# Unexplored Faces of Robustness and Out-of-Distribution: Covariate Shifts in Environment and Sensor Domains

Eunsu Baek Keondo Park Jiyoon Kim Hyung-Sin Kim
Seoul National University
{beshu9407, gundo0102, iamkjy, hyungkim}@snu.ac.kr

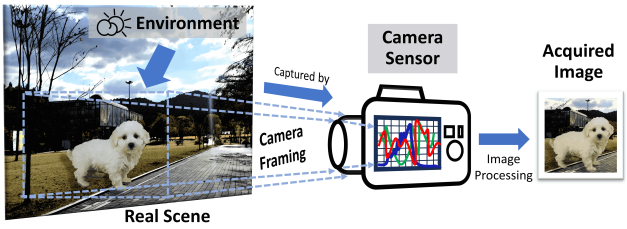
## Abstract

*Computer vision applications predict on digital images acquired by a camera from physical scenes through light. However, conventional robustness benchmarks rely on perturbations in digitized images, diverging from distribution shifts occurring in the image acquisition process. To bridge this gap, we introduce a new distribution shift dataset, ImageNet-ES, comprising variations in environmental and camera sensor factors by directly capturing 202k images with a real camera in a controllable testbed. With the new dataset, we evaluate out-of-distribution (OOD) detection and model robustness. We find that existing OOD detection methods do not cope with the covariate shifts in ImageNet-ES, implying that the definition and detection of OOD should be revisited to embrace real-world distribution shifts. We also observe that the model becomes more robust in both ImageNet-C and -ES by learning environment and sensor variations in addition to existing digital augmentations. Lastly, our results suggest that effective shift mitigation via camera sensor control can significantly improve performance without increasing model size. With these findings, our benchmark may aid future research on robustness, OOD, and camera sensor control for computer vision. Our code and dataset are available at https://github.com/Edw2n/ImageNet-ES.*

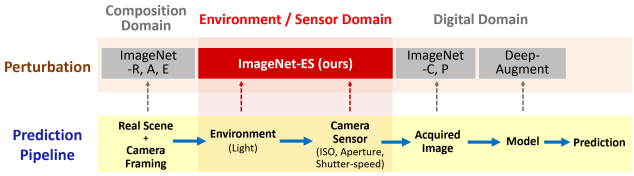
## 1. Introduction

The human vision system processes visual information by capturing light through the eyes and interpreting it within the brain. While proper training of our brains is undoubtedly crucial, addressing eyesight or light-related challenges necessitates equipping ourselves with customized- or sunglasses rather than relying solely on cognitive enhancement.

Similarly, in many computer vision frameworks, as depicted in Figure 1a, a camera serves as the 'eyes,' capturing authentic scenes through the play of light and generating digital images. These images are then interpreted by a deep neural network (*i.e.* the brain), as illustrated in Figure 1b. Continuous efforts towards improving AI systems to match




(a) Real-world image acquisition process. Variations in the environmental and camera sensor factors can cause significant covariate shifts in the acquired image.




(b) Real-world image prediction pipeline with existing perturbation benchmarks at each phase. *ImageNet-ES* first investigates the environment and sensor domains directly, instead of mimicking via digital perturbation.

Figure 1. Motivation and contribution of *ImageNet-ES*, the first benchmark on the necessary but unexplored faces of image covariate shifts: environment and camera sensor domains.

the robustness of the human vision system predominantly focus on the 'brain' component. Existing robustness benchmarks evaluate the resilience of model predictions against perturbations in *digitized* images [7, 10, 20, 25]. Various techniques, such as domain generalization/adaptation and out-of-distribution (OOD) detection, have refined deep learning models to handle distribution shifts [1, 6, 11, 13, 15, 19, 21, 22, 26, 30, 31, 37, 42, 42].

However, the implications of distribution shifts resulting from the *image acquisition process* (*i.e.* eyes), caused by variations in real-world light and camera sensor operations, remain unexplored. The absence of a benchmark introduces uncertainty regarding the generalizability of observed robustness in synthetic data to real-world applications. Moreover, the synergistic interplay between the camera sensor and the model has not been investigated. Therefore, current approaches may risk inefficiency by attempting to address eyesight/light problems through over-training the brain.

