WiSE-OD: Benchmarking Robustness in Infrared Object Detection

Heitor R. Medeiros Atif Belal Masih Aminbeidokhti
Eric Granger Marco Pedersoli
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA)
International Laboratory on Learning Systems (ILLS)
Dept. of Systems Engineering, ETS Montreal, Canada

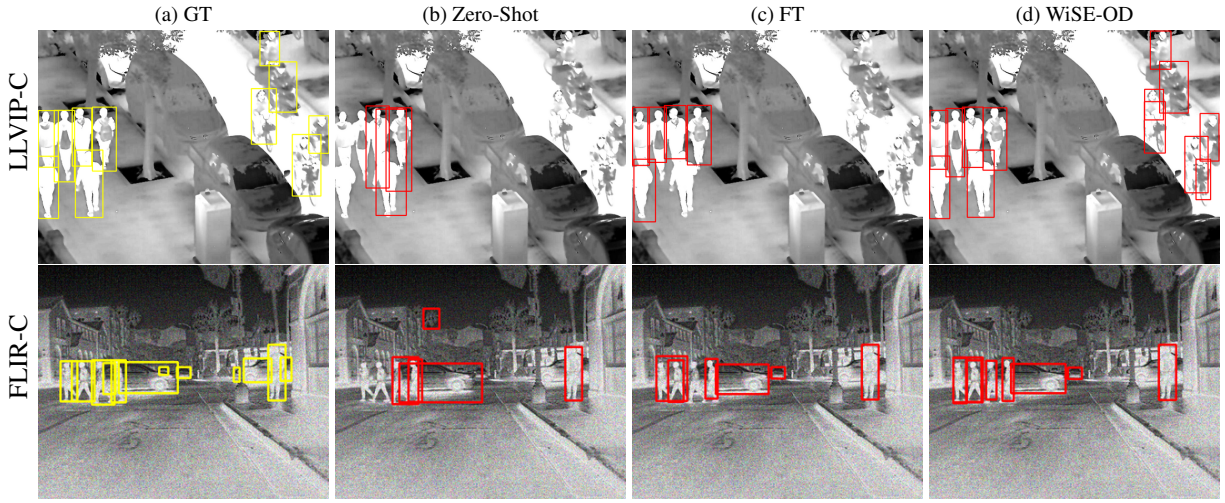


Figure 1. **Robustness of Infrared Object Detection on LLVIP-C and FLIR-C datasets.** In the first row, LLVIP-C has a brightness corruption severity level of 5; in the second row, FLIR-C shot noise corruption has a severity level of 2. In (a) ground-truth boxes (yellow); (b) zero-shot COCO; (c) fine-tuning (FT); (d) WiSE-OD with Faster R-CNN.

Abstract

Object detection (OD) in infrared (IR) imagery is critical for low-light and nighttime applications. However, the scarcity of large-scale IR datasets forces models to rely on weights pre-trained on RGB images. While fine-tuning on IR improves accuracy, it often compromises robustness under distribution shifts due to the inherent modality gap between RGB and IR. To address this, we introduce LLVIP-C and FLIR-C, two cross-modality out-of-distribution (OOD) benchmarks built by applying corruptions to standard IR datasets. Additionally, to fully leverage the complementary knowledge from RGB and infrared-trained models, we propose WiSE-OD, a weight-space ensembling method with two variants: WiSE-OD$_{ZS}$, which combines RGB zero-shot and IR fine-tuned weights, and WiSE-OD$_{LP}$, which blends zero-shot and linear probing. Evaluated using four RGB-pretrained detectors and two robust baselines on our benchmark and in the real-world out-of-distribution M3FD dataset, our WiSE-OD improves robustness across modalities and to corruption in synthetic and real-world distribution shifts without any additional training or inference costs. Our code is available at: https://github.com/heitorrapela/wiseod.

1. Introduction

In recent years, deep learning (DL) has achieved significant success across various computer vision tasks, including object detection (OD) [26] using thermal infrared (IR) imaging [14, 15]. Unlike visible spectrum imaging (RGB), which relies on reflected light, thermal IR imaging captures the heat emitted by objects, allowing it to function independently of lighting conditions. This makes IR-based OD highly effective in challenging environments with limited or absent visible light, such as night-time surveillance and autonomous driving cars [20]. Despite these advantages, IR

OD models must maintain consistent performance and reliable predictions under variations in input due to occlusions, viewpoint shifts, or image degradation. Ensuring such robustness is therefore essential for real-world applications in surveillance [2, 17], autonomous vehicles [16], and defense [11], where fluctuating environmental conditions can impact sensor inputs and compromise system reliability.

Without large-scale IR pre-training datasets, OD models for IR typically initialize from powerful models pre-trained on large-scale RGB datasets (e.g., COCO [7]), followed by fine-tuning on IR data. While this pipeline yields strong in-domain (ID) performance, where ID refers to test samples similar to the training data, it often compromises robustness against out-of-domain (OOD) samples [3]. OOD samples differ significantly from the training data, leading to performance degradation. This deterioration stems from the fact that the fine-tuning process tends to cause the model to prioritize task-specific information at the expense of broader knowledge acquired during pre-training. As a result, the model struggles to generalize to new or diverse scenarios [23]. This issue is further amplified by the already substantial modality shift between RGB and IR, making robust transfer learning even more difficult [13]. In classification, several techniques have been proposed to improve robustness under distribution shifts, including linear probing (LP), LP followed by fine-tuning (LP-FT) [6], and weight-space ensembling (WiSE-FT) [23]. While these methods effectively enhance robustness in classification, they either are not directly applicable to object detection or remain underexplored [23]. This gap arises from the complexity of the detection architectures and objectives, which require both precise localization and accurate classification. Moreover, cross-modality adaptation from RGB to infrared (IR) introduces additional challenges due to the modality shift and the scarcity of large-scale IR data [13–15].

To address these challenges, this paper introduces two key components: (i) a novel cross-modality RGB/IR corruption benchmark and (ii) two efficient approaches to improve average IR performance under OOD scenarios without additional training or inference cost. Our benchmark, LLVIP-C and FLIR-C, applies common corruption transforms to the original LLVIP and FLIR datasets to evaluate cross-modality OOD performance. Using this benchmark, we assess three families of IR detectors fine-tuned from RGB pre-trained models against standard robust fine-tuning baselines. Our analysis shows that traditional methods underperform under corruption. Therefore, we introduce WiSE-OD, a simple approach that preserves the original detection head to combine zero-shot and fine-tuned weights, yielding WiSE-OD$_{ZS}$ and its linear probing variation WiSE-OD$_{LP}$. Results show that these weight-space ensembling methods exhibit significant robustness. Additional analysis across various levels of corruption demonstrates that these methods improve average IR model performance by preserving ID accuracy from fine-tuning and OOD robustness from zero-shot weights, which explains their effectiveness. Our method also performs well under real-world shifts such as IR images captured at night, indoors, and under fog or rain, when the original models were trained solely on daytime IR images.

