixels.

Bias class disagreement ensemble In order to enhance the robustness of our approach, we propose to utilize an output-level ensemble of two 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 as pixels of OOD instances.

3.3.4 Solution 6:

Hanlin Tian, Kenta Matsui, Tianhao Wang, Fahmy Adan
h.tian22@imperial.ac.uk



Figure 4. Data augmentation in Solution6. (Left) Original image from the training dataset. (Right) Corresponding image post-application of our weather data augmentation process.

Data Augmentation We introduced a data augmentation technique by employing a combination of white noise, Gaussian blur, and color alterations to simulate snowfall and its associated environmental changes. Our focus remained primarily on snowy scenarios due to time constraints. Addressing the challenges posed by rain and fog conditions remains a potential avenue for future work.

We introduced white particle-like noise to depict falling snowflakes in the data augmentation process. By varying the size of these particles, we simulated snowflakes appearing at different distances. We also introduced a slight vertical blur to mimic the motion of falling snow. We modified the color palette to colder tones in the following process, replicating the subdued and darker colors often associated with snowy scenes. We randomly selected and augmented 10% of the entire dataset to simulate snowy conditions. Figure 4 shows the example of the output.

Xception Backbone In our approach, we integrated the Xception backbone [8] into our segmentation model, where it serves as a feature extractor. This choice involved replacing our prior ResNet backbone [19] while retaining the DeeplabV3+ head [5] in our model configuration. The high-level features extracted by the Xception backbone are seamlessly fed into the subsequent segmentation layers of our model. The incorporation of the Xception backbone led to a significant improvement in our model's overall performance. Notably, we observed that the mAUROC score, which is a pivotal metric for gauging our model's quality, surged from 0.7861 to 0.8281. This remarkable enhancement underscores the efficacy of the Xception backbone and attests to the substantial advantages of integrating advanced architectures in confronting the challenges of intricate segmentation tasks.

OHEM Cross-Entropy Loss Our custom loss function, named OhemCELoss, is an implementation of Online Hard Example Mining (OHEM) with Cross-Entropy loss. This loss function is designed to focus on the hardest samples to classify, identified by a threshold on the computed loss values, and is particularly beneficial in segmentation tasks where extreme weather conditions lead to challenging pixels.

Initially, the standard Cross-Entropy loss between the logits and ground truth labels is computed without reduction:

$$L_i = -\log\left(\frac{\exp(x_{i,y_i})}{\sum_j \exp(x_{i,j})}\right) \tag{1}$$

where L_i is the loss for the i-th sample, $x_{i,j}$ represents the logits, and y_i is the true label for the i-th sample.

Hard examples are identified as those samples where the loss exceeds a dynamically computed threshold (thr):

$$\text{thr} = \max\left(\text{thr}, \min\left(L_{(\text{sorted})}[\text{min_kept}], L_{(\text{sorted})}[\text{end}]\right)\right) \tag{2}$$

where $L_{(\text{sorted})}$ is the sorted vector of losses.

The final OHEM loss is the mean loss value over these selected hard examples:

$$\text{OHEM Loss} = \frac{1}{N}\sum_{i=1}^{N} L_i \cdot \mathbb{I}(L_i > \text{thr}) \tag{3}$$

where N is the total number of examples, and $\mathbb{I}(\cdot)$ is the indicator function, which is 1 if the condition inside is true, and 0 otherwise.

In practice, this results in the model focusing more intently on the most challenging parts of the images—often, these are regions with intricate textures, object boundaries, or extreme weather effects. By setting an ignore label parameter (e.g., to 255), we explicitly instruct the model to disregard certain regions, which is useful for handling out-of-distribution (OOD) areas in the segmentation task.

3.3.5 Solution 7: Energy Score with BN adaptation

Zhitong Gao, Xuming He
gaozht@shanghaitech.edu.cn

The method adopts the energy score as the base strategy for OOD object detection, which aligns theoretically with the probability density of the inputs and offers better performance than softmax scores [32]. However, applying energy scores directly as a post-hoc approach often suffers from the presence of special weather conditions in the test set.