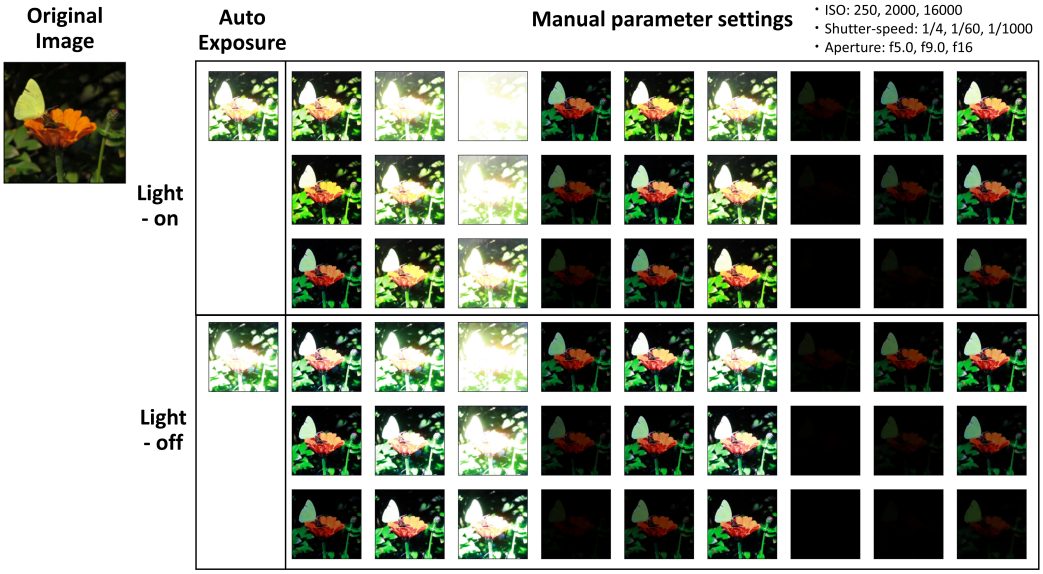



Figure 2. Representative Examples of *ImageNet-ES*. In contrast to conventional robustness benchmarks that rely on digital perturbations, we directly capture 202k images by using a real camera in a controllable testbed. The dataset presents a wide range of covariate shifts caused by variations in light and camera sensor factors.

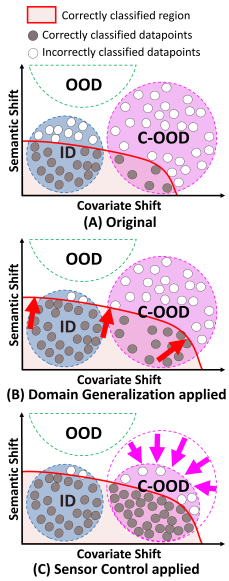



Figure 3. Robustness improvement scenario to cover real-world C-OOD

This work aims to narrow the gap between synthetic and real-world data by investigating the impact of environmental and camera sensor factors. Instead of relying on digital perturbation, we construct a controllable testbed, *ES-Studio*. This testbed allows us to directly capture images using a physical camera with varying sensor parameters (ISO, shutter speed and aperture) and different light conditions (on/off), resulting in a novel dataset called *ImageNet-ES*.

*ImageNet-ES* consists of 202k images covariate-shifted from 2,000 samples in TinyImageNet [38]. For example, Figure 2 shows 56 variations for a single sample, captured in *ES-Studio* under different light and camera sensor settings. These example images illustrate a broad spectrum of distribution shifts, suggesting that model robustness observed in conventional benchmarks might not necessarily generalize to our *ImageNet-ES* benchmark. Furthermore, some of the captured images even lose essential visual features due to severe perturbation, making them impractical for model prediction. This implies that, as shown in Figure 3, restricting distribution shifts in the image acquisition phase via camera sensor control can be more practical than solely focusing on model improvement.