Our main contributions can be summarized as follows:

- A new benchmark, LLVIP-C and FLIR-C, is introduced to advance the evaluation of robust cross-modality OD between RGB and IR. This benchmark is essential for measuring detector performance across diverse, real-world conditions. Within this framework, we comprehensively evaluate three widely used object detection models: Faster R-CNN, FCOS, and RetinaNet, each initialized with COCO pre-trained weights.
- We propose WiSE-OD with two variants: WiSE-OD$_{ZS}$ and WiSE-OD$_{LP}$, an efficient OD technique that combines zero-shot and fine-tuned weights to enhance robustness under real-world and synthetic distribution shifts.
- Extensive experiments on the proposed benchmark and a real-world IR OOD dataset demonstrate significant gains for WiSE-OD over four OD frameworks (Faster R-CNN, FCOS, RetinaNet, and YOLOv8).

2. Related Works

Object detection. OD is one of the most challenging computer vision tasks [10], especially due to many different environmental conditions [16]. The objective of OD is to localize with a bounding box and provide labels for all objects in an image [24]. Commonly, detectors can be categorized into two groups: one-stage and two-stage detectors. The most famous two-stage detector is Faster R-CNN [18], which first generates regions of interest and then uses a second classifier to confirm object presence within those regions. In contrast, one-stage detectors eliminate the proposal generation stage, targeting real-time inference. Among one-stage detectors, RetinaNet [8] uses focal loss to address the class imbalance. Also, models such as FCOS [21] have emerged in this category, eliminating predefined anchor boxes to enhance inference efficiency. The YOLO family is a prominent line of one-stage detectors, with YOLOv8 adopting an anchor-free design with a decoupled head and offering a strong speed–accuracy trade-off. Our work focuses on these four detectors: Faster R-CNN, RetinaNet, FCOS, and YOLOv8.

Robustness in Object Detection. Robustness in OD refers to the model's ability to maintain performance despite variations in input conditions. Hendrycks & Dietterich [3] proposed diverse corruptions for classification datasets,

resulting in ImageNet-C and CIFAR10-C. Michaelis et al. [16] extended this to OD, proposing Pascal-C, COCO-C, and Cityscapes-C with a study on corruption severity and detector performance. Beghdadi et al. [1] introduced additional local transformations for RGB OD on COCO, and Mao et al. [12] proposed COCO-O with six types of natural distribution shifts. Despite growing efforts for RGB OD robustness, IR OD still lacks such benchmarks. In this direction, Josi et al. [5] applied classification corruptions to IR for person ReID. Given the widespread use of IR in surveillance and autonomous driving, a robustness benchmark for IR OD is essential.

Robust Fine-Tuning. The deep learning community has explored various fine-tuning (FT) strategies to improve robustness in classification tasks. A common approach is linear probing (LP), where the backbone is frozen and only the head is trained. Kumar et al. [6] extended this with LP-FT, which first trains a linear head before unfreezing the backbone for full fine-tuning. Wortsman et al. [23] proposed WiSE-FT, which ensembles the weights of a zero-shot pre-trained model and its fine-tuned counterpart in weight space, showing strong performance under distribution shifts on ImageNet.

In this work, we adapt these robustness techniques, which were originally developed for classification, to the more challenging cross-modality object detection setting, offering simple yet effective strategies to mitigate corruption effects.

3. Background

In this section, we introduce preliminary definitions that are necessary to understand this work, and subsequently, we define our proposed benchmark.

Object Detection. Consider a set of training samples $\mathcal{D} = \{(x_i, B_i)\}$, where $x_i \in \mathbb{R}^{W\times H\times C}$ are images with spatial resolution $W \times H$ and C channels, and $B_i = \{b_0, b_1, ..., b_N\}$ is a set of bounding boxes corresponding to the image x_i. Each bounding box can be represented as $b = (c_x, c_y, w, h, o)$ where c_x and c_y are the center coordinates of the bounding box with size $w \times h$ and o is the class label. During training we aim to learn a parameterized function $f_\theta : \mathbb{R}^{W\times H\times C} \rightarrow \mathcal{B}$, with $\mathcal{B}$ being the family of sets B_i and θ the model's parameter vector. The optimization of f_θ is guided by a combination of a regression $\mathcal{L}_r$ and classification $\mathcal{L}_c$ loss, i.e., l_2 loss and binary cross-entropy, respectively. The loss function for object detection can be represented as:

$$\mathcal{L}_d(\theta) = \frac{1}{|\mathcal{D}|} \sum_{(x,B)\in\mathcal{D}} \mathcal{L}_c(f_\theta(x), B) + \lambda\mathcal{L}_r(f_\theta(x), B). \quad (1)$$

Robustness to Corruption. Corruption robustness measures a classifier's average performance under classifier-agnostic input distortions [3]. Let $f : \mathcal{X} \rightarrow \mathcal{Y}$ be a classifier trained on samples drawn from a distribution $\mathcal{Q}$, and let $C = \{c : \mathcal{X} \rightarrow \mathcal{X}\}$ be a set of corruption functions (e.g., noise, blur, contrast) and $\mathcal{E} = \{e : \mathcal{X} \rightarrow \mathcal{X}\}$ a set of additional perturbation functions. The classifier's clean accuracy is $\mathbb{P}_{(x,y)\sim\mathcal{Q}}\big(f(x) = y\big)$. Its corruption robustness, i.e., its expected accuracy under all compositions of one corruption and one perturbation, can be represented as:

$$\mathbb{E}_{c\sim C}\,\mathbb{E}_{e\sim\mathcal{E}}\Big[\,\mathbb{P}_{(x,y)\sim\mathcal{Q}}\big(f\big(e(c(x))\big) = y\big)\Big].$$

Weight-space Ensembling. Given a mixing coefficient $\lambda \in [0, 1]$, weight-space ensembling can be defined as the following function:

$$f_{\text{wse}}(\theta_i, \theta_j; \lambda) \;=\; (1-\lambda)\,\theta_i \;+\; \lambda\,\theta_j, \qquad (2)$$

which computes the element-wise convex combination of two parameter vectors θ_i and θ_j. The resulting ensemble parameter $\theta_{\text{ens}} = f_{\text{wse}}(\theta_i, \theta_j; \lambda)$ is then used to initialize the model for prediction. A notable example of this technique is WiSE-FT [23]. Moreover, weight-space ensembling builds on principles of output-space ensemble averaging [4] and has demonstrated improved OOD robustness on classification benchmarks [22].

