Bridge: Basis-Driven Causal Inference Marries VFMs for Domain Generalization

Supplementary Material

Contents

A. Overview

This supplemental material presents a detailed description of the Diverse Weather DroneVehicle dataset, the ***Bridge's*** specific structure of Stable Diffusion Backbone, and further experimental results in more detail.

B. Implementation details

Table 1 summarizes the training hyperparameters of our model across different datasets. All experiments are conducted on NVIDIA A40 GPUs. Moreover, since our framework does not involve fine-tuning VFMs, it can also be trained on consumer-grade GPUs such as the RTX 3090 or 4090, as long as the graphics memory is at least 20 GB.

For data augmentation, we follow the settings used in Boost [3] and GDD [4], incorporating both image-level augmentations (color and spatial transformations) and domain-level augmentations, including FDA [13], Histogram Matching, and Pixel Distribution Matching.

C. Overview of the Diverse Weather DroneVehicle Dataset

We further analyzed the distribution of manually annotated weather labels in the DroneVehicle dataset, as summarized in Table 2. Notably, images captured under foggy conditions exhibit a distinct bounding box distribution compared to other weather scenarios. Specifically, foggy scenes have fewer annotated objects and smaller average bounding box areas (mean ± std: 1,408 ± 1,646 pixels) relative to clear, dark, or extremely dark conditions. This variability introduces an additional challenge for domain-generalized object detection, requiring models to adapt not only to **weather-induced appearance variations** but also to inconsistent **object scales** across diverse conditions.

Figure 1 shows the average RGB histograms for the entire Diverse Weather DroneVehicle dataset under different weather conditions. These histograms illustrate the mean pixel intensity distributions in clear, dark, foggy, and extreme dark environments. As shown, the most challenging condition is the extreme dark scenario, where the average RGB pixel values cluster around $\mathbf{15}$, indicating severely limited illumination.

Figure 2 shows several sample images from DroneVehicle. We enhance the extreme dark images using the latest state-of-the-art low-light enhancement model, HVI-CIDNet [12], and denote the results as *enhanced extreme dark*. We can see that, even with such an advanced image restoration model, these images still contain substantial noise. We believe that establishing this DG benchmark can make a valuable contribution to the remote sensing community by drawing more attention to this highly challenging scenario in UAV imagery, and promoting the development of more robust and generalizable models for, e.g., traffic safety, disaster rescue, and precision agriculture, under challenging perception conditions.

D. Details of FACL

In this subsection, we provide more details about the comparison between the recent state-of-the-art front-door approach and our proposed model to block confounders in detection tasks.

GOAT [11] FACL *(CVPR'24)*, which leverages a multi-head attention mechanism to approximate the expectation terms in the front-door formulation. Specifically, the cross-attention operation is employed to estimate the expectations over the potential input features $\boldsymbol{x}'$ and the mediator features $\boldsymbol{m}$, respectively. The overall process can be formulated as follows:

$$\mathbb{E}_{x'}[\boldsymbol{x}'] \approx \sum_{x'} P(\boldsymbol{x}'|\boldsymbol{g}_1)\boldsymbol{x}' = \sum_i \frac{\exp(\boldsymbol{g}_1 \boldsymbol{x}_i'^T)}{\sum_j \exp(\boldsymbol{g}_1 \boldsymbol{x}_j'^T)} \boldsymbol{x}_i', \quad (1)$$

$$\mathbb{E}_{m|x}[\boldsymbol{m}] \approx \sum_{m} P(\boldsymbol{m}|\boldsymbol{g}_2)\boldsymbol{m} = \sum_i \frac{\exp(\boldsymbol{g}_2 \boldsymbol{m}_i^T)}{\sum_j \exp(\boldsymbol{g}_2 \boldsymbol{m}_j^T)} \boldsymbol{m}_i, \quad (2)$$

where $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ denote the query embeddings obtained from two learnable projection functions acting on the input features $\boldsymbol{x}$. Here, $\boldsymbol{x}'$ represents the potential input samples from the entire representation space, which are differ-

Table 1. Training hyperparameters for different datasets.