With the *ImageNet-ES* dataset, we conduct an extensive empirical study on OOD detection and domain generalization. Furthermore, we explore the potential of camera sensor control in addressing real-world distribution shifts. Our study unveils a series of noteworthy findings as follows:

- **OOD definition:** Covariate-shifted data (C-OOD) have been categorized entirely as either OOD or in-distribution (ID). However, C-OOD data in *ImageNet-ES* exhibit widespread OOD scores in most metrics, including both ID and OOD. Model-Specific OOD (MS-OOD) [1] is more proper for fine-grained labeling of our C-OOD data.
- **OOD detection:** State-of-the-art (SOTA) OOD detection methods, focusing on distinguishing semantic shifts, falter in *ImageNet-ES*. OOD detection should be improved to incorporate real-world covariate shifts together.
- **Domain generalization:** Existing digital augmentations do not incorporate distribution shifts in *ImageNet-ES*. Learning environment/sensor-domain perturbations in *ImageNet-ES* with existing augmentations improves model robustness, even in conventional benchmarks.
- **Potential of sensor control:** Camera sensor control can significantly improve prediction accuracy by mitigating distribution shifts. With sensor control, EfficientNet can perform comparably to much heavier transformer models.
- **Direction of sensor control:** High-quality images in terms of model prediction do not necessarily align with human aesthetics but rather with what the model learns from training data. Sensor control should be grounded in the features that the model (not the human) prefers.

Overall, future research on OOD detection and model robustness requires more thorough evaluations, including environmental and camera sensor variations. Furthermore, it is valuable to explore camera sensor control so that acquired images contain more features preferred by the model.

## 2. Related Work

### 2.1. Robustness Benchmarks

A number of benchmarks have employed various digital perturbations to assess image classifier robustness or OOD

detection methods. Notably, ImageNet-C and -P [10] simulate environmental and adversarial perturbations through blur, noise, brightness, etc. ImageNet-A and -O [14] limit spurious cues using adversarial perturbations. Several datasets utilize visual renditions to change real scenes, such as art, cartoons, patterns, toys, paintings, etc. [13, 29, 36]. SVSF [13] or ImageNet-E [20] changes camera views or image compositions.

While these benchmarks aim to incorporate real-world distribution shifts, such as camera framing, their approaches are limited to the indirect simulation of actual shifts via perturbing already-acquired digital images. Recent studies have highlighted that SOTA OOD detection methods face challenges due to a lack of knowledge about the real-world OOD distributions [28] and experience performance degradation in near-OOD, shifted benchmarks [18]. Building on prior work, our *ImageNet-ES* dataset directly modifies physical light and camera sensor parameters, which provide another type of real-world distribution shifts and demystify the relationship between digital and physical manipulations.

### 2.2. Out-of-Distribution (OOD) Detection

Out-of-distribution (OOD) detection is the task of identifying test data that come from a distribution different from the distribution of training data, due to either semantic shift (S-OOD) or covariate shift (C-OOD) [18].

OOD studies have focused on detecting samples with semantic shifts (S-OOD) that do not belong to any of the classes present in the training set. A number of methods determine the OOD score based on the decision-making component of classifiers [11, 15, 21, 22]. These techniques are more robust when class-agnostic information needs to be carefully considered, but vulnerable to significant semantic shifts or overconfidence issues [30, 37]. To alleviate these problems, other methods calculate the OOD score based on features the model learned [6, 19, 30, 31, 37]. Rigorous efforts in this area have achieved nearly perfect performance.

However, prior work has relatively unexplored how to handle covariate-shifted (C-OOD) samples. A handful of studies have considered entire C-OOD examples as in-distribution (ID) to enhance classifier robustness against covariate shifts [39, 40, 42]. Some studies have taken opposite approaches, treating all C-OOD samples as OOD to make OOD detection more generalizable to non-semantic shifts [16]. To address the problem of the rough treatment of entire covariate-shifted data as ID or OOD, more recent studies provide fine-grained categorization of C-OOD samples into ID and OOD, based on their own definitions [1, 33]. Notably, Averly and Chao have proposed a unified criterion that incorporates both S-OOD and C-OOD data based on model prediction results, called Model-Specific OOD (MS-OOD) [1]. MS-OOD reveals the problems of existing methods but does not provide a solution.

