DrIFT: Autonomous Drone Dataset with Integrated Real and Synthetic Data, Flexible Views, and Transformed Domains (Supplementary Material)

Fardad Dadboud[1], Hamid Azad[1], Varun Mehta[2], Miodrag Bolic[1], Iraj Mantegh[2]
[1]University of Ottawa, [2]National Research Council Canada

1. Introduction

This supplementary material contains important information that could not be included in the main paper due to space constraints and aims to support the discussions in the main paper. The structure of the main paper is followed.

2. DrIFT Dataset

2.1. Dataset Characteristics and Statistics

Fig. 1 in the main paper displays a variety of backgrounds from our dataset, including the sky, trees, and ground during three distinct seasons (fall, winter, and summer) or adverse weather conditions (foggy, snowy, and rainy). Fig. 1 demonstrates that the DrIFT possesses $47,991$ image frames. As discussed in Subsec. 3.1 "The DrIFT Story" in the main paper, we attempted to keep the balance between training and validation sets for almost all domains: $3,000$ frames for training and 300 frames for validation. This standard practice facilitates a proper platform for evaluating the UDA algorithms. Fig. 2 shows the number of existing background samples in each domain. It is important to note that, as shown in the last three rows of Fig. 2, the dataset includes only a validation set for the aerial-real domains, without a corresponding training set. Additionally, it is noteworthy that our adverse weather domains only contain a sky background. Hence we have avoided reporting metrics for tree and ground backgrounds within these domains in Tab. 1.

Fig. 3 depicts drones' relative size and location distribution in real and synthetic domains. The center point, width, and height are normalized to the image width and height. For the aerial-real data in Fig. 3c, the means of the relative width and height are approximately 0.015, whereas in Fig. 3d, representing aerial-synthetic data, these values are around 0.02 and 0.015, respectively. The width and height of the ground-real, illustrated in Fig. 3g, are about 0.03, although for the ground-synthetic shown in Fig. 3h, these are approximately 0.02. These numbers indicate that we deal with extremely small objects in comparison to other applications, *e.g.*, autonomous land vehicles [2, 11]. It makes DrIFT more challenging in terms of training the detector models.

3. DrIFT Benchmark

3.1. Methodology

3.1.1 Uncertainty Estimation

In [8, 9], the researchers employed a gradient function and introduced the concept of self-learning gradient as a metric to evaluate the uncertainty of each detection. If we consider the supervised learning scenario the gradient of the loss function for each detection is $g(\boldsymbol{X}_\mathrm{i}, \boldsymbol{y}_\mathrm{i}^\mathrm{j}) = \nabla_{\boldsymbol{\omega}} \mathcal{L}(\hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{j}, \boldsymbol{y}_\mathrm{i}^\mathrm{j})$. The $\boldsymbol{\omega}$ is the network's weight vector. If the ground truth, $\boldsymbol{y}_\mathrm{i}^\mathrm{j}$, is replaced with detection, $\hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{j}$, and the detection is replaced with its candidates, $\hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{k}$, the self-gradient metric would be

$$g^{cand}(\boldsymbol{X}_\mathrm{i}, \hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{j}) = \sum_{\hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{k} \in \mathbb{C}_{\hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{j}}} \nabla_{\boldsymbol{\omega}} \mathcal{L}(\hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{k}, \hat{\boldsymbol{y}}_\mathrm{i}^\mathrm{j}). \tag{1}$$

The self-gradient metric, $g^{cand}(.)$, referred to as **Grad-loss**, operates as a characteristic that signifies the degree of epistemic uncertainty, which is the focal point for investigating DS. **Grad-loss-localization** is called the corresponding localization term of the loss, although **Grad-loss-classification** points to the classification term in the loss. Nevertheless, it does not inherently encompass the true essence of uncertainty. To consider other methods, we employ a technique based on MC-dropout to capture the inherent uncertainty associated with each detection. In this approach, we activate dropout at inference time and run our model for N_{mcdo} times. Let us consider the output of the model at each iteration $\hat{\mathbb{Y}}_\mathrm{i}^\mathrm{m}$ where $\mathrm{m} \in \{1, \ldots, \mathrm{N}_{mcdo}\}$. Initially, we create an $|\hat{\mathbb{Y}}_\mathrm{i}^1|$-length list corresponding to all output detections in the first iteration. Subsequently, we perform some NMS like the one in Eq. 1 of the main paper to have a candidate list and assign the best candidate with the highest IoU to each list. A detection in each iteration is only allowed to be a member of one list, and a new list is created if there is no option with a higher IoU threshold.