4. OD IR Robustness Benchmark

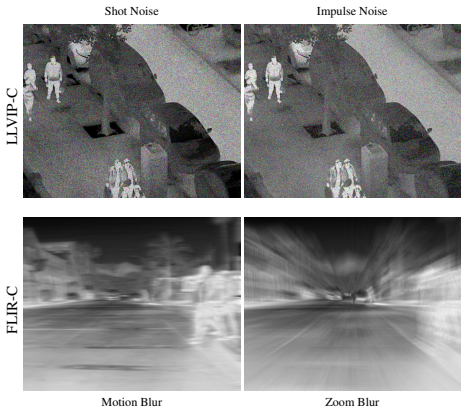


Figure 2. **LLVIP-C and FLIR-C examples.** First row, we have one example from the LLVIP-C test set with two different corruptions: Shot Noise, and Impulse Noise with a severity level of 5. In the second row, we have one example from the FLIR-C test set with Motion Blur and Zoom Blur with a severity level of 5.

4.1. Benchmark Datasets

For our proposed robust IR OD benchmark with corruptions, we explore two classical datasets containing paired RGB and infrared images: LLVIP and FLIR. Additionally, we explore the M3FD dataset, which has real-world shifts.

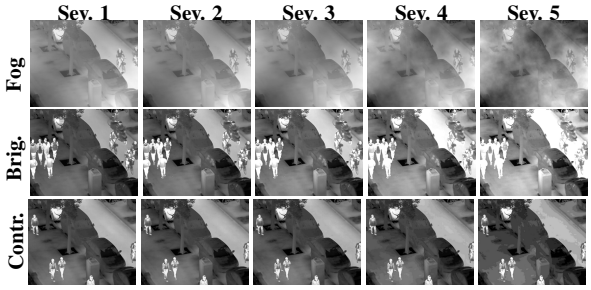


Figure 3. **Examples of fog perturbations at different severity levels for LLVIP.** Each column shows the effect of increasing corruption severity (1–5) on infrared images. Rows: top-fog, middle-brightness, and bottom-contrast. Higher severities introduce stronger degradations, simulating real-world challenging conditions.

LLVIP: LLVIP is a surveillance dataset composed of $12,025$ paired IR and RGB images for training and $3,463$ paired IR and RGB images for testing. The resolution of images is 1280×1024 pixels, and annotations consist of bounding boxes around pedestrians. **FLIR ALIGNED:** For the FLIR dataset, we used the sanitized and aligned paired sets provided [25], which contains $4,129$ paired IR and RGB images for training, and $1,013$ paired IR and RGB images for testing. The FLIR images are captured by a front-mounted car camera at a resolution of 640×512 pixels, and annotations contain bicycles, dogs, cars, and people. **M3FD (Real-World OOD):** M3FD [9] is a paired RGB-IR benchmark with well-aligned images captured by a calibrated dual-sensor rig. It contains 4,200 aligned pairs at 1024×768 resolution and six classes (person, car, bus, motorcycle, truck, lamp). We considered the day IR images as training data, and for testing, we considered fog, night, rain, and indoor images. The high resolution, alignment quality, and scenario diversity make M3FD a strong testbed for robustness in multi-modality detection and fusion.

LLVIP-C and FLIR-C: In this section, we present our two corrupted benchmarks: LLVIP-C and FLIR-C, derived from the LLVIP and FLIR datasets. In Figure 2, the first row shows an LLVIP-C test example corrupted with Shot Noise and Impulse Noise at severity level 5. The second row shows an FLIR-C test example corrupted with Motion Blur and Zoom Blur at severity level 5. As illustrated qualitatively, severity level 5 is too strong for the FLIR images, already compressed JPEGs, and both zero-shot and fine-tuned models perform worse on FLIR-C than on LLVIP-C at this level. Therefore, we recommend a maximum corruption severity of 2 for FLIR-C based on qualitative and quantitative results. For the following experiments, we use severity level 5 for LLVIP-C and severity level 2 for FLIR-C. In Figure 3, we show the impact of different severity levels (pre-defined levels) defined by [16] on the IR images going from severity 1 to severity 5.

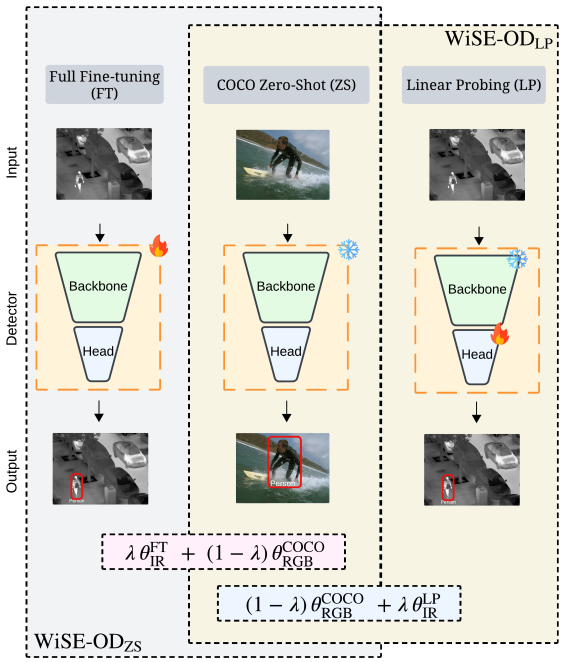


Figure 4. **Our proposed method: WiSE-OD and its variants.** In the large grey box, we have WiSE-OD$_{ZS}$ with the equation inside the pink square, and WiSE-OD$_{LP}$ in the yellow large box with the equation inside the blue square.

5. WiSE-OD

Our proposed method, WiSE-OD (f_{wod}) in Figure 4, extends the idea of WiSE-FT to object detection setting. Let $\theta_{\text{RGB}}^{\text{COCO}}$ and $\theta_{\text{IR}}^{\text{FT}}$ denote the parameters of the RGB pre-trained COCO detector and the fully fine-tuned IR detection models, respectively. WiSE-OD constructs a new detector by interpolating these parameter vectors in weight space:

$$f_{\text{wod}}(\theta_{\text{RGB}}^{\text{COCO}}, \theta_{\text{IR}}^{\text{FT}}; \lambda) \;=\; (1-\lambda)\,\theta_{\text{RGB}}^{\text{COCO}} \;+\; \lambda\,\theta_{\text{IR}}^{\text{FT}}. \qquad (3)$$

The resulting interpolated model inherits both the broad generalization of large-scale COCO pre-training and the modality-specific accuracy of IR fine-tuning, yet requires no extra modules or change to the inference pipeline, only a one-time weight merge. We evaluate two variants: WiSE-OD$_{ZS}$ uses $\theta_{\text{IR}}^{\text{FT}}$ (full fine-tuning), and WiSE-OD$_{LP}$ uses $\theta_{\text{IR}}^{\text{LP}}$ (linear probing on the detection head with a frozen backbone). Both variants consistently improve robustness under domain shift and common corruptions, while maintaining the same inference cost as a single detector. This weight-space ensembling is model-agnostic and can be extended to fuse multiple checkpoints or modalities by hierarchical interpolation.