Looking forward, OOD detection should be improved to reliably handle both S-OOD and C-OOD data with a well-defined OOD score and detection method, which requires support from proper benchmarks. *ImageNet-ES* can contribute to this aspect by providing realistic C-OOD samples.

### 2.3. Domain Generalization

Domain generalization focuses on improving the robustness of models to distribution shifts in testing domains. To this end, Hendrycks *et al*. identified that using larger models and artificial data augmentations (called DeepAugment) can improve model robustness [13]. While many augmentation techniques [5, 12, 41] have shown to improve the robustness, their evaluation scope is limited to digital corruptions. More recently, foundation models have demonstrated success in learning effective feature representations through architectural changes [24], discriminative self-supervised pre-training [9, 26, 44], or large uncurated data [26].

However, prior work has focused on digital distribution shifts (*e.g*. pixelate or gaussian noise etc.), scene and camera composition shifts. On the other hand, our *ImageNet-ES* addresses other types of distributional shifts, such as those arising from the image acquisition process.

## 3. Background

As illustrated in Fig. 1, an image is influenced by three primary aspects at the point of its capture. Firstly, the term *composition* pertains to the arrangement, organization, and layout of visual elements within the frame of the image. Composition is subject to dynamic alterations caused by the movement of objects, addition or removal of objects, or other modifications. Camera operations, such as zooming or tilting, can also impact the resulting image. Secondly, *environment* signifies lighting conditions. For example, light can be scattered by dust or smoke, leading to image blurring. The position and intensity of light can also affect the image's quality. Finally, *camera sensor* generates an image from the light. The captured image can dynamically fluctuate according to sensor parameter settings.

This work focuses on variations in environmental and sensor factors without changing the composition.

### 3.1. Camera Operation for Image Acquisition

Before digitization, image variations can be introduced during the camera's acquisition process, which involves the following steps: (1) light reception, (2) sensor conversion, (3) image signal processing, and (4) final image creation.

Firstly, light is captured from the scene and environment. This light, the primary source of image variation in photography, plays a crucial role in determining the quality and characteristics of the image. Next, the captured light hits the camera sensor, which converts the light into an electrical signal. The types and settings of the sensor can influence

the image, with different sensors responding differently to the same light conditions. The electrical signal undergoes processing by the camera's internal systems. This processing commonly includes operations such as noise reduction, white balance adjustment, and color grading. Finally, the processed signal is converted into an image.

While image signal processing techniques can introduce various perturbations and contribute to the quality improvement of the final image, these results are *fundamentally bounded* by the original electrical signal generated by the sensor from the lighting conditions. Therefore, despite numerous existing perturbations through post-processing, investigating the impact of environmental and sensor factors has additional value.

### 3.2. Light Factor in Environment Domain

In the environmental domain, changes in lighting conditions significantly impact the captured image. For example, an object photographed under bright overhead lighting may cast a strong shadow, altering the object's appearance. Similarly, an object photographed in low light may lack sufficient detail. Furthermore, changes in the color of the light, such as transitioning from daylight to artificial light, can affect how colors appear in the image. These variations present challenges for deep learning models, which often rely on consistent lighting for accurate image recognition.

### 3.3. Light Factor in Camera Sensor Domain

The camera sensor has three main parameters, ISO, shutter speed, and aperture, which influence light levels in an image while also impacting various aspects of the captured scene.

- **ISO** adjusts the sensitivity of the camera sensor to light. Higher ISO values increase brightness but may introduce additional noise to the image.
- **Shutter speed** governs the duration that the camera's shutter remains open. Slower shutter speed allows more light to reach the sensor, resulting in a brighter image but also motion blur. Conversely, faster shutter speed can produce a darker image and freeze motion.
- **Aperture** determines the size of the lens opening, regulating light entry and affecting the image's depth of field. A larger aperture brightens the image but leads to a shallower depth of field, concentrating focus on a limited portion of the scene.