Setting	BDD100K	FoggyCityscape	DWD	R2A	Drone
Optimizer	SGD	SGD	SGD	SGD	SGD
Learning rate	1e-4	1e-4	1e-4	1e-4	1e-4
Weight decay	1e-5	1e-5	1e-5	1e-5	1e-5
Batch size	4	4	16	4	8
Warmup iter	500	500	500	500	500
Iter	20000	20000	20000	20000	12000
LR Scheduler	[16000, 19000]	[16000, 19000]	[18000]	[16000, 19000]	[10000]
Data augmentation: RandomResize, RandomCrop, RandomFlip, RandColor, RandomErasing, FDA [13]					

Table 2. Statistics of manually annotated weather labels in the DroneVehicle dataset. **BBoxes** denotes the total number of annotated objects, **Area** indicates the area of each bounding box in pixels (mean ± standard deviation), and the last five columns show the number of BBoxes per object category.

Weather	Img	BBoxes	Area (mean ± std)	Car	Truck	Freight Car	Bus	Van
Clear	8,881	132,942	2,830 ± 3,597	110,815	11,016	4,131	4,408	2,572
Dark	13,553	207,535	2,493 ± 2,504	183,059	6,190	5,061	6,905	6,320
Extreme Dark	4,965	85,006	3,079 ± 3,329	72,955	3,750	3,236	3,466	1,599
Foggy	1,040	27,087	1,408 ± 1,646	22,950	1,167	972	554	1,444

Table 3. DG Results (%) on five benchmark.

Method	BDD	Foggy	DWD	Drone	R2A
Baseline	57.8	57.7	48.6	47.1	72.7
FACL [11]	58.5	60.5	48.2	46.8	71.6
Ours	**58.9**	**61.6**	**50.8**	**48.4**	**73.3**

Table 4. DG Results (%) on five benchmarks using ResNet101 Backbone.

Method	BDD	Foggy	DWD	Drone	R2A
ResNet101 [5]	44.0	47.2	35.9	31.2	35.5
Ours	**45.3**	**49.5**	**37.0**	**32.1**	**36.8**

Table 5. Performance comparison with other methods under In-Domain and Cross-Domain settings.

Method	In-Domain	Cross-Domain	
	Cityscapes	BDD	Foggy
GDD [4]	59.8	50.1	46.6
Boost [3]	61.1	50.7	49.3
Ours	**61.8** (+0.7)	**53.6** (+2.9)	**53.1** (+3.8)

ent from the current inputs $\mathcal{X} = \boldsymbol{x}$, and are used to construct the cross-sampling expectation term in the front-door adjustment. The attention weights $P(\boldsymbol{x}'|\boldsymbol{g}_1)$ and $P(\boldsymbol{m}|\boldsymbol{g}_2)$ correspond to the normalized similarity scores between the query and key features, allowing the model to approximate the in-sampling and cross-sampling expectations through the attention mechanism.

However, directly sampling $\boldsymbol{x}'$ from the entire training set via K-means clustering is almost infeasible for dense prediction tasks. The clustering space is extremely high-dimensional, containing an enormous number of feature instances with continuous distributions and lacking semantic consistency across local regions. Therefore, we apply *dropout* to intervene on $\boldsymbol{x}$ and obtain $\boldsymbol{x}'$. The stochastic masking introduced by dropout acts as an *intervention* operator on the features. Table 3 reports the DG results on five benchmark datasets. While the dropout-based FACL [11] improves over the baseline, our method still consistently achieves higher performance across all datasets, particularly on dense prediction tasks such as DWD and Drone.