Metrics: Following the methodology for benchmarking robustness in OD [16], we select AP$_{50}$ as our detection performance metric for both LLVIP-C and FLIR-C evaluations. We also report the dataset-specific performance (P), defined as AP$_{50}$ on the original target dataset (infrared), and

mean performance under corruption, mPC, defined as:

$$mPC = \frac{1}{N_c} \sum_{c=1}^{N_c} P_c, \tag{4}$$

where P_c is the AP_{50} under corruption, and mPC is the average over all the corruptions. In our case, $N_c = 14$ since we decided to remove the glass blur corruption because it lacks a fast implementation for our benchmark.

Baseline models: In our study, we utilize four OD architectures: Faster R-CNN, FCOS, RetinaNet, and YOLOv8, all initialized with COCO pre-trained weights. These models are trained on the COCO dataset, which contains 80 object categories, providing strong initial performance for most detection tasks and facilitating subsequent fine-tuning. We evaluate the following robust fine-tuning methods using our proposed benchmark:

1. **Zero-Shot (ZS) –** Unmodified detectors used directly for deployment without any fine-tuning.
2. **Linear probing (LP) –** Train the classification and regression heads on top of a frozen backbone by minimizing the detection loss.
3. **Full fine-tuning (FT) –** Update both the detection heads and the backbone parameters by minimizing the detection loss.
4. **LP-FT –** A two-stage process in which we first apply linear probing and then perform full fine-tuning initialized from the LP stage.
5. **Weight-ensembling –** Two variants, WiSE-OD_{ZS} and WiSE-OD_{LP}, which interpolate parameters between the zero-shot or linear-probing models and the fully fine-tuned models, respectively.

6. Experiments and Results

6.1. Training protocol

For this work, we split each training dataset into 80% for training and 20% for validation, reserving the original test set for final evaluation. All models were implemented in PyTorch, optimized with the Adam optimizer, and trained on an NVIDIA A100 GPU. We set a maximum training budget of 200 epochs for all detectors; in practice, fine-tuning typically converges within 10–20 epochs, depending on the model and dataset. We used a cosine annealing scheduler on the training loss and applied early stopping based on validation AP_{50}. Our 14 corruption types and severities follow the ImageNet-C [3]/COCO-C [16] protocol (see supplementary material for details). We provide code for reproducibility.

6.2. Benchmark Quantitative Results

In this section, we measured the mPC performance of all the proposed baselines for our benchmark on LLVIP-C with a severity level of 5 and the FLIR-C dataset with a severity level of 2. Results are shown in Table 2 for Faster R-CNN, FCOS, and RetinaNet under zero-shot, FT, LP, LP-FT, WiSE-OD_{ZS}, and WiSE-OD_{LP}. We see from Table 2 that, on average, WiSE-OD_{ZS} with λ fixed at 0.5, i.e., equal weighting of Zero-Shot and FT, outperforms all other baselines without the need to tune any hyperparameters. For instance, on LLVIP-C, WiSE-OD_{ZS} improved mPC by 18.68 over FT and by 4.12 over LP for Faster R-CNN. In most cases, our proposed variant WiSE-OD_{LP} outperformed WiSE-OD_{ZS} for Faster R-CNN and RetinaNet.

6.3. Detection performance per corruption

In this section, we evaluated the benchmark per corruption. In Table 1, we show the in-domain performance (evaluation on infrared of the LLVIP dataset), we name "Original", which is the original LLVIP infrared test set. Then, we have the corruptions for the LLVIP-C and the mPC metric for the Faster R-CNN detector; the same methodology was used for FLIR and FLIR-C. The original performance is measured in terms of AP_{50} for Faster R-CNN; additional results for FCOS, RetinaNet, and all the detectors are provided in the supplementary material. As described in Table 1, the in-domain performance for LP (91.82) and LP-FT (92.18) is lower than the FT (93.63), but the mPC is much higher than the zero-shot and FT. The WiSE-OD_{ZS} and WiSE-OD_{LP} were able to outperform the others with in-domain of 96.06 and 96.24, respectively, and for the mPC, WiSE-OD_{LP} achieves 75.83 and WiSE-OD_{ZS} 75.08. For WiSE-OD_{ZS} this corresponds to an increase of 18.68 over FT and 4.12 mPC over LP. For FLIR-C, we also have good improvements compared to the others. It is important to mention that the WiSE-OD_{ZS} is a training-free technique, and for this table, λ is fixed at 0.5, same for WiSE-OD_{LP}, but this variation needs the LP model instead of the zero-shot. In contrast, IR-adapted detectors already degrade substantially, which limits ensemble gains. For example, on LLVIP-C FT collapses ($AP_{50} = 0.00$), while WiSE-OD_{LP} recovers to 14.3 (+14.3). On FLIR-C, FT remains strong (75.5) and WiSE-OD_{LP} improves to 79.7 (+4.2). Accordingly, we keep a fixed λ to avoid target-data tuning.

6.4. Performance over different corruption levels

In this section, we measured the per AP_{50} performance for Faster R-CNN, FCOS, and RetinaNet over different corruption severity levels for the benchmark. Here, in Figure 5, we provided (a) Frost for LLVIP-C, (b) Fog for FLIR-C for Faster R-CNN, and (c) different APs for per-class analysis (person, car, truck) in FLIR-C. When the corruption severity level increases, e.g., from 1 to 5 in LLVIP-C, we can see a large drop in the zero-shot and FT, while the WiSE-OD_{ZS} is more stable and can bring more robustness to the final model. Some corruptions have more impact than oth-

Table 1. **AP_{50} performance over the perturbations on different datasets.** For LLVIP-C with severity level 5, and FLIR-C with severity level 2 for Faster R-CNN.