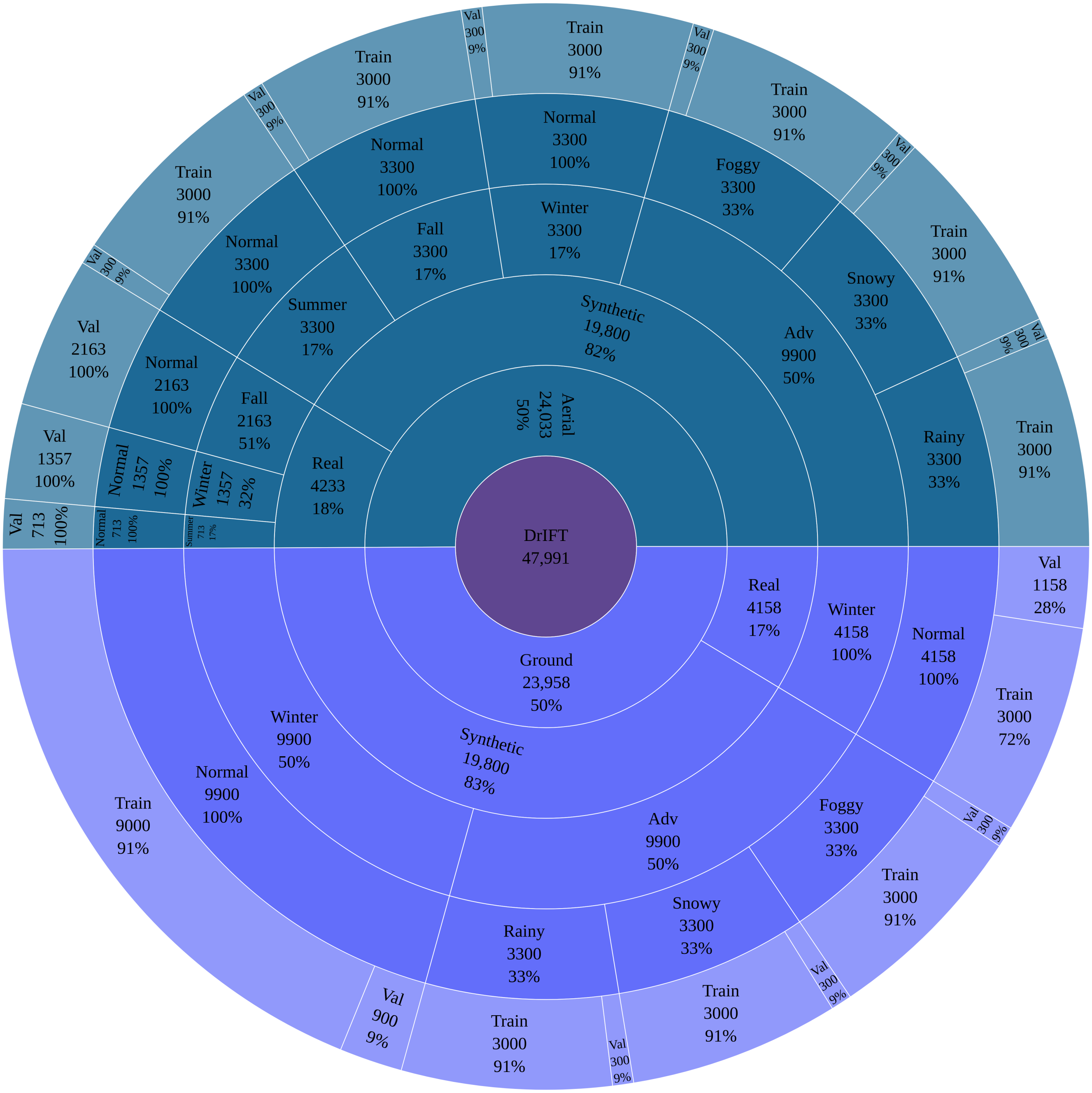


Figure 1. Hierarchical sunburst chart of the DrIFT dataset: The DrIFT dataset contains aerial and ground views in real-world and simulated environments. There are numerous domains based on the various seasons and weather. The chart displays the number and percentage of the samples within the parent category. Adv: adverse

Ultimately, we calculate the standard deviation of localization parameters $\sigma_{\boldsymbol{b}}^{\rm q}$ and the entropy of the mean of the classification probability vector $H_{cls}^{\rm q}$, $\rm q \in \{1, \ldots, N_{mcdo\_out}\}$, respectively. If we assume we have N_{mcdo_out} lists of outputs, we can compute the uncertainty for each list as follows:

$$\sigma_b^{\rm q} = \sqrt{\frac{1}{\rm c_q}\sum_{\rm n=1}^{\rm c_q}(\boldsymbol{b}_n - \bar{\boldsymbol{b}})^2}, H_{cls}^{\rm q} = -\sum_{\rm n=1}^{\rm c_q} \bar{s}_n * \log \bar{s}_n. \quad (2)$$

Here, $\rm c_q$ is the number of members in each list, q is the index of each list, and $\bar{.}$ denotes the mean of the underly-