To overcome this challenge, the method introduces a novel instance-wise test-time domain adaptation technique, which enables the backbone network adapts the batch normalization (BN) layers according to image statistics at test time. Specifically, the overall framework operates in two sequential phases: first, it detects and mitigates the effects of unknown weather conditions through the adaptive BN; next, it performs OOD detection by computing a pixel-wise energy score map of the model's outputs. The entire process unfolds through the following specific steps:

Model Training The method does not modify the training process and can use any standard closed-world segmentation training procedures. In this work, a pretrained Deeplab v3+

with Resnet 101 model, provided by the official benchmark GitHub repository [2], is used.

BN Adaptation When exposed to online test data, the model employs a test-time adaptation module to identify and adapt to special weather conditions that might be present in the data. Inspired by the technique of transductive batch normalization [36], we implement the adaptation by updating the BN statistics of the network. Specifically, the running mean and standard deviation μ^r, σ^r, are mixed with the mean and standard deviation $\mu(x), \sigma(x)$ for each test input x. The updated mean and standard deviation are computed as $\hat{\mu} = \alpha * \mu(x) + (1-\alpha) * \mu^r$, $\hat{\sigma}_l^2 = \alpha * \sigma^2(x) + (1-\alpha) * (\sigma^r)^2$. Here, α is the mixing coefficient, representing the probability that the input image originates from an unknown weather condition. This coefficient is estimated by considering the Kullback–Leibler (KL) divergence between the feature distributions of the test image and the training images within the Normalization layers: $KL(N(\mu(x), \sigma(x)) || N(\mu^r, \sigma^r))$. The mixing coefficient α is calculated by averaging the KL divergence values over all BN layers and normalizing the result to a 0-1 range using the sigmoid function.

Prediction and OOD detection After the BN Adaptation, the method proceeds to compute the segmentation network's predictions along with an energy score as the basis for OOD detection. Specifically, the logits of the model output are denoted as η_i, where i identifies the pixel. The inlier prediction is obtained through: $\arg\max_c \exp(\eta_i^c) / \sum_n^{N_c} \exp(\eta_i^n)$, and the energy score is calculated using the energy function: $\log \sum_n^{N_c} \exp(\eta_i^n)$. This score represents a quantitative measure of the model's certainty regarding its predictions and is employed to detect out-of-distribution (OOD) instances within the test data.

Overall, the method synergizes the energy scores with a BN adaptation strategy to address the unique challenges posed by special weather conditions, which provides a robust and effective solution for OOD detection in the wild.

3.3.6 Solution 8: Polynomial Calibration

Quentin Bouniot, Hossein Moghaddam
quentin.bouniot@telecom-paris.fr

We apply our technique on the baseline model, i.e. DeepLabV3+ [5] with ResNet101 backbone [19]. Among the techniques employed to quantify uncertainty, one that consistently demonstrates commendable performance involves the estimation of the posterior distribution $P(\boldsymbol{\theta}|\mathcal{D})$ of the network [3, 12, 34]. Subsequently, N DNNs are sampled according to this posterior, denoted as $F_{\boldsymbol{\theta}_j}(\cdot)_{j=1}^N$, and their performance is averaged as follows:

$$\hat{\mathbf{y}}_i = P(Y = \mathbf{y}_i | X = \mathbf{x}_i) = 1/N \sum_j F_{\boldsymbol{\theta}_j}(\mathbf{x}_i) \quad (4)$$

The Deep Ensembles [26] (DEns) technique simplistically estimates this posterior, often yielding impressive results due to its capability, highlighted in [50], to estimate diverse modes of the posterior. However, the training of multiple DNNs makes DEns challenging. Various strategies have been proposed to streamline DEns training [11, 18, 27, 49], but due to the time-intensive nature of training, the number of samples for ensemble training remains limited. Alternatively, the Monte Carlo (MC) dropout [15] method estimates a singular mode of the posterior, offering a more accessible solution as it requires only one training. Consequently, this approach can furnish a larger sample set. However, both these techniques introduce increased complexity to the inference process.