	LLVIP-C					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Original	71.21 ± 0.02	93.68 ± 0.86	91.82 ± 0.15	92.18 ± 0.03	96.06 ± 0.22	96.24 ± 0.03
Gaussian Noise	59.24 ± 0.07	67.46 ± 7.45	75.12 ± 0.12	72.51 ± 0.28	86.68 ± 0.44	85.45 ± 0.76
Shot Noise	51.48 ± 0.14	64.83 ± 7.79	70.82 ± 0.27	69.89 ± 0.26	85.26 ± 0.50	85.25 ± 0.12
Impulse Noise	56.62 ± 0.07	71.32 ± 6.33	78.31 ± 1.13	75.20 ± 0.86	88.54 ± 0.33	88.40 ± 0.15
Defocus Blur	47.90 ± 0.08	80.48 ± 3.60	84.31 ± 0.24	83.12 ± 0.05	89.74 ± 0.98	90.40 ± 0.03
Motion Blur	26.39 ± 0.23	78.32 ± 3.18	77.15 ± 0.33	75.13 ± 0.05	86.81 ± 0.71	87.02 ± 0.32
Zoom Blur	02.47 ± 0.02	11.18 ± 1.56	24.65 ± 0.39	17.46 ± 0.02	22.83 ± 2.44	27.08 ± 0.01
Snow	33.65 ± 0.01	13.46 ± 4.45	69.92 ± 0.14	69.34 ± 0.13	65.97 ± 1.90	65.28 ± 2.70
Frost	33.25 ± 0.38	47.32 ± 3.45	68.00 ± 0.27	66.93 ± 0.42	75.87 ± 0.39	74.85 ± 0.29
Fog	59.60 ± 0.10	50.90 ± 10.07	87.05 ± 0.07	87.33 ± 0.39	84.51 ± 3.80	88.17 ± 0.06
Brightness	41.77 ± 0.03	35.36 ± 6.97	71.47 ± 0.34	76.45 ± 0.14	82.10 ± 1.20	82.61 ± 0.92
Contrast	47.48 ± 0.04	00.00 ± 0.00	51.53 ± 0.03	48.93 ± 0.02	10.57 ± 3.82	14.30 ± 1.82
Elastic transform	52.42 ± 0.18	92.41 ± 0.93	86.30 ± 0.25	88.98 ± 0.14	94.72 ± 0.07	94.85 ± 0.14
Pixelate	03.95 ± 0.01	87.69 ± 2.67	65.35 ± 0.09	65.71 ± 0.04	85.06 ± 3.32	84.33 ± 0.05
JPEG compression	57.22 ± 0.02	88.93 ± 1.69	83.58 ± 0.07	83.32 ± 0.24	92.59 ± 1.22	93.73 ± 0.03
mPC	40.96	56.40	70.96	70.02	75.08	**75.83**

	FLIR-C					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Gaussian Noise	31.21 ± 0.29	28.07 ± 2.91	41.99 ± 0.39	39.83 ± 0.44	42.49 ± 4.48	39.33 ± 0.36
Shot Noise	25.26 ± 0.12	15.73 ± 2.05	33.24 ± 0.23	33.15 ± 0.43	30.45 ± 3.96	35.84 ± 0.46
Impulse Noise	17.69 ± 0.03	13.22 ± 2.27	26.15 ± 0.46	25.58 ± 0.46	22.51 ± 2.78	26.72 ± 0.18
Defocus Blur	25.32 ± 0.22	52.47 ± 0.99	44.57 ± 0.12	45.29 ± 0.20	54.08 ± 1.74	56.50 ± 1.22
Motion Blur	25.01 ± 0.25	51.71 ± 2.12	43.01 ± 0.41	42.79 ± 0.48	51.03 ± 2.16	57.24 ± 0.28
Zoom Blur	08.98 ± 0.05	17.97 ± 0.90	15.17 ± 0.02	14.17 ± 0.07	16.93 ± 1.00	19.34 ± 0.17
Snow	09.84 ± 0.14	07.86 ± 2.01	16.94 ± 0.31	19.57 ± 0.58	13.94 ± 2.66	16.31 ± 0.23
Frost	21.96 ± 0.50	33.87 ± 4.69	36.15 ± 0.29	37.82 ± 0.43	37.97 ± 3.63	38.80 ± 2.87
Fog	56.36 ± 0.28	73.61 ± 0.06	71.90 ± 0.37	71.26 ± 0.15	78.68 ± 1.24	78.10 ± 1.09
Brightness	64.41 ± 0.26	75.18 ± 0.99	74.92 ± 0.20	74.24 ± 0.10	79.72 ± 0.35	78.42 ± 0.17
Contrast	54.59 ± 0.04	75.47 ± 1.29	71.11 ± 0.13	70.60 ± 0.30	78.36 ± 1.06	79.72 ± 0.15
Elastic transform	41.88 ± 0.24	69.68 ± 1.15	64.49 ± 0.18	64.29 ± 0.06	73.39 ± 0.40	73.83 ± 0.54
Pixelate	38.67 ± 0.11	54.91 ± 6.21	55.61 ± 0.09	55.53 ± 0.13	61.12 ± 3.13	56.54 ± 0.01
JPEG compression	50.24 ± 0.14	57.55 ± 3.04	64.36 ± 0.30	62.82 ± 0.17	66.65 ± 0.89	65.70 ± 0.27
mPC	33.67	44.80	47.11	46.92	50.52	**51.59**

ers for each dataset; for instance, in FLIR-C, the noise corruption affected the performance more due to the original low-quality images. In the IR modality, the contrast corruption affects the detection performance more because IR images already have low contrast when compared to natural RGB images. A similar trend of stability of WiSE-OD$_{ZS}$ for other detectors and corruptions over zero-shot and FT is shown in the supp. materials.

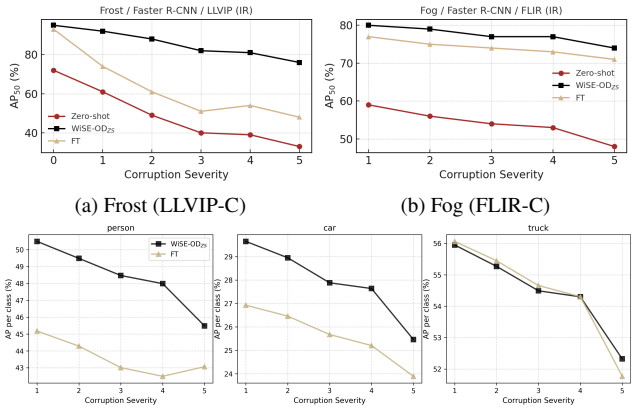
6.5. Activation map analysis

In this section, we qualitatively analyze activation maps of Faster R-CNN under corruptions on LLVIP-C for the zero-shot model, WiSE-OD$_{ZS}$, and FT. As shown in Figure 6, Grad-CAM [19] highlights impulse noise (top) and zoom blur (bottom), with ground-truth boxes in red (additional examples are in the supplementary). While FT and zero-shot often fail to detect the person under heavy corruptions, WiSE-OD$_{ZS}$ activates more strongly on object regions, in-

Table 2. **Detection performance for the OD IR Robustness Benchmark.** mPC metric for LLVIP-C with severity 5 and FLIR-C with severity level 2.