While manual control of these parameters is possible, most cameras are equipped with automatic exposure control. The **auto exposure** function calculates the optimal exposure settings for a given scene. However, it is important to note that the optimal settings are for human aesthetic, which may not align with those optimal for model predictions.

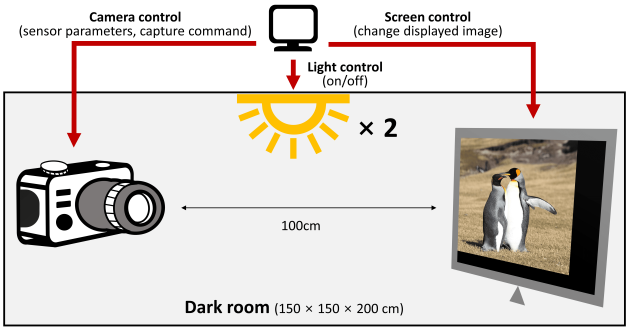



Figure 4. Illustration of the *ES-Studio* setup

## 4. *ES-Studio* and *ImageNet-ES*

To compensate the missing perturbations in current datasets, we construct a new testbed, ***ES-Studio*** (**E**nvironment and camera **S**ensor perturbation **Studio**). It can control physical light and camera sensor parameters during data collection. Utilizing *ES-Studio*, we compile ***ImageNet-ES***, a novel dataset comprising 202,000 samples of perturbed data from the environment and camera sensor domains.

### 4.1. *ES-Studio* Design Considerations and Setup

In constructing our data collection studio, we prioritize two main considerations: 1) ensuring reproducibility and 2) capturing real-world perturbations, particularly those related to light factors in both the environment and camera sensor domains. Specifically, being the first effort to gather such real-world perturbations, it is crucial for our data collection process to be reproducible, facilitating and promoting future research in this area. To achieve this, we have employed *ES-Studio*, providing individual control over environment and sensor parameters involved in image acquisition.

The construction of *ES-Studio* is depicted in Figure 4. First, we established a completely dark room to eliminate any external light during the data collection process. The room is equipped with four main components: (1) a large screen to display the reference dataset, (2) a camera with adjustable parameters for ISO, shutter speed, and aperture, (3) two ceiling lamps to manipulate environmental light, and (4) a desktop and Wi-Fi network to manage above components. More details of *ES-Studio* settings are in Appendix.

### 4.2. *ImageNet-ES* dataset

#### 4.2.1 Sampling Process for Target Datasets

We choose Tiny-ImageNet [38], a 200-class subset derived from ImageNet-1K, as our reference dataset. This dataset not only provides a diverse range of categories but also demands less computational power for experiments compared to ImageNet-1K. We randomly select ten images from each category in the validation set of Tiny-ImageNet. Subsequently, we divide these images into two halves, utilizing

Table 1. Environment and Sensor specifics of *ImageNet-ES* collection

| Dataset | Original samples | Light | Camera sensor | ISO | Shutter speed | Aperture | Captured images |
|---|---|---|---|---|---|---|---|
| Validation | 1,000<br>(5 samples/class) | On/Off | Auto exposure (5 shots)<br>Manual (64 options) | Auto<br>200/800/3200/12800 | Auto<br>(0"4')/(1/20')/(1/160')/(1/1250') | Auto<br>f5.0/f9.0/f13/f20 | 10,000<br>128,000 |
| Test | 1,000<br>(5 samples/class) | On/Off | Auto exposure (5 shots)<br>Manual (27 options) | Auto<br>250/2000/16000 | Auto<br>(1/4')/(1/60')/(1/1000') | Auto<br>f5.0/f9.0/f16 | 10,000<br>54,000 |

the first five for validation and the remaining five for testing purposes. To ensure visual fidelity, each sampled image maintains a resolution greater than $375 \times 500$ pixels, preventing distortion when displayed on the screen. In total, we systematically sample 2,000 images.

#### 4.2.2 Data Collection

Table 1 provides a comprehensive overview of the collected data. We display each sampled reference image on the screen and take its picture multiple times while varying the environmental and camera sensor factors.