Smith and Gal [45] interestingly amalgamated these two Bayesian Neural Network approximations to develop Dropout Ensembles (DropEns), reaping the benefits of both preceding techniques. Nonetheless, DropEns may encounter calibration prediction issues. Several calibration techniques for trained DNN predictions have been explored [24, 41, 44] to address this issue, commonly referred to as post-hoc calibration. Among these, we evaluated temperature scaling [17] (TS), a method involving a temperature parameter $\tau \geq 0$ to the DNN logits. Specifically, if we denote $F_{\boldsymbol{\theta}_j}(\mathbf{x}_i) = \text{soft}\left(G_{\boldsymbol{\theta}_j}(\mathbf{x}_i)\right)$, where soft$(\cdot)$ represents the softmax function and $G_{\boldsymbol{\theta}_j}(\cdot)$ signifies the logit of $F_{\boldsymbol{\theta}_j}(\cdot)$, temperature scaling entails tuning τ to achieve the optimal ECE for $F^{\tau}_{\boldsymbol{\theta}_j}(\mathbf{x}_i) = \text{soft}\left(G_{\boldsymbol{\theta}_j}(\mathbf{x}_i)/\tau\right)$.

Typically, any monotonically increasing function can be employed for this purpose. In our approach, we use a polynomial function, considering that we normalize the logits by subtracting the minimum value for each data point to ensure their positivity. Thus, the normalized logit becomes $G^{\text{norm}}_{\boldsymbol{\theta}_j}(\cdot)$. We then apply the polynomial temperature scaling (PTS) function incorporating parameters τ_1, τ_2, and τ_3, leading to the following DNNs:

$$F^{\tau_1,\tau_2,\tau_3}_{\boldsymbol{\theta}_j}(\mathbf{x}_i) = \text{soft}\Big(G^{\text{norm}}_{\boldsymbol{\theta}_j}(\mathbf{x}_i)/\tau_1 + (G^{\text{norm}}_{\boldsymbol{\theta}_j}(\mathbf{x}_i)/\tau_2)^2 + (G^{\text{norm}}_{\boldsymbol{\theta}_j}(\mathbf{x}_i)/\tau_3)^3\Big) \quad (5)$$

An overview of our findings is presented in Tab 2. It's evident that DEns and standalone MC Dropout exhibit commendable performances. However, they fall short of effectively capturing the underlying uncertainty inherent in the DNNs. Notably, the DropEns outperforms its counterparts in terms of performance, and its calibrated version further enhances its effectiveness.

[2] https://github.com/ENSTA-U2IS/DeepLabV3Plus-MUAD-Pytorch

Technique	mAUROC ↑	mAUPR ↑	mFPR ↓	mECE ↓	mIoU↑
Baseline	0.7337	0.1790	0.5253	0.2880	0.3690
MC Dropout	0.7794	0.1778	0.4791	0.1380	0.4358
DEns	0.7399	0.1884	0.4787	0.2521	0.3764
DropEns	0.8195	0.2243	0.4015	0.1093	0.4547
DropEns+TS	0.8219	0.2261	0.3988	0.1005	0.4544
DropEns+PTS	**0.8356**	**0.2502**	**0.3898**	**0.0942**	**0.4558**

Table 2. Ablation study for Solution8.

3.3.7 Solution 9: Uncertainty Calibrated Mask2Former

Shyam Nandan Rai, Fabio cermelli and Carlo Masone
shyam.rai@polito.it

Mask architectures [6, 7] have proven to be an effective method for universal segmentation and anomaly segmentation [38, 42]. However, these networks are not well calibrated, resulting in poor uncertainty estimates. Hence, we calibrate the Mask2Former architecture [6] to give better uncertainty estimates and identify OOD objects. For this purpose, we introduce the following modifications to the architecture: a) during training, we change the loss function from cross-entropy to focal loss [30] as it allows us to learn well-calibrated models, and b) at inference, we propose a way to perform temperature scaling in mask-architecture. After integrating modification into mask architecture, we call the solution Uncertainty Calibrated Mask2Former (UC-M2F). It is also important to note that our method uses no additional OOD training data. In the subsequent sub-sections, we will delve into each of these modifications.

Preliminary Mask2Former consists of three major parts: a) a *backbone* acting as a feature extractor, b) a *pixel-decoder* that upsamples the low-resolution features from the backbone to produce high-resolution *per-pixel embeddings*, and c) a *transformer decoder*, that takes the image features to output a fixed number of object queries consisting of *mask embeddings* and their associated *class scores* $C \in \mathbb{R}^{N \times N_C}$. The final *class masks* $M \in \mathbb{R}^{N \times (H \times W)}$ are obtained by multiplying the mask embeddings with the per-pixel embeddings obtained from the pixel-decoder. N represents the number of output masks.