Detector	LLVIP-C					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Faster R-CNN	40.96	56.40	70.96	70.02	75.08	**75.83**
FCOS	36.11	61.17	63.91	60.26	**76.50**	75.95
RetinaNet	37.50	61.13	60.37	61.11	73.69	**74.39**

Detector	FLIR-C					
	Zero-Shot	FT	LP	LP-FT	WiSE-OD$_{ZS}$	WiSE-OD$_{LP}$
Faster R-CNN	33.67	44.80	47.11	46.92	50.52	**51.59**
FCOS	28.85	41.07	38.92	38.14	**47.13**	46.76
RetinaNet	28.27	42.71	36.71	36.45	45.35	**47.53**



(a) Frost (LLVIP-C) (b) Fog (FLIR-C)

(c) AP per class (FLIR-C)

Figure 5. **AP_{50} versus corruption severity.** (a) Frost on LLVIP-C (IR) and (b) Fog on FLIR-C (IR). Curves compare Faster R-CNN in Zero-shot, WiSE-OD_{ZS}, and FT settings; y-axis shows AP_{50} (%). Severity increases left-to-right (0–5 for LLVIP-C, 1–5 for FLIR-C). (c) FLIR-C per-class AP_{50} under fog for *person*, *car*, and *truck*. WiSE-OD_{ZS} maintains a higher level of performance across severities.

dicating greater robustness. Although performance varies across corruption types, tuning λ can balance zero-shot and FT contributions for specific real-world needs. Overall, WiSE-OD_{ZS} preserves more complete object regions across LLVIP-C and FLIR-C, especially under fog and low contrast, showing that ensembling retains complementary cues and stabilizes predictions.

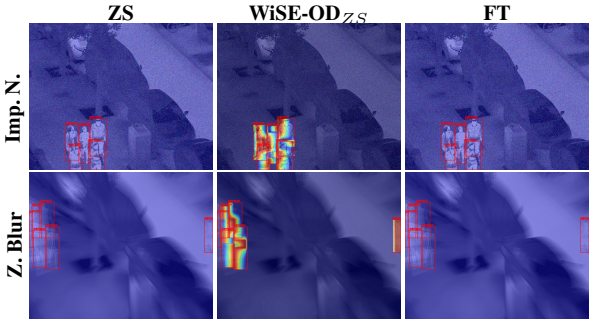


Figure 6. **Activation map analysis on LLVIP-C.** Rows: impulse noise (top) and zoom blur (bottom; severity 5). Columns: Zero-shot (ZS), WiSE-OD_{ZS}, and FT. Ground-truth boxes in red.

6.6. Real-world OOD Scenario

In this section, we validate robustness on the M3FD dataset [9], a multi-scenario, multi-modality benchmark designed for fusing infrared and visible modalities in object detection. M3FD contains diverse real-world IR scenes across *fog, rain, indoor, day, and night conditions*, providing challenging evaluation scenarios for robustness studies. Models are trained on day images and evaluated on the other splits, simulating a realistic OOD deployment. As shown in Tab. 4, WiSE-OD consistently improves over FT and ZS: **+10.8** mAP (raining), **+7.2** (fog), **+7.1** (night), and **+0.2** (indoor), for an average gain of **+6.3**. These gains are achieved without retraining or access to target data, highlighting WiSE-OD's practicality. Importantly, the consistent improvements on M3FD demonstrate that our approach is not limited to synthetic benchmarks (LLVIP-C, FLIR-C), but generalizes to real-world distribution shifts.

Table 3. **Ablation of λ over LLVIP-C and FLIR-C dataset for Faster R-CNN.** Where $\lambda = 0.0$ represents the zero-shot model, $\lambda = 0.5$ represents default WiSE-OD_{ZS} and $\lambda = 1.0$ represents the fine-tuning model. For LLVIP-C, the severity level is 5.

LLVIP-C

	$\theta(\lambda=0.0)$	$\theta(\lambda=0.2)$	$\theta(\lambda=0.5)$	$\theta(\lambda=0.8)$	$\theta(\lambda=1.0)$
Original	71.21 ± 0.02	93.88 ± 0.28	96.06 ± 0.22	95.41 ± 0.60	93.68 ± 0.86
Gaussian Noise	59.24 ± 0.07	86.52 ± 0.40	86.68 ± 0.44	78.47 ± 3.43	67.46 ± 7.45
Shot Noise	51.48 ± 0.14	83.86 ± 0.70	85.26 ± 0.50	76.42 ± 3.74	64.83 ± 7.79
Impulse Noise	56.62 ± 0.07	86.93 ± 0.65	88.54 ± 0.33	80.94 ± 2.70	71.32 ± 6.33
Defocus Blur	47.90 ± 0.08	88.41 ± 0.31	89.74 ± 0.98	85.85 ± 2.55	80.48 ± 3.60
Motion Blur	26.39 ± 0.23	81.10 ± 0.44	86.81 ± 0.71	83.24 ± 1.70	78.32 ± 3.18
Zoom Blur	02.47 ± 0.02	27.97 ± 0.62	22.83 ± 2.44	14.46 ± 1.82	11.18 ± 1.56
Snow	33.65 ± 0.01	67.67 ± 0.77	65.97 ± 1.90	38.50 ± 2.61	13.46 ± 4.45
Frost	33.25 ± 0.38	72.31 ± 0.23	75.87 ± 0.39	65.10 ± 0.57	47.32 ± 3.45
Fog	59.60 ± 0.10	89.79 ± 0.43	84.51 ± 3.80	64.47 ± 9.62	50.90 ± 10.0
Brightness	41.77 ± 0.03	82.38 ± 0.17	82.10 ± 1.20	62.81 ± 3.16	35.36 ± 6.97
Contrast	47.48 ± 0.04	50.59 ± 4.48	10.57 ± 3.82	00.77 ± 0.31	00.00 ± 0.00
Elastic transform	52.42 ± 0.18	89.92 ± 0.51	94.72 ± 0.07	94.32 ± 0.34	92.41 ± 0.93
Pixelate	03.95 ± 0.01	66.27 ± 3.14	85.06 ± 3.32	88.97 ± 2.35	87.69 ± 2.67
JPEG compression	57.22 ± 0.02	87.87 ± 0.89	92.59 ± 1.22	91.86 ± 1.38	88.93 ± 1.69
mPC	40.96	75.82	75.08	66.15	56.40