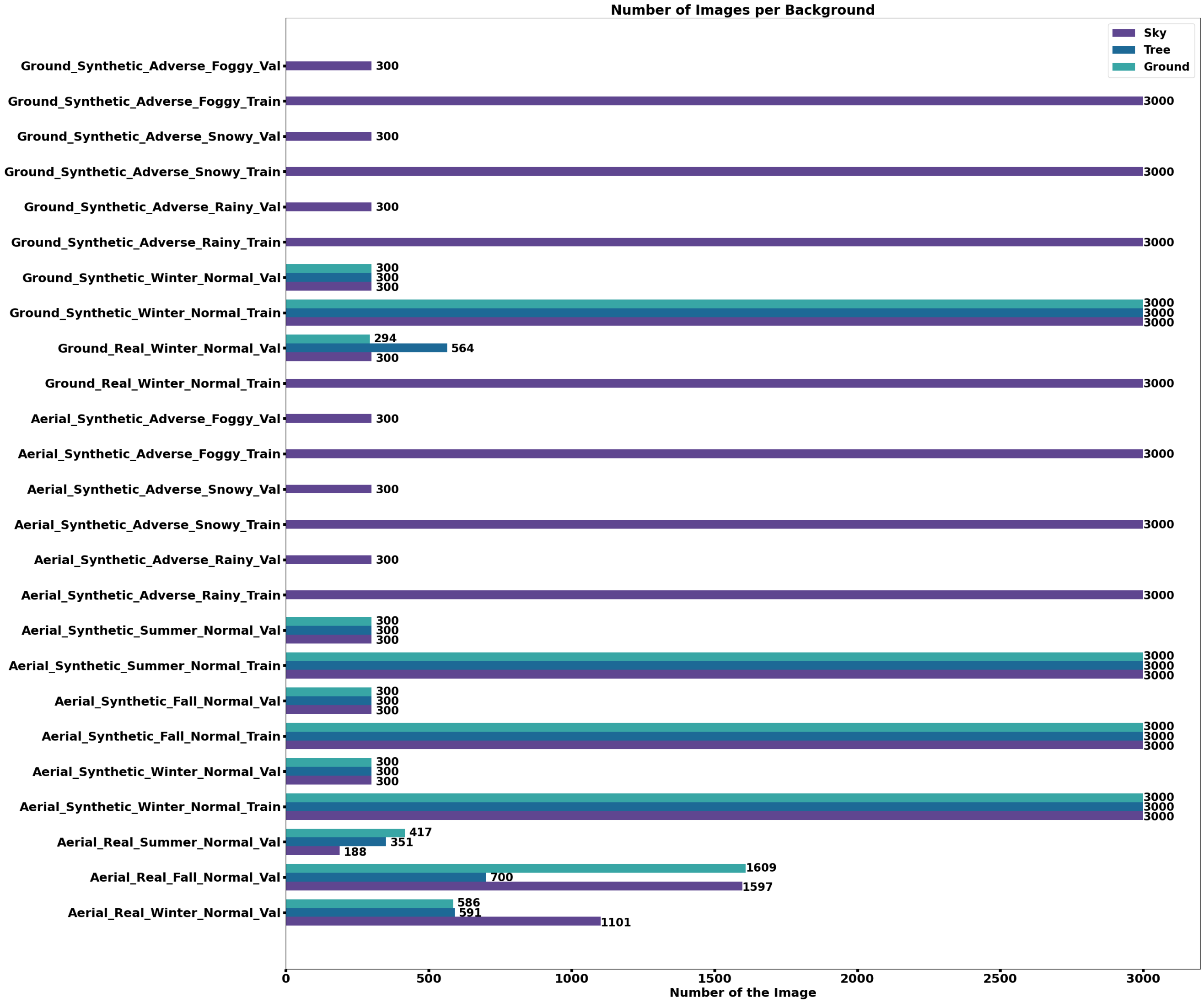


Figure 2. Number of existing background samples in each domain: We aim to maintain an equal number of background samples for each domain's training and validation sets, unless the distribution of real data prevents us from adhering to this guideline.

ing variable. This technique is referred to as **MCDO-NMS** which is divided into **MCDO-NMS-localization**, referred to as σ^{q}_{b}, and **MCDO-NMS-classification**, referred to as H^{q}_{cls} in Eq. (2). Inspired by [6], which suggests averaging individual uncertainties as one possible aggregation solution, we take a weighted average of classification entropy. Similarly, we sum the square residuals of localization parameters, take a weighted average, and calculate the square root at the end.

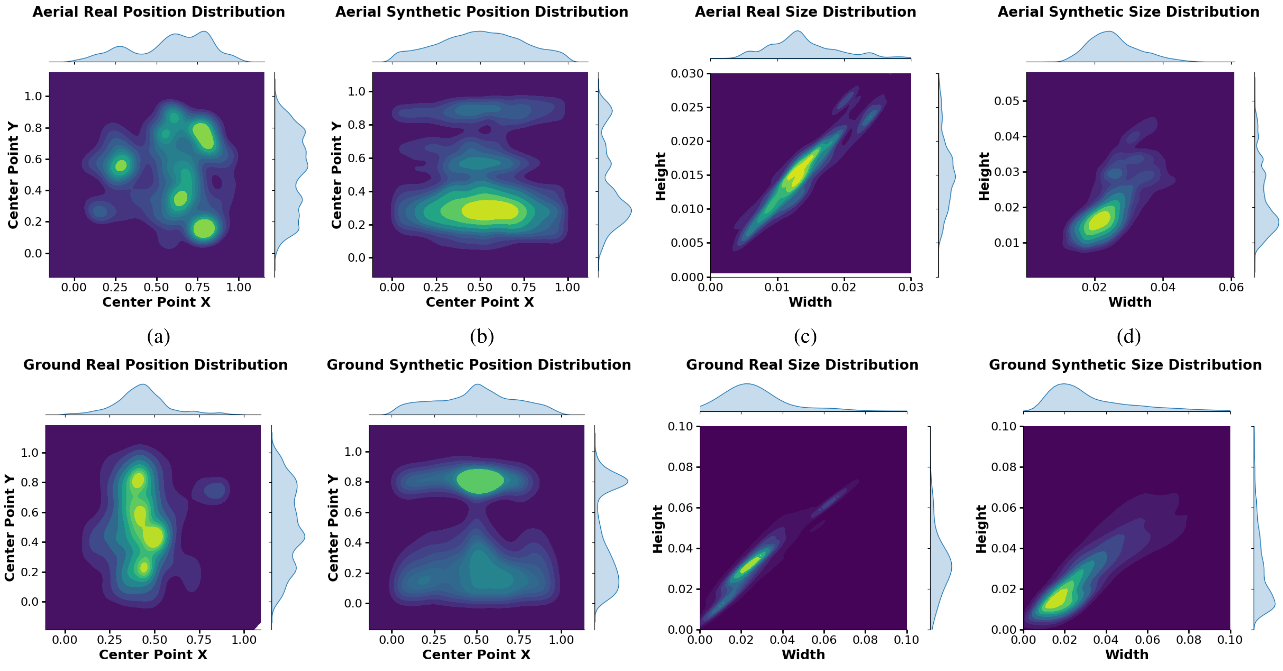
3.2. Benchmark Scenarios

3.2.1 Normalization

Normalization has been done for each metric by subtracting a reference value, which is the value of the metric for the source domain, and then dividing by the same reference value. The normalization is mathematically expressed by

$$M^{\mathrm{i}}_{\mathrm{norm}} = \frac{M^{\mathrm{i}} - M_{\mathrm{r}}}{M_{\mathrm{r}}}. \tag{3}$$