Training Loss Mask2Former is trained with a cross-entropy loss to predict C and both a binary cross-entropy and a dice loss to predict M. However, [35] points out that training deep networks on standard cross-entropy loss results in miscalibrated network and suggests using focal loss [30] can lead to a very well-calibrated network. So, we change the training loss for the class scores to focal loss

$$L_{fl}(x_i) = -\alpha(1 - \hat{y}^i)^\gamma \log(\hat{y}^i) \tag{6}$$

where α and γ are the scaling and focusing factors, respectively.

Temperature scaling Temperature scaling is another way to improve the calibration of a network without affecting accuracy. However, temperature calibration in Mask2Former is not trivial due to the nature of its output calculation given:

$$g(x_i) = \max^{N_C} \left(\text{softmax}(C)^T \cdot \text{sigmoid}(M)\right) \tag{7}$$

We now have three options to calibrate the network: dividing C by temperature, dividing M by temperature, or both. We found that dividing C by the temperature gives the best results. Formally, eq. 7 becomes:

$$g(x_i) = \max^{N_C} \left(\text{softmax}(C/t)^T \cdot \text{sigmoid}(M)\right) \tag{8}$$

Training Details Our UC-M2F architecture uses Swin-B as the feature backbone, and the decoder is kept the same as Mask2Former. The network is trained with an initial learning rate of 1e-4 and batch size of 8 for 90 thousand iterations on AdamW with a weight decay of 0.05. Then, fine-tuned for 5 thousand iterations having a learning rate of 1e-5. We use an image crop of 400 $\times$ 800 with large-scale jittering along with a random scale ranging from 0.1 to 2.0. In the focal loss, γ is kept at 2, and α is 10. The best temperature value is found to be 0.5.

4. Conclusion

Results of UNCV2023 challenge were presented. The main theme of the challenge is to evaluate the uncertainty quantification performance of the semantic segmentation models based on the MUAD dataset. The goal is to discover uncertainty quantification solutions in autonomous driving scenarios and assess the impact of different sources of uncertainty on model performance.

As per the solutions submitted by participants, a variety of model architectures, backbones, data augmentation techniques, and model ensembles emerged as prevalent choices among most teams. Additionally, approaches involving region normalization, novel loss functions, test time batch norm adaptation, and model calibration were also brought into the spotlight. These diverse strategies collectively offer valuable insights into the practical deployment of robust semantic segmentation algorithms and the concurrent quantification of uncertainty within intricate urban environments.

We view these approaches as a repository of strategies akin to a "bag of tricks". The prospect of optimizing the synergy among these methods and exploring potential conflicts between them holds significant promise for future research endeavors.

Acknowledgements

We gratefully acknowledge the support of DATAIA Paris-Saclay institute and AID Project ACoCaTherm which supported the creation of the dataset.

References

[1] Roberto Alcover-Couso, Juan C. SanMiguel, and Escudero-Viñolo Marcos. Biased class disagreement: detection of out of distribution instances by using differently biased semantic segmentation models. In *ICCV Workshop*, 2023. 5

[2] Victor Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Triggering failures: Out-of-distribution detection by learning from local adversarial attacks in semantic segmentation. In *ICCV*, 2021. 2

[3] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *ICML*, 2015. 7

[4] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Mathieu Salzmann, Pascal Fua, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. *NeurIPS*, 2021. 1, 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2, 6, 7

[6] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 3, 5, 8

[7] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *Neural Information Processing Systems*, 2021. 8

[8] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 6

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. In *arXiv*, 2017. 3

[10] Giancarlo Di Biase, Hermann Blum, Roland Siegwart, and Cesar Cadena. Pixel-wise anomaly detection in complex driving scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16918–16927, 2021. 2

[11] Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13539–13548, 2021. 7

[12] Gianni Franchi, Andrei Bursuc, Emanuel Aldea, Séverine Dubuisson, and Isabelle Bloch. TRADI: Tracking deep neural network weight distributions. In *ECCV*, 2020. 7

[13] Gianni Franchi, Xuanlong Yu, Andrei Bursuc, Emanuel Aldea, Severine Dubuisson, and David Filliat. Latent discriminant deterministic uncertainty. In *ECCV*, 2022. 2