E. Additional Ablation Study

In addition to VFMs, we also evaluate our ***Bridge*** framework on a classic backbone, ResNet-101, to further verify its generalization ability and robustness. Table 4 reports the results on five DG benchmarks. As shown, ***Bridge*** consistently improves performance over the ResNet-101 across all datasets, with gains of +1.3% on BDD, +2.3% on Foggy, +1.1% on DWD, +0.9% on Drone, and +1.3% on R2A, demonstrating that our ***Bridge*** framework is effective not only with VFMs but also with classic backbones like ResNet-101 across diverse domain shifts.

F. Additional Quantitative Experiments

Table 5 shows the model's performance on in-domain testing using the Cityscapes benchmark. It can be observed that, although our method improves in-domain performance by only 0.7 mAP, it achieves significant gains of **2.9** and **3.8**

Table 6. Generalization Detection Results (%) on Cityscapes-Corruption Benchmark.

Methods	Noise			Blur				Weather			Digital					mPC ↑
	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Bright	Contrast	Elastic	JPEG	Pixel	
GDD *(Diff. Detector, SD)* [4] *(CVPR'25)*	20.3	23.2	17.2	26.8	21.7	23.7	3.4	16.6	24.2	32.5	34.4	30.6	33.7	29.1	24.4	24.1
Boost *(Diff. Detector, SD)* [3] *(ICCV'25)*	23.1	26.6	20.6	29.7	24.5	25.5	4.1	18.3	28.2	**37.2**	**39.2**	**35.5**	**37.4**	32.7	28.8	27.4
Ours (Diff. Detector, SD)	**27.3**	**29.7**	**26.4**	**30.3**	**29.4**	**27.8**	**11.5**	**23.6**	**27.3**	34.3	35.7	33.1	35.0	**33.7**	**32.2**	**29.2**