For a set of values of a metric $\mathcal{M} = \{M^1, M^2, \ldots, M^{\mathrm{n}}\}$ and corresponding reference value M_{r}, the normalized set is $\mathcal{M}_{\mathrm{norm}} = \{M^1_{\mathrm{norm}}, M^2_{\mathrm{norm}}, \ldots, M^{\mathrm{n}}_{\mathrm{norm}}\}$. The subtraction of the reference value M_{r} ensures that the data is centered around zero, and the subsequent division by M_{r} scales the data, making it comparable or suitable for further analysis. Fig. 4 is illustrated using normalized values of different metrics. All metrics are normalized to their values for the source domain. Positive values indicate increases.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3. Center point (x, y), height, and width distributions through training sets for both aerial and ground view with a separation for both real and synthetic parts of our dataset. The height and width distributions show the relatively small size of the objects available in DrIFT.

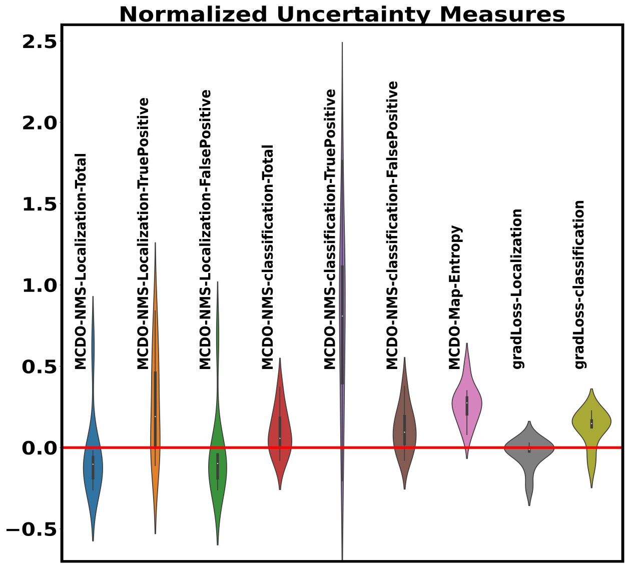


Figure 4. Violin plot of uncertainty metrics: The metrics are normalized to the source domain's uncertainty. MCDO-map always shows positive values, increasing along with DSs, while most of the metrics show negative values.

3.2.2 Violin box plot

Violin box plot [3] is a graphical representation that combines aspects of both box plots and kernel density plots. It provides a concise and informative way to visualize the distribution, central tendency, and spread of a variable. In the violin box plot:

- The central box represents the interquartile range (IQR) of the data, with the line inside indicating the median.
- The ”violin” shape surrounding the box displays the probability density function of the data, providing insights into the distribution's shape.
- Wider sections of the violin indicate higher data density, while narrower sections represent lower density.
- Outliers, if any, are often displayed as individual points.

3.2.3 Pearson correlation coefficient

Pearson correlation coefficient [1], denoted by ρ, is a measure of the linear relationship between two variables $\mathcal{M}^1$ and $\mathcal{M}^2$. It is defined as the ratio of the covariance of $\mathcal{M}^1$ and $\mathcal{M}^2$ to the product of their standard deviations,

$$\rho_{\mathcal{M}^1\mathcal{M}^2} = \frac{cov(\mathcal{M}^1, \mathcal{M}^2)}{\sigma_{\mathcal{M}^1}\sigma_{\mathcal{M}^2}}. \tag{4}$$

The Pearson correlation coefficient ranges from -1 to 1. A value of 1 indicates a perfect positive linear relationship, 0 indicates no linear relationship, and -1 indicates a perfect negative linear relationship. Positive values indicate that as one variable increases, the other variable tends to increase as well. Negative values indicate that as one variable increases, the other variable tends to decrease. This coefficient has been used in Fig. 3 of the main paper to analyze the relationships between different metrics.

3.2.4 Kullback-Leibler (KL) divergence

KL divergence [4] is a measure of how one probability distribution diverges from a second, expected probability distribution. In the DrIFT benchmark, it serves as a metric to quantify the distance between feature map distributions of different domains with the source domain, which is ground-synthetic-winter-normal-sky. The KL divergence is defined

$$D_{\text{KL}}(\boldsymbol{FMD}_{\text{target}} || \boldsymbol{FMD}_{\text{source}}) = \sum_{\text{i}}^{\text{N}} \boldsymbol{FMD}^{\text{i}}_{\text{target}} \log\left(\frac{\boldsymbol{FMD}^{\text{i}}_{\text{target}}}{\boldsymbol{FMD}^{\text{i}}_{\text{source}}}\right), \tag{5}$$

in which N is the cardinality of the feature map distributions, $|\boldsymbol{FMD}_{\text{target}}|$. We assume the two distributions have the same size, $|\boldsymbol{FMD}_{\text{target}}| = |\boldsymbol{FMD}_{\text{source}}|$. $\boldsymbol{FMD}_{\text{target}}$ is the feature map distribution of each domain that is taken as the target domain, and $\boldsymbol{FMD}_{\text{source}}$ is the source domain's feature map distribution. i is the index of existing elements in each domain's feature map distribution.