[14] 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 *BMVC*, 2022. 1, 2

[15] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. 7

[16] Matej Grcić, Petra Bevandić, Zoran Kalafatić, and Siniša Šegvić. Dense anomaly detection by robust learning on synthetic negative data. *arXiv preprint arXiv:2112.12833*, 2021. 2

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *ICML*, 2017. 7

[18] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M Dai, and Dustin Tran. Training independent subnetworks for robust prediction. *arXiv preprint arXiv:2010.06610*, 2020. 7

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 6, 7

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017. 2

[21] Lukas Hoyer, Dengxin Dai, and Luc Van Gool. Hrda: Context-aware high-resolution domain-adaptive semantic segmentation. In *ECCV*. Springer, 2022. 5

[22] Lukas Hoyer, Dengxin Dai, Haoran Wang, and Luc Van Gool. Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11721–11732, 2023. 5

[23] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *BMVC*, 2017. 2

[24] Amr Khalifa and Ibrahim Alabdulmohsin. Improving the post-hoc calibration of modern neural networks with probe scaling. 2021. 7

[25] Tchoya Florence Koné, Eric Bonjour, Eric Levrat, Frédérique Mayer, and Stéphane Géronimi. Safety demonstration of autonomous vehicles: A review and future research questions. In *Complex Systems Design & Management: Proceedings of the Tenth International Conference on Complex Systems Design & Management, CSD&M Paris 2019*, pages 176–188. Springer, 2020. 1

[26] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. 7

[27] Olivier Laurent, Adrien Lafage, Enzo Tartaglione, Geoffrey Daniel, Jean-Marc Martinez, Andrei Bursuc, and Gianni Franchi. Packed-ensembles for efficient uncertainty estimation. In *International Conference on Learning Representations*, 2023. 7

[28] Jamy Li, Xuan Zhao, Mu-Jung Cho, Wendy Ju, and Bertram F Malle. From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Technical report, SAE Technical Paper, 2016. 1

[29] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018. 2

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 8

[31] Krzysztof Lis, Krishna Nakka, Pascal Fua, and Mathieu Salzmann. Detecting the unexpected via image resynthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2152–2161, 2019. 2

[32] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020. 6

[33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 3

[34] Wesley J Maddox, Timur Garipov, Dmitry Vetrov, Andrew Gordon Wilson, et al. A simple baseline for bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019. 7

[35] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip HS Torr, and Puneet K Dokania. Calibrating deep neural networks using focal loss. In *NeurIPS*, 2020. 8

[36] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 7

[37] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *AAAI*, volume 29, 2015. 2

[38] Nazir Nayal, Mısra Yavuz, João F. Henriques, and Fatma Güney. Rba: Segmenting unknown regions rejected by all. In *IEEE/CVF International Conference on Computer Vision*, 2023. 8

[39] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. 2

[40] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *IROS*, 2016. 1

[41] Amir Rahimi, Thomas Mensink, Kartik Gupta, Thalaiyasingam Ajanthan, Cristian Sminchisescu, and Richard Hartley. Post-hoc calibration of neural networks by g-layers. *arXiv preprint arXiv:2006.12807*, 2020. 7

[42] Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, and Barbara Caputo. Unmasking anomalies in road-scene segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2023. 8

[43] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *ICCV*, 2021. 1, 2

[44] Andy Shih, Dorsa Sadigh, and Stefano Ermon. Long horizon temperature scaling. *arXiv preprint arXiv:2302.03686*, 2023. 7

[45] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *arXiv preprint arXiv:1803.08533*, 2018. 7

[46] Tao Sun, Mattia Segu, Janis Postels, Yuxuan Wang, Luc Van Gool, Bernt Schiele, Federico Tombari, and Fisher Yu. SHIFT: a synthetic driving dataset for continuous multi-task domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21371–21382, June 2022. 1

[47] Yu Tian, Yuyuan Liu, Guansong Pang, Fengbei Liu, Yuanhong Chen, and Gustavo Carneiro. Pixel-wise energy-biased abstention learning for anomaly segmentation on complex urban driving scenes. In *European Conference on Computer Vision*, pages 246–263. Springer, 2022. 2

[48] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023. 3

[49] Yeming Wen, Dustin Tran, and Jimmy Ba. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In *ICLR*, 2020. 7

[50] Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. *arXiv preprint arXiv:2002.08791*, 2020. 7

[51] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 5

[52] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 3, 5

[53] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 1