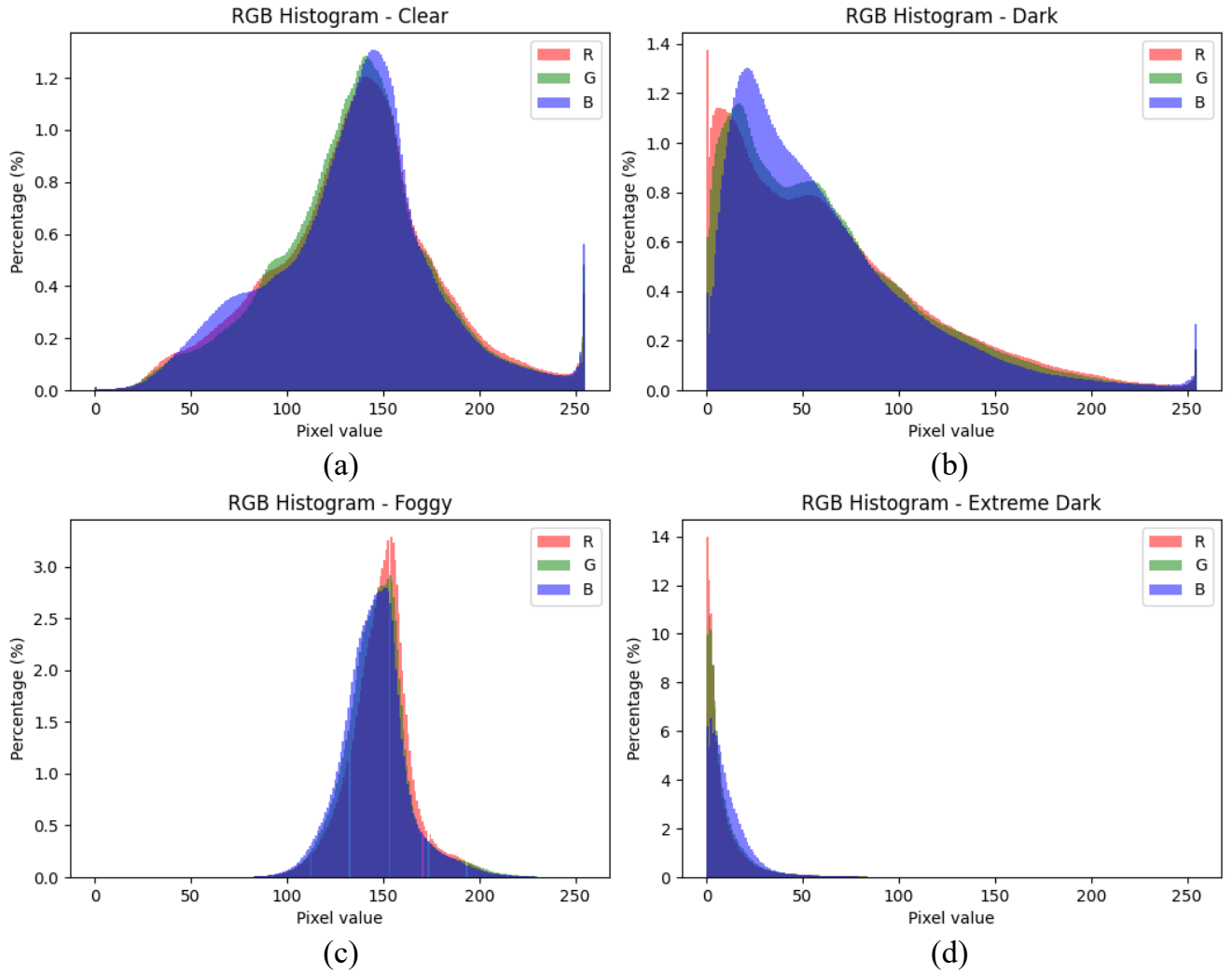

Figure 1. RGB histograms under varying conditions: (a) clear (balanced), (b) low-light (left-shifted peaks), (c) foggy (high-intensity concentration), and (d) extremely dark (sharp peak near zero).

mAP in **cross-domain tests**. This further demonstrates the strong generalization ability of the ***Bridge*** across different domains.

In addition to the five DG benchmarks evaluated in the manuscript, we further follow the OADG [7] setting, evaluating our method on Cityscapes-C [8] with 15 corruption types. Table 6 presents the detailed metrics under different corruptions, where our ***Bridge*** achieves state-of-the-art performance.

G. Generalization Beyond Object Detection

To further verify the generality of ***Bridge***, we extend our framework beyond object detection to two additional vision tasks: semantic segmentation and image classification.

Semantic Segmentation. Table 7 reports the cross-domain semantic segmentation results. Specifically, we train Mask2Former with a frozen DINOv3 backbone on Cityscapes and evaluate it on BDD100K and Mapillary [9].