3.3. Experiments and Results

The ground-synthetic-winter-normal-sky is taken as the source domain all over the paper and supplementary material unless we specify other domains.

3.3.1 Implementation Details

For the object detector in this work, the Faster R-CNN [7] architecture with a VGG16 [10] in the mmdetection platform [5] has been utilized. For generalization and MC-dropout uncertainty evaluation implementation, the dropout has been activated within the VGG. The experiments were run on a Desktop with a Geforce RTX 3090 and a High-performance computing cluster providing 4 x NVidia A100 (40 GB memory). For the vanilla network training that was started from scratch, we used a stochastic gradient descent optimizer for 73 epochs, for which the learning rate was 0.24 for a batch size of 6 on each GPU. For adaptation training, the vanilla network is used as the pre-trained weights. The learning rate has been decreased to 10^{-5}, and the discriminator, which is a simple convolutional neural network, has been trained by an Adam optimizer with a learning rate of 10^{-6}. The codes and details will be available on an online platform.

The objective of Tab. 1 and Tab. 2 is to compare our uncertainty estimation method, **MCDO-map**, with various uncertainty estimation metrics mentioned in the main paper. In Tab. 1, the source domain was ground-synthetic-winter-normal-sky, while ground-real-winter-normal-sky served as the source domain in Tab. 2. To provide a comprehensive explanation, we utilized Fig. 6 to discover a meaningful relation between different uncertainty evaluation metrics, AP, D-ECE, and KL divergence metric (which measures the distance between feature map distributions of different domains relative to the source domain) using the Pearson correlation coefficient. The findings in Fig. 6 could be summarized as follows:

- MCDO-map exhibits the highest positive correlation (0.81) with KL divergence, indicating its superior capability to capture DSs. A greater level of shift, reflected by increased distance or KL divergence, correlates with higher values of MCDO-map.
- As an uncertainty evaluation metric, a negative correlation with AP is expected, implying that higher AP values correspond to lower uncertainty levels. In this context, MCDO-NMS-Loc-Total, MCDO-NMS-Loc-FP, and MCDO-map yield the best results.
- Positive correlations between D-ECE and most uncertainty evaluation metrics suggest that increased uncertainty tends to coincide with calibration errors.
- A positive correlation between D-ECE and AP indicates that even with higher AP values, the model may exhibit over or under-confidence, compromising its reliability.
- Positive correlations between D-ECE and most uncertainty evaluation metrics, such as 0.36 for MCDO-map, suggest that higher levels of uncertainty are associated with calibration errors.

Consequently, MCDO-map emerges as a wise choice for our UDA algorithm to capture DSs effectively.

To enhance the understanding of our results, we present three examples of the outputs generated by the trained Faster R-CNN model on the ground-synthetic-winter-normal-sky domain, depicted in Fig. 7. In Fig. 7a, the drone with a sky background exhibits low uncertainty, as indicated by the blue bounding box on the entropy map. We observe non-zero std values only at the edge of the bounding box in the std map (inside blue, red at the edge). However,