FLIR-C

	$\theta(\lambda=0.0)$	$\theta(\lambda=0.2)$	$\theta(\lambda=0.5)$	$\theta(\lambda=0.8)$	$\theta(\lambda=1.0)$
Original	65.52 ± 0.07	77.49 ± 0.10	82.20 ± 0.07	80.18 ± 0.11	77.57 ± 0.24
Gaussian Noise	31.21 ± 0.29	42.62 ± 2.92	42.49 ± 4.48	34.85 ± 3.85	28.07 ± 2.91
Shot Noise	25.26 ± 0.12	33.91 ± 2.98	30.45 ± 3.96	21.88 ± 2.93	15.73 ± 2.05
Impulse Noise	17.69 ± 0.03	24.85 ± 2.26	22.51 ± 2.78	16.96 ± 2.17	13.22 ± 2.27
Defocus Blur	25.32 ± 0.22	44.57 ± 2.40	54.08 ± 1.74	55.00 ± 0.98	52.47 ± 0.99
Motion Blur	25.01 ± 0.25	40.63 ± 2.17	51.03 ± 2.16	53.85 ± 2.19	51.71 ± 2.12
Zoom Blur	08.98 ± 0.05	13.72 ± 0.97	16.93 ± 1.00	18.32 ± 1.09	17.97 ± 0.90
Snow	09.84 ± 0.14	14.55 ± 2.07	13.94 ± 2.66	10.36 ± 2.48	07.86 ± 2.01
Frost	21.96 ± 0.50	33.37 ± 2.39	37.97 ± 3.63	36.47 ± 4.00	33.87 ± 4.69
Fog	56.36 ± 0.28	72.17 ± 0.86	78.68 ± 1.24	78.11 ± 1.49	73.61 ± 0.06
Brightness	64.41 ± 0.26	75.68 ± 0.19	79.72 ± 0.35	77.79 ± 1.24	75.18 ± 0.99
Contrast	54.59 ± 0.04	71.38 ± 0.95	78.36 ± 1.06	78.02 ± 1.09	75.47 ± 1.29
Elastic transform	41.88 ± 0.24	63.89 ± 0.37	73.39 ± 0.40	72.51 ± 0.58	69.68 ± 1.15
Pixelate	38.67 ± 0.11	55.23 ± 2.37	61.12 ± 3.13	58.87 ± 4.58	54.91 ± 6.21
JPEG compression	50.24 ± 0.14	63.21 ± 0.56	66.65 ± 0.89	63.27 ± 2.36	57.55 ± 3.04
mPC	33.67	46.41	50.52	48.30	44.80

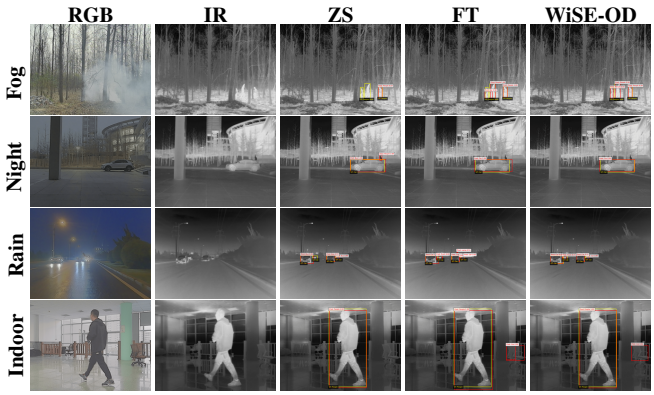


Figure 7. **Qualitative comparison on M3FD under adverse conditions.** Rows: fog, night, rain, indoor. Columns: RGB, IR, Zero-Shot (ZS), Fine-Tuning (FT), and WiSE-OD_{ZS}.

Table 4. **Detection performance of four detectors across M3FD splits.** Models are trained on IR day images; WiSE-OD is applied with λ=0.5. Results are reported for Rain, Night, Fog, Indoor, and average across conditions.

Method	Rain			Night			Fog			Indoor			Avg.		
	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP	AP_{50}	AP_{75}	mAP
Faster R-CNN															
Zero-shot	13.1	6.1	6.7	14.3	8.3	8.2	87.4	81.7	68.4	5.8	5.7	4.4	30.2	25.4	22.0
Fine-tuning	28.0	11.8	13.8	**47.8**	**30.9**	**29.4**	93.0	77.5	66.2	**59.8**	32.8	33.2	57.2	38.3	35.7
WiSE-OD	**29.4**	**13.2**	**14.9**	43.3	30.3	27.9	**97.2**	**84.5**	**74.2**	59.2	**36.2**	**33.9**	**57.3**	**41.1**	**37.7**
FCOS															
Zero-shot	10.6	5.5	5.6	12.3	7.2	7.1	83.4	74.8	64.0	4.3	4.0	3.3	27.7	22.9	20.0
Fine-tuning	21.7	9.5	10.7	**42.3**	**28.3**	**27.9**	93.5	78.7	65.9	**61.6**	32.1	31.7	**54.8**	37.2	34.1
WiSE-OD	**25.5**	**11.7**	**13.0**	34.7	22.0	21.6	**96.0**	**86.8**	**73.0**	56.4	**35.2**	**33.7**	53.2	**38.9**	**35.3**
RetinaNet															
Zero-shot	10.6	4.9	5.4	12.9	7.7	7.4	91.4	84.6	69.6	9.0	6.9	6.3	31.0	26.0	22.2
Fine-tuning	**29.3**	11.5	13.5	**45.0**	**28.8**	**27.3**	96.5	82.8	67.7	55.3	**34.4**	**32.9**	**56.5**	39.4	35.3
WiSE-OD	27.2	**12.7**	**13.7**	40.1	27.1	25.8	**97.6**	**89.9**	**73.6**	**56.7**	33.6	32.8	55.4	**40.8**	**36.5**
YOLOv8n															
Zero-shot	28.4	22.4	20.5	37.3	34.6	31.8	88.9	84.9	73.8	22.1	22.1	19.3	44.2	41.0	36.4
Fine-tuning	29.2	23.1	21.7	49.1	37.7	36.6	92.3	83.6	70.3	**55.1**	**43.4**	**38.5**	56.4	46.9	41.8
WiSE-OD	**42.2**	**35.5**	**32.5**	**55.5**	**46.1**	**43.7**	**95.3**	**90.9**	**77.5**	48.8	43.6	38.7	**60.4**	**54.0**	**48.1**

6.7. WiSE-OD$_{ZS}$: Ablation study on λ

In this section, we extensively conducted studies about the λ value to combine the zero-shot RGB COCO pre-training weights of Faster R-CNN, FCOS, and RetinaNet with the FT IR under the respective datasets LLVIP-C and FLIR-C. Evaluating the performance of such weight ensembling WiSE-OD$_{ZS}$ under the different corruption settings. Here, in the main manuscript, we show the results for Faster R-CNN in Table 3 for some values of λ, and we provide the detailed ablation and additional results in the supplementary material. It is important to mention that in the rest of the main manuscript, λ was fixed to 0.5, while here, we wanted to further investigate the potential of the WiSE-OD$_{ZS}$ over the different corruptions. In Table 3, the best results for in-domain performance were with $\lambda = 0.5$, while the best out-of-domain was $\lambda = 0.2$, which shows that for Faster R-CNN the zero-shot model can bring robustness for the model, but some corruptions, such as pixelate, are better when the λ is higher.