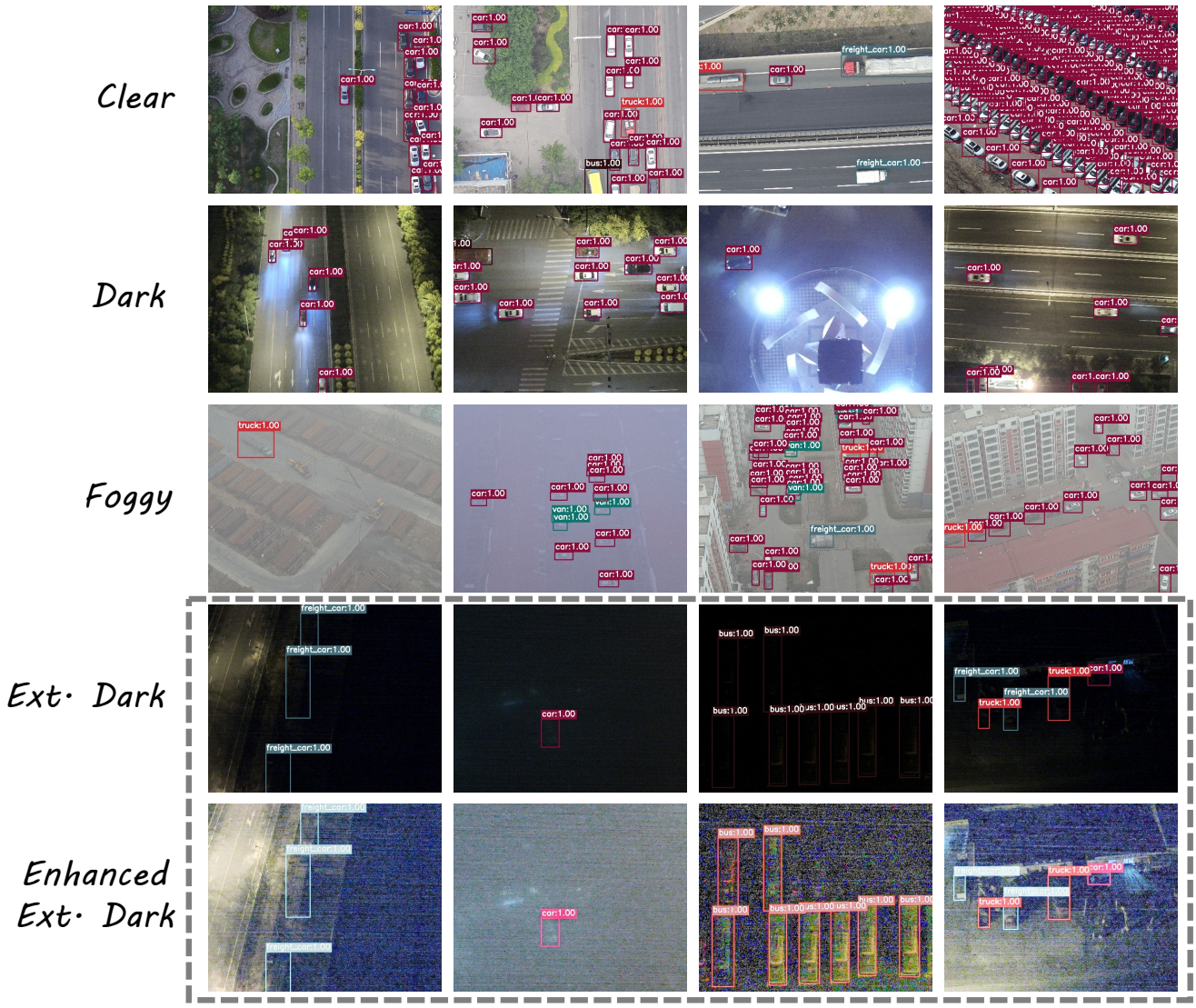

Figure 2. Examples of different weather conditions. The bottom row shows enhanced images generated using HVI-CIDNet [12], which improve visibility for illustration purposes only.

Bridge consistently outperforms the baseline on both target domains, improving the average performance from 69.85 to 70.78. These results indicate that our method generalizes well to dense prediction tasks beyond object detection.

Table 7. Cross-domain semantic segmentation results from Cityscapes to BDD100K and Mapillary.

Method	BDD100K	Mapillary	Avg.
Baseline	65.25	74.45	69.85
Ours	**65.77**	**75.78**	**70.78**

Image Classification. We further evaluate ***Bridge*** on the DomainBed [2] benchmark for domain-generalized image classification. Using frozen DINOv2 trained with ERM [10] as the baseline, we adapt our learnable basis projection strategy to the classification setting. On the challenging TerraIncognita [1] dataset, our method improves the performance from 56.8 ± 0.6 to $\mathbf{58.2 \pm 0.6}$ under the training-domain-validation model selection protocol, further demonstrating the task-level generalizability of ***Bridge***.