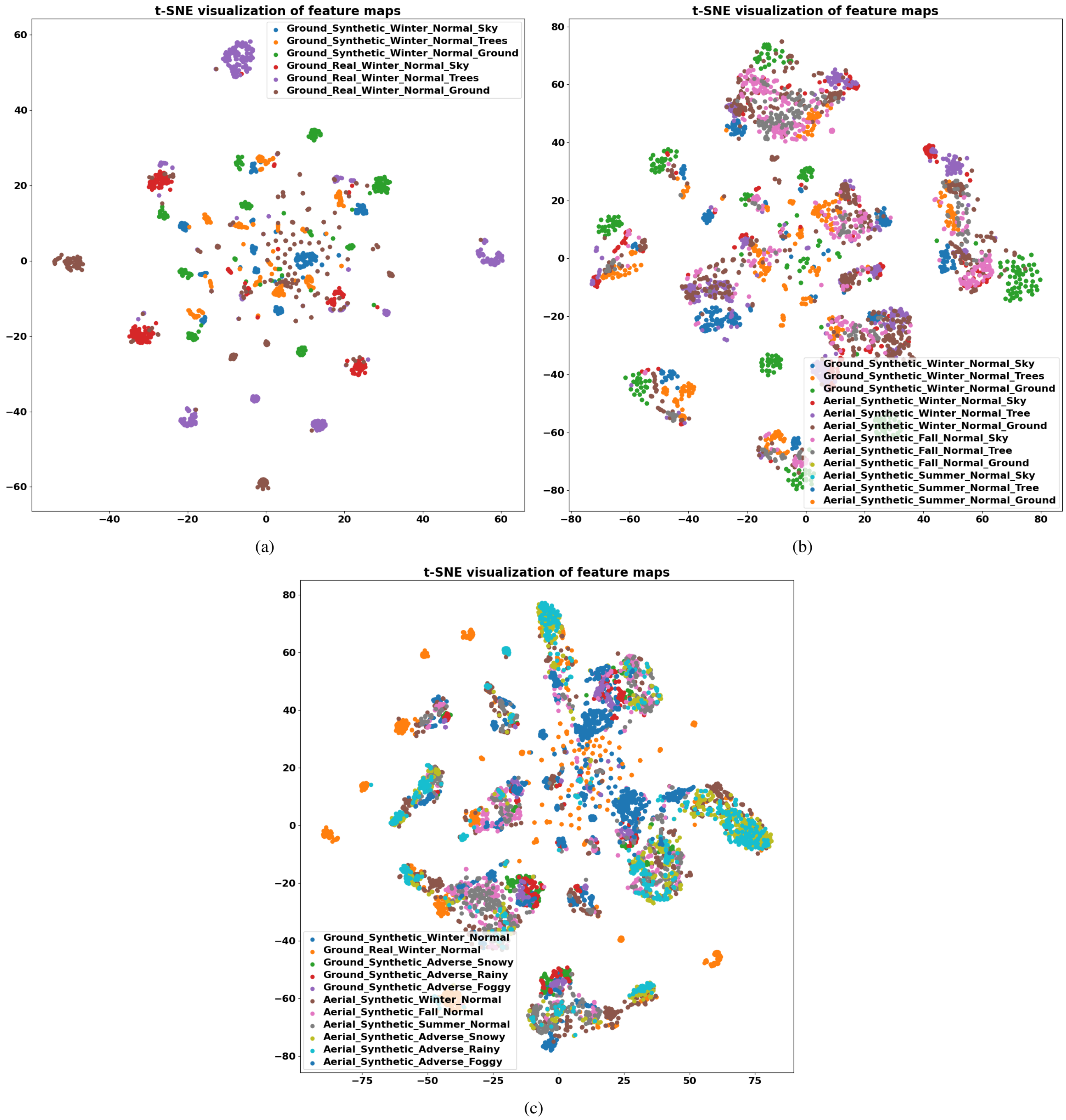
Validation Domain					MCDO-NMS×10^{-3}						MCDO-Map	grad-loss×10^{-3}	
					Localization			Classification			×10^{-4}	Loc.	Cls.
View	Source	Season	Weather	BG	Total	TP	FP	Total	TP	FP			
ground	synthetic	winter	normal	-	107	83	107	457	354	459	3582	433	610
ground	synthetic	winter	normal	sky	107	63	107	436	141	439	-	445	564
ground	synthetic	winter	normal	tree	108	93	108	512	523	512	-	451	680
ground	synthetic	winter	normal	ground	106	104	107	440	476	422	-	397	670
ground	real	winter	normal	-	183	56	186	492	313	496	5355	477	708
ground	real	winter	normal	sky	173	56	180	495	298	508	-	434	692
ground	real	winter	normal	tree	162	100	163	442	689	432	-	451	680
ground	real	winter	normal	ground	189	59	189	490	578	490	-	502	717
ground	synthetic	adverse	rainy	sky	98	67	99	401	112	404	3866	390	634
ground	synthetic	adverse	snowy	sky	79	58	79	454	183	456	4686	458	658
ground	synthetic	adverse	foggy	sky	85	83	85	449	241	484	4454	335	642
aerial	synthetic	winter	normal	-	100	78	101	436	335	438	4287	373	601
aerial	synthetic	winter	normal	sky	101	62	103	407	166	417	-	444	500
aerial	synthetic	winter	normal	tree	104	74	104	499	463	499	-	415	659
aerial	synthetic	winter	normal	ground	98	90	99	426	427	426	-	341	628
aerial	synthetic	fall	normal	-	95	99	95	509	351	515	4680	424	650
aerial	synthetic	fall	normal	sky	92	105	92	523	312	529	-	435	653
aerial	synthetic	fall	normal	tree	100	59	101	492	386	495	-	391	658
aerial	synthetic	fall	normal	ground	102	107	102	456	453	456	-	427	611
aerial	synthetic	summer	normal	-	95	92	95	525	299	532	4677	424	657
aerial	synthetic	summer	normal	sky	94	92	94	538	269	543	-	434	658
aerial	synthetic	summer	normal	tree	108	104	112	478	222	505	-	364	642
aerial	synthetic	summer	normal	ground	93	116	91	465	464	465	-	410	663
aerial	synthetic	adverse	rainy	sky	100	092	101	440	238	446	4084	468	554
aerial	synthetic	adverse	snowy	sky	85	67	85	598	298	605	4835	445	690
aerial	synthetic	adverse	foggy	sky	105	116	103	468	390	477	4394	445	635

Table 1. Comparison of our uncertainty estimation metric, **MCDO-map**, with other methods for the Faster R-CNN trained on the source domain. Each row shows the validation domain. In each row, three different methods have evaluated the uncertainty level. MCDO-NMS reported separately for TP and FP detections. MCDO-map works better than other methods in terms of capturing DSs effectively. Loc.: localization, Cls.: Classification

Validation Domain					MCDO-NMS×10^{-3}						MCDO-Map	grad-loss×10^{-3}	
					Localization			Classification			×10^{-4}	Loc.	Cls.
View	Source	Season	Weather	BG	Total	TP	FP	Total	TP	FP			
ground	real	winter	normal	-	84	31	88	405	146	425	5355	299	621
ground	real	winter	normal	sky	173	56	180	495	298	508	-	434	692
ground	real	winter	normal	tree	86	34	86	464	489	464	-	358	675
ground	real	winter	normal	ground	71	79	71	377	388	377	-	454	564

Table 2. Comparison of our uncertainty estimation metric, **MCDO-map**, with other methods for the Faster R-CNN trained on ground-real-winter-normal-sky domain. Each row shows the validation domain. In each row, three different methods have evaluated the uncertainty level. MCDO-NMS reported separately for TP and FP detections. MCDO-map works better rather than other methods in terms of capturing DSs effectively. Loc.: localization, Cls.: Classification

a few false detections occur during the MCDO iterations, resulting in non-zero values in both maps around the intersection of the tree and ground. Moving to Fig. 7b, the drone with a tree background demonstrates higher uncer-