We observe that $\lambda = 0.5$ consistently provides a favorable trade-off, even for FLIR-C, it was the best ID. For LLVIP-C, WiSE-OD achieved the best mean performance under corruption ($\lambda = 0.2$, mPC 75.82), outperforming both the pure fine-tuned model ($\lambda = 1.0$, mPC 56.40) and the zero-shot model ($\lambda = 0.0$, mPC 40.96). Notably, $\lambda = 0.5$ performs best under heavy corruptions such as Gaussian noise, Fog, and Brightness shifts, scenarios where both FT and ZS individually struggle. For example, under Fog and Brightness, $\lambda = 0.5$ yields 84.51 and 82.10, respectively, while $\lambda = 1.0$ achieves only 50.90 and 35.36. This highlights the benefit of WiSE-OD in preserving complementary robustness features from both models. Interestingly, $\lambda = 0.8$ performs well in several cases but shows more variability, suggesting that moderate ensembling (rather than heavily biasing toward FT) is more robust under distribution shifts. These results justify the use of $\lambda = 0.5$ as a robust default and motivate future work on adaptive λ selection strategies.

7. Conclusion

In this work, we presented a new benchmark for IR OD robustness based on the work of Hendrycks and Dietterich [3], targeting traditional IR datasets such as LLVIP and FLIR, as well as real-world shifts such as M3FD. Our new benchmark is a challenging setting for IR robustness with the introduction of LLVIP-C and FLIR-C. Furthermore, we conducted an extensive study of different robust fine-tuning strategies over our proposed benchmark and in real-world OOD data. Additionally, we presented the WiSE-OD method and its variants WiSE-OD$_{ZS}$ and WiSE-OD$_{LP}$, both of which surpass traditional robustness strategies while also increasing in-domain performance across different detectors, such as Faster R-CNN, FCOS, RetinaNet, and YOLOv8. Our extensive study shows that our simple WiSE-OD strategy can mitigate performance drops without any additional training cost.

Main limitations. WiSE-OD assumes access to both a zero-shot and a fine-tuned (or linearly probed) model, which can be infeasible in constrained deployments. A fixed mixing coefficient ($\lambda = 0.5$) works well on average but is not uniformly optimal; we deliberately avoid tuning λ on held-out target data to reflect deployment realities, which may leave corruption-specific gains unrealized. The method also inherits weaknesses from its base models; for instance, if either the zero-shot or FT model underperforms, ensemble gains are limited.

Failure cases. Robustness degrades under extreme corruptions, like severe snow, heavy blur, very low brightness, and especially *low contrast* with FLIR-C most affected. In such scenes, both base models often miss or fragment objects, and the ensemble propagates these errors, yielding uncertain activations and unstable boxes. Mitigation likely requires corruption-aware ensembling or adaptive λ.

Future work. A natural extension is *adaptive* weight-space ensembling: predict λ per image or corruption using lightweight signals (e.g., confidence/entropy, or a small gating network). Integrate WiSE-OD with domain-generalization to better handle unseen IR conditions without target labels. Finally, evaluate additional sensors (depth, multispectral) to assess generality beyond IR.

Acknowledgments

This work was supported by Distech Controls Inc., the Natural Sciences and Engineering Research Council of Canada, the Digital Research Alliance of Canada, and Mitacs.

References

[1] Ayman Beghdadi, Malik Mallem, and Lotfi Beji. Benchmarking performance of object detection under image distortions in an uncontrolled environment. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2071–2075. IEEE, 2022. 3

[2] Thomas Dubail, Fidel Alejandro Guerrero Peña, Heitor Rapela Medeiros, Masih Aminbeidokhti, Eric Granger, and Marco Pedersoli. Privacy-preserving person detection using low-resolution infrared cameras. In *European Conference on Computer Vision*, pages 689–702. Springer, 2022. 2

[3] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 2, 3, 5, 8

[4] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 3

[5] Arthur Josi, Mahdi Alehdaghi, Rafael MO Cruz, and Eric Granger. Multimodal data augmentation for visual-infrared person reid with corrupted data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 32–41, 2023. 3

[6] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022. 2, 3

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[8] 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*, pages 2980–2988, 2017. 2

[9] Jinyuan Liu, Xin Fan, Zhanbo Huang, Guanyao Wu, Risheng Liu, Wei Zhong, and Zhongxuan Luo. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5802–5811, 2022. 4, 7

[10] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International journal of computer vision*, 128:261–318, 2020. 2

[11] Xiaoqiong Liu, Yunhe Feng, Shu Hu, Xiaohui Yuan, and Heng Fan. Benchmarking the robustness of uav tracking against common corruptions. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 465–470. IEEE, 2024. 2

[12] Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, and Hui Xue. Coco-o: A benchmark for object detectors under natural distribution shifts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6339–6350, 2023. 3

[13] Heitor Rapela Medeiros, Masih Aminbeidokhti, Fidel Alejandro Guerrero Peña, David Latortue, Eric Granger, and Marco Pedersoli. Modality translation for object detection adaptation without forgetting prior knowledge. In *European Conference on Computer Vision*, pages 51–68. Springer, 2024. 2

[14] Heitor Rapela Medeiros, Fidel A Guerrero Pena, Masih Aminbeidokhti, Thomas Dubail, Eric Granger, and Marco Pedersoli. Hallucidet: Hallucinating rgb modality for person detection through privileged information. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1444–1453, 2024. 1

[15] Heitor R Medeiros, Atif Belal, Srikanth Muralidharan, Eric Granger, and Marco Pedersoli. Visual modality prompt for adapting vision-language object detectors. *arXiv preprint arXiv:2412.00622*, 2025. 1, 2

[16] 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. 2, 3, 4, 5

[17] Anitha Ramachandran and Arun Kumar Sangaiah. A review on object detection in unmanned aerial vehicle surveillance. *International Journal of Cognitive Computing in Engineering*, 2:215–228, 2021. 2

[18] 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*, 28:91–99, 2015. 2

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6

[20] Michael Teutsch, Angel D Sappa, and Riad I Hammoud. Computer vision in the infrared spectrum: challenges and approaches. *Challenges and Approaches*, 2021. 1

[21] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 2

[22] Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR, 2022. 3

[23] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022. 2, 3

[24] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. *Dive into deep learning*. Cambridge University Press, 2023. 2

[25] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280. IEEE, 2020. 4

[26] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 1