H. Instability of $\mathcal{P}(\mathcal{Z} \mid \mathcal{X})$

As mentioned in Section 3.1.2 of the manuscript, in this section, we explain why $\mathcal{P}(\mathcal{Z} \mid \mathcal{X})$ becomes unstable under limited sample settings.

Theorem 1 (Instability of $\mathcal{P}(\mathcal{Z} \mid \mathcal{X})$ under finite samples)**.** *Consider a fixed input $\mathcal{X} = x$, and let the confounder $\mathcal{Z}$*

take finitely many values $\{z_1, \ldots, z_K\}$. Assume that $\mathcal{X}$ and $\mathcal{Z}$ are independent, so that the true conditional probabilities are uniform:

$$p_k := \mathcal{P}(\mathcal{Z} = z_k \mid \mathcal{X} = x) = \frac{1}{K}, \quad k = 1, \ldots, K.$$

Assume we observe n independent samples $\{\mathcal{Z}_i\}_{i=1}^n$ drawn from $\mathcal{P}(\mathcal{Z} \mid \mathcal{X} = x)$. Define the empirical frequency

$$\hat{p}_k := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{\mathcal{Z}_i = z_k\}.$$

Then, for any $\delta > 0$,

$$\mathcal{P}(\hat{p}_k - p_k \geq \delta) \leq \exp\left(-2n\delta^2\right).$$

Proof. For the fixed k, define indicator random variables $X_i^{(k)} := \mathbf{1}\{\mathcal{Z}_i = z_k\}$. Each $X_i^{(k)}$ is independent and bounded in $[0, 1]$, with expectation $\mathbb{E}[X_i^{(k)}] = p_k = 1/K$. Hence,

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^{(k)}.$$

By Hoeffding's inequality [6] for bounded independent variables,

$$\mathcal{P}(\hat{p}_k - p_k \geq \delta) \leq \exp\left(-2n\delta^2\right),$$

which completes the proof.

Discussion. Although theoretically $\mathcal{X}$ and $\mathcal{Z}$ are independent and all p_k are equal to $1/K$, finite sample fluctuations can make the empirical $\hat{p}_k$ deviate significantly from p_k. In particular, small n allows some $\hat{p}_k$ to be temporarily overestimated, which can cause the mixture

$$\mathcal{P}(\mathcal{Y} \mid \mathcal{X}) = \sum_{\mathcal{Z}} \mathcal{P}(\mathcal{Y} \mid \mathcal{X}, \mathcal{Z})\, \mathcal{P}(\mathcal{Z} \mid \mathcal{X}), \tag{3}$$

to be dominated by one confounder component purely due to sampling noise, not any causal effect of $\mathcal{X}$ on $\mathcal{Z}$.

References

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision*, pages 456–473, 2018. 4

[2] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 4

[3] Boyong He, Yuxiang Ji, Zhuoyue Tan, and Liaoni Wu. Boosting domain generalized and adaptive detection with diffusion models: Fitness, generalization, and transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1912–1923, 2025. 1, 2, 3

[4] Boyong He, Yuxiang Ji, Qianwen Ye, Zhuoyue Tan, and Liaoni Wu. Generalized diffusion detector: Mining robust features from diffusion models for domain-generalized detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9921–9932, 2025. 1, 2, 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 2

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963. 5

[7] Wooju Lee, Dasol Hong, Hyungtae Lim, and Hyun Myung. Object-aware domain generalization for object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2947–2955, 2024. 3

[8] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 3

[9] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4990–4999, 2017. 3

[10] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 4

[11] Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. Vision-and-language navigation via causal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13139–13150, 2024. 1, 2

[12] Qingsen Yan, Yixu Feng, Cheng Zhang, Guansong Pang, Kangbiao Shi, Peng Wu, Wei Dong, Jinqiu Sun, and Yanning Zhang. Hvi: A new color space for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5678–5687, 2025. 1, 4

[13] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020. 1, 2