(a)

(b)

(c)

Figure 5. 2D representation of Faster R-CNN's, trained on the source domain, last layer feature maps distribution for different domains by utilizing t-SNE. a) For the target domains, synthetic data is changed to real data with sky, tree, and ground backgrounds. b) The view and season have been changed to aerial view, and fall and summer, respectively, as well as different backgrounds. c) All defined domains are taken into account without separating the different backgrounds.

tainty. The bounding box exhibits some red areas in the entropy map, accompanied by non-zero standard deviation values inside the bounding box. Once again, false detections contribute to non-zero values in the maps. Finally, in Fig. 7c, the drone with a ground background is detected with the highest level of uncertainty among these cases, corresponding to way too red color for the bounding box in the entropy map and nonzero values within the bounding box

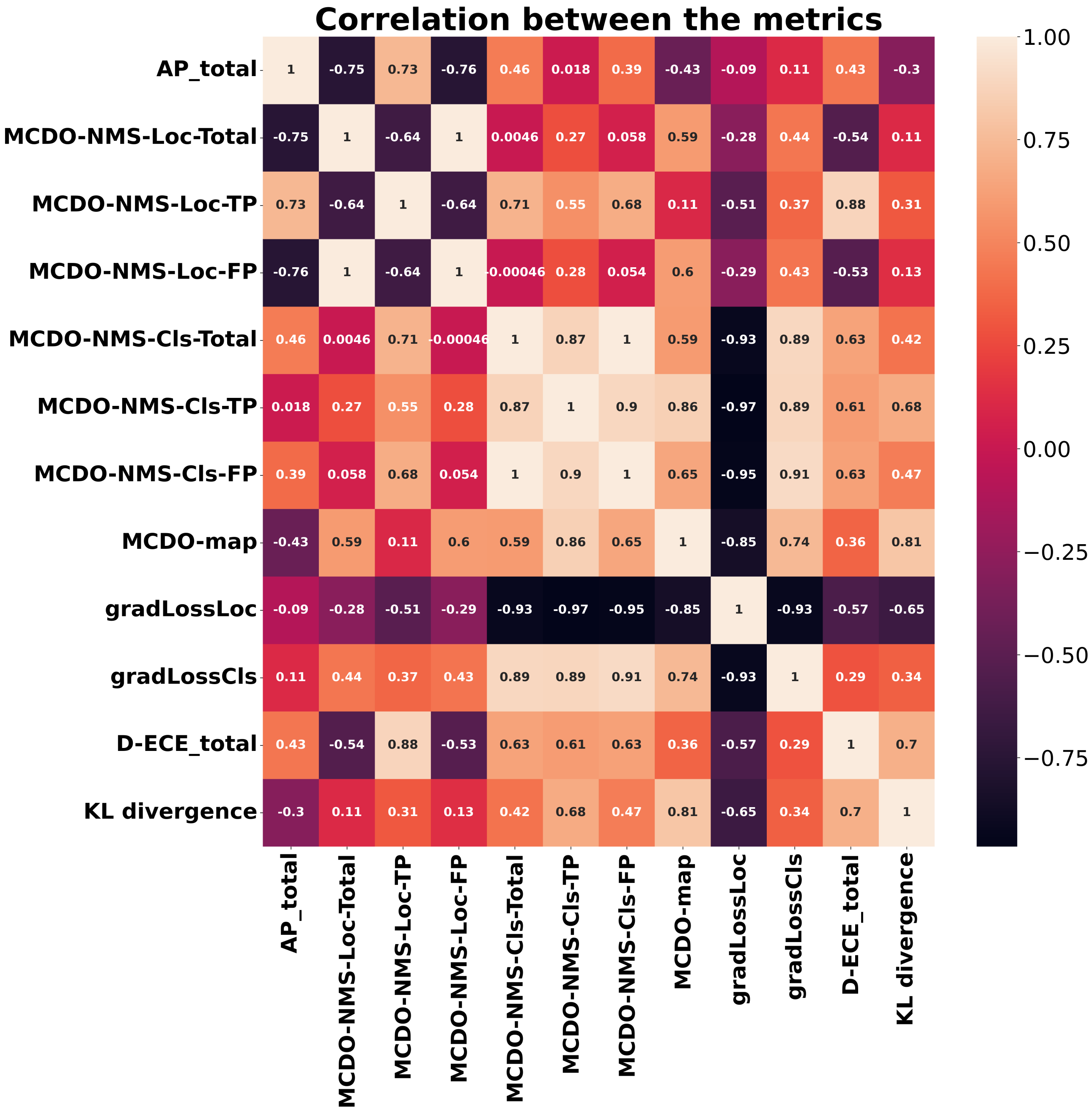
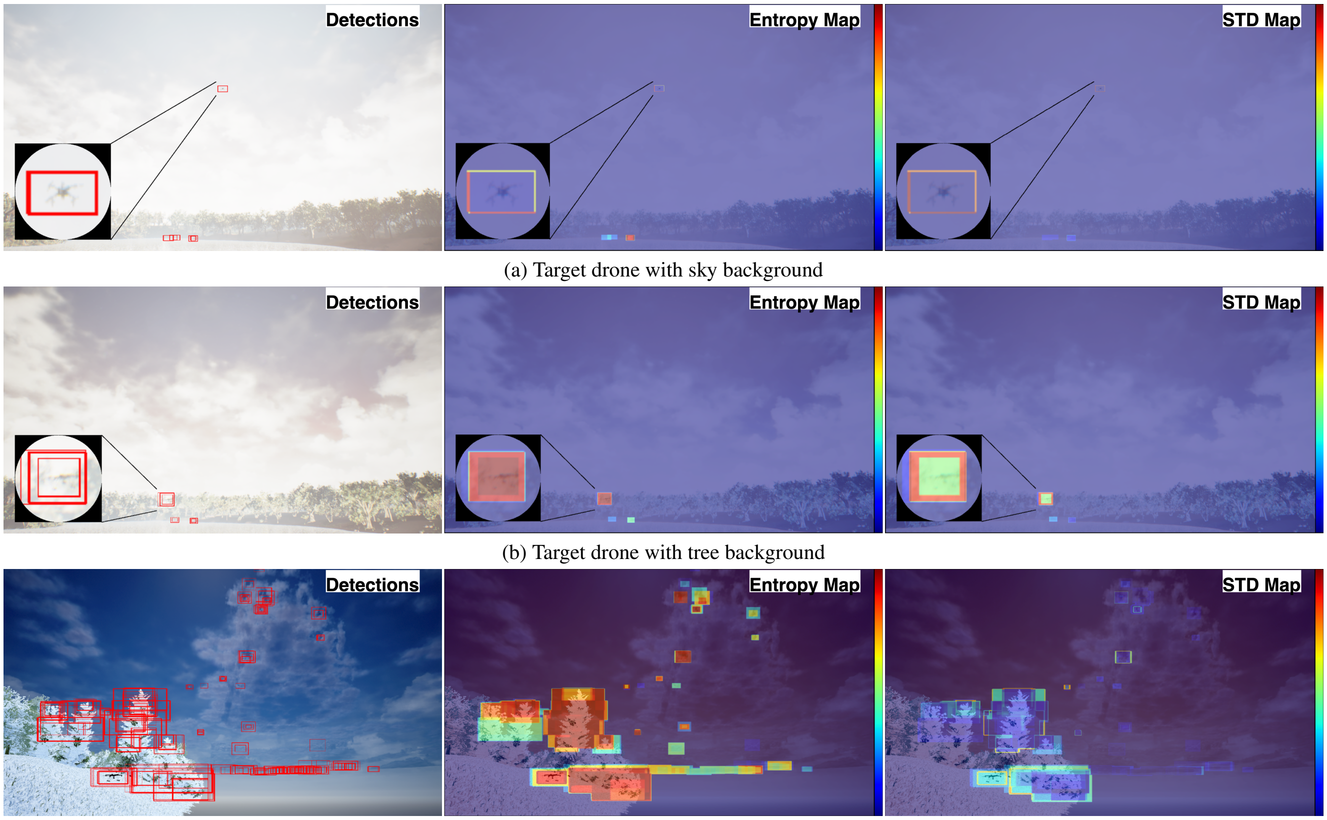