We consider two options for the environmental factor: lights in the "on" and "off" states. For camera sensor control, we use both auto exposure and manual parameter settings. Under auto exposure, the camera autonomously determines each sensor parameter. Given that the auto-controlled parameters can be different at each time, we capture each sample five times to observe the average effect. For manual parameter setting, we use four different options for ISO, shutter and aperture during the validation split, and three options during the test split, leading to 64/27 variations in the validation/test split. To ensure the integrity of our data collection process, we implemented pauses between setting changes. Specifically, we introduced a one/seven/ten-second pause between each parameter option, between changes in light options, and between sample image changes, respectively. A detailed log is recorded for each image, serving debugging purposes.

#### 4.2.3 Data Processing and Validation

The next step involves cropping the valid image area from the collected images. The valid image area is determined through a systematic process: First, we display a visually discriminative reference image on the screen and capture the screen with the camera. We extract crucial information for the captured image, including the left top point, width and height of the reference image. Then, for other images, we determine the valid area of each image by using the digitally calculated ratio of each image to the reference image. Finally, we set the padding to the determined valid area and crop the captured image accordingly.

To validate the *ImageNet-ES* collection, we conduct a subjective validation approach. For each reference sample, we aggregate all images taken under different settings and concatenate them into a single image along with the original sample. This composite image is then reviewed by three individuals to ensure that all images are captured consistently. The validation process also confirms that the collected images align accurately with the original image.

## 5. Experiments

We design experiments to evaluate the impact of distribution shifts within the environmental and camera sensor domains. The experiments include widely used methods for OOD detection and domain generalization.

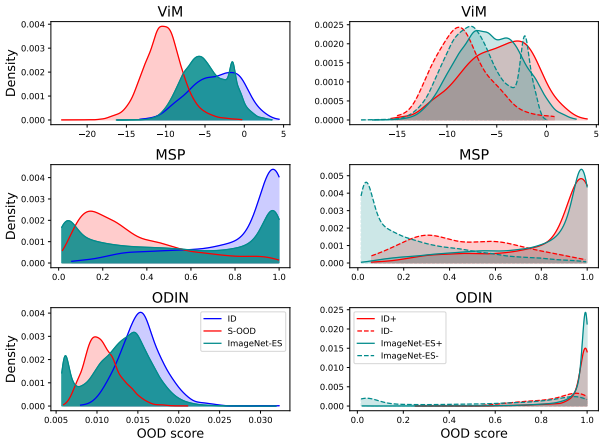
### 5.1. OOD Detection

We validate OOD detection techniques on *ImageNet-ES*: ViM [37], ReAct [30], ASH [6], MSP [11] and ODIN [21]. They report SOTA performance and serve as baseline methods in recent OOD studies [1, 42]. Likewise, EfficientNet-B0 [32] is selected as the underlying model for OOD detection, given its widespread use in OOD studies. Training details and evaluations for other models are in Appendix.

#### 5.1.1 Evaluation of OOD Definition

Most OOD detection techniques are developed under a framework that focuses on detecting samples with semantic shifts (S-OOD). Under this framework, all samples from *ImageNet-ES* (*i.e.* C-OOD data) should be classified as either OOD or In-Distribution (ID) in their entirety. To assess the validity of the semantics-centric OOD definition under *ImageNet-ES*, we analyze the distribution of OOD scores of ViM [37], MSP [11] and ODIN [21] on ID (Tiny-ImageNet [38]), S-OOD (Texture-O [4]) and *ImageNet-ES* (C-OOD) datasets, as in Figure 5a. While OOD detection techniques provide clearly distinguished OOD scores for ID (blue region) and S-OOD (red region), the scores on *ImageNet-ES* (green region) are widely spread across the entire spectrum between OOD and ID. The results show that treating entire C-OOD data in *ImageNet-ES* as either ID or OOD leads to significant detection errors. It is risky to directly apply the semantics-centric framework in the presence of C-OOD data.

We also evaluate an alternative framework, called MS-OOD (Model-Specific OOD) [1]. In this framework, OOD is defined by considering model-specific acceptance (MS-A) or rejection (MS-R): (1) MS-A includes ID and C-OOD samples that are correctly classified by the model, denoted as ID+ and C-OOD+. (2) On the other hand, MS-R includes all S-OOD samples, as well as ID and C-OOD samples that are misclassified by the model, denoted as ID− and C-OOD−. Within this MS-OOD framework, the objective




(a) Semantics-centric framework (b) MS-OOD framework

Figure 5. OOD score distribution with semantics-focused and MS-OOD frameworks. Tiny-ImageNet [38] and Texture [4] are used for the ID and S-OOD datasets, respectively. *ImageNet-ES* serves as a C-OOD dataset.

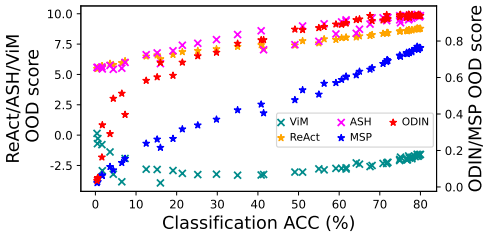



Figure 6. Each point represents the OOD score measured on the single parameter setting of *ImageNet-ES*.

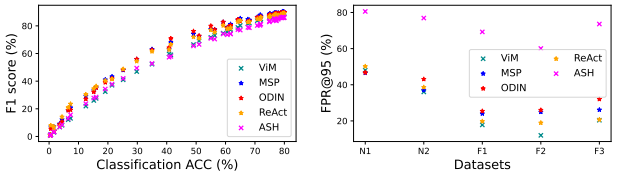
of OOD detection methods is to accept correctly predicted examples and reject incorrectly predicted examples.

Figure 5b presents the distribution of OOD scores measured on ID and *ImageNet-ES* (C-OOD) datasets under the MS-OOD framework. ViM [37], the current SOTA method for S-OOD detection, still exhibits a significant overlapping area between *ImageNet-ES+* and *ImageNet-ES−*. This confirms that methods developed to detect S-OOD cannot handle C-OOD data properly solely by modifying the underlying framework. On the other hand, MSP, an older method usually serving as a baseline, shows a clearer score separation between *ImageNet-ES+* and *ImageNet-ES−*.

### 5.1.2 Evaluation of OOD Detection Methods

Next, we evaluate 54 manual environmental/sensor variations in the *ImageNet-ES* test set in terms of classification accuracy and OOD scores. The OOD scores are obtained using five methods (MSP [11], ODIN [21], ReAct [30], ASH [6] and ViM [37]) within the MS-OOD framework.

Figure 6 showcases both accuracy and OOD scores for each setting, averaged over 1,000 samples out of 200 classes. Given that the MS-OOD framework defines OOD based on model prediction results, the OOD score is expected to increase with classification accuracy. Our results reveal that the older methods, MSP [11] and ODIN [21],




(a) C-OOD (*ImageNet-ES*) (b) S-OOD

Figure 7. Performance of OOD methods with C-OOD and S-OOD. (a) Each point is the F1 score measured on a parameter setting of *ImageNet-ES*. (b) N1: SSB-hard [35], N2: NINCO [3], F1: iNaturalist [34], F2: Texture [4], F3: Openimage-O [37]

provide relatively desirable correlation between accuracy and OOD score. In contrast, more recent methods (ASH [6], ReAct [30] and ViM [37]) demonstrate a weaker relationship between accuracy and OOD score. Particularly, ViM shows a rapid increase in OOD scores as accuracy approaches zero; ViM tends to accept numerous samples as ID even when they are misclassified by the model.

In addition, Figure 7a shows the OOD detection performance for each environmental/sensor setting of the *ImageNet-ES* test set, in terms of F1 score used in [1]. The results reveal that MSP and its advanced versions, ODIN and ReAct, consistently outperform ViM and ASH. Meanwhile the latest ViM and ASH emerge as the least effective among the five. The detection errors observed in ViM can be attributed to its inability to recognize unseen features in C-OOD. A more thorough explanation is in Appendix.

For further insight, Figure 7b employs five S-OOD datasets as benchmarks, comprising two for near-OOD (SSB-hard [35], NINCO [3]) and three for far-OOD (iNaturalist [34], Texture [4], and Openimage-O [37]). In contrast to the results in *ImageNet-ES* (Figure 7a), Figure 7b underscores that ViM provides superior performance compared to MSP, ODIN, ReAct and ASH, confirming the effectiveness of latest methods in detecting semantic shifts.

Overall, our findings suggest that the evolution of OOD detection methods over recent years might have been biased towards S-OOD handling, potentially retrograding in terms of C-OOD handling. As of now, there is no single method that excels in both C-OOD and S-OOD detection. Given the importance of addressing covariate shifts in real-world applications, future research on OOD detection should integrate and advance both S-OOD and C-OOD aspects.

## 5.2. Domain Generalization

In this section, we investigate the impact of domain generalization techniques on enhancing the robustness in the environment and sensor domains. As a baseline, we finetune ResNet-50 using only the subset of ImageNet (IN) that precisely matches the images corresponding to the validation split from *ImageNet-ES*, incorporating composition-related augmentations such as crop, resize and flip.

For comparison schemes, we employ both basic and ad-

Table 2. Evaluation with different robustness enhancing strategies. The result is based on ResNet-50. (IN: ImageNet)

| ID | Comp.aug | Basic digital aug | Advanced digital aug | Incl. *ImageNet-ES* | Eval dataset IN | IN-C | *ImageNet-ES* |
|---|---|---|---|---|---|---|---|
| 1 | ✓ | | | | 85.8 | 51.0 | 49.6 |
| 2 | ✓ | ✓ | | | 85.8 | 51.7 | 50.4 |
| 3 | ✓ | ✓ | ✓ | | 85.5 | 57.4 | 49.1 |
| 4 | ✓ | | | ✓ | **86.0** | 51.8 | **55.8** |
| 5 | ✓ | ✓ | | ✓ | 85.8 | 51.4 | 54.5 |
| 6 | ✓ | ✓ | ✓ | ✓ | 84.0 | **57.9** | 53.7 |

vanced digital augmentations. Basic augmentations include color-jitter, solarize and posterize, while advanced augmentations include DeepAugment [13] and AugMix [12]. To explore the potential of finetuning with our real-world perturbations, we replace half of the finetuning images with randomly sampled images from the validation set of *ImageNet-ES*. We exclude some images that are too far distorted to be identifiable, utilizing an image similarity metric LPIPS [43]. We evaluate each finetuning result on the test sets of IN, IN-C [10] and *ImageNet-ES*. Since *ImageNet-ES* contains only a subset of images from 200 classes from IN, we use the same subset of IN and IN-C corresponding to the test set of *ImageNet-ES* for fair comparison.

The evaluation results are summarized in Table 2. Experiment 2 shows that basic augmentations lead to improved accuracy in both IN-C and *ImageNet-ES*. However, our findings in experiment 3 contradict prior work [12, 13]. While more aggressive augmentations, such as AugMix and DeepAugment, significantly improve robustness on IN-C, these strategies result in performance degradation when predicting images with our real-world perturbations.

Experiments 4 to 6 evaluate the impact of learning augmentations in environmental and sensor domains in addition to conventional digital augmentations. Learning these real-world perturbations consistently improves robustness in *ImageNet-ES*, demonstrating its effectiveness in real-world applications. Furthermore, experiments 4 and 6 show that adopting environmental/sensor augmentations further improves robustness on the conventional benchmark IN-C.

In summary, these results verify that augmentations in the environmental and sensor domains can provide additional valuable information absent in conventional augmentation schemes. The impact becomes more significant in real-world applications involving cameras.

## 5.3. Sensor Parameter Control

To investigate the impact of sensor parameter control on model performance, we evaluate the performance of three different subsets of *ImageNet-ES* test split. **Auto exposure** includes 10,000 samples captured with the default auto exposure (AE) setting. **All params** includes 54,000 samples captured with 27 different manual parameters settings. **Best** includes 2,000 samples captured with the manual parameter