Figure 6. Correlation heatmap of metrics in DrIFT: MCDO-map exhibits the highest positive correlation (0.81) with KL divergence, indicating its superior capability to capture DSs. MCDO-NMS-Loc-Total, MCDO-NMS-Loc-FP, and MCDO-map yield the top three negative correlations with AP. MCDO-map emerges as the best metric in terms of capturing DSs effectively.

in the std map. However, a significant number of false detections around trees contribute to a considerable level of uncertainty in the maps, reflecting the low AP for trees and, consequently, higher uncertainty in this domain. Detailed AP and uncertainty values for trees are provided in Tab. 2 of the main paper.



(a) Target drone with sky background

(b) Target drone with tree background

(c) Target drone with ground background

Figure 7. Visual results of the trained faster R-CNN on the Ground-Synthetic-Winter-Normal-Sky domain: In (a), the drone with a sky background exhibits a low level of uncertainty, as indicated by the bounding box being entirely blue in the entropy map, with non-zero std values only at the edge. In (b), the drone with a tree background demonstrates a higher level of uncertainty, with red areas in the entropy map and non-zero standard deviation values within the bounding box. In (c), the drone with a ground background is detected with the highest level of uncertainty in this figure, with an intense red coloration within the box in the entropy map and high std.

References

[1] Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4, 2009. 4

[2] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[3] Jessica Hullman, Paul Resnick, and Eytan Adar. Hypothetical outcome plots outperform error bars and violin plots for inferences about reliability of variable ordering. *PloS one*, 10(11):e0142444, 2015. 4

[4] Solomon Kullback. Kullback-leibler divergence, 1951. 5

[5] MMDetection Contributors. OpenMMLab Detection Toolbox and Benchmark, Aug. 2018. 5

[6] Kemal Oksuz, Tom Joy, and Puneet K Dokania. Towards building self-aware object detectors via reliable uncertainty quantification and calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9263–9274, 2023. 3

[7] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5

[8] Tobias Riedlinger, Matthias Rottmann, Marius Schubert, and Hanno Gottschalk. Gradient-based quantification of epistemic uncertainty for deep object detectors. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3921–3931, 2023. 1

[9] Tobias Riedlinger, Marius Schubert, Karsten Kahl, and Matthias Rottmann. Uncertainty quantification for object detection: output-and gradient-based approaches. In *Deep Neural Networks and Data for Automated Driving: Robustness, Uncertainty Quantification, and Insights Towards Safety*, pages 251–275. Springer International Publishing Cham, 2022. 1

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[11] 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*, pages 21371–21382, 2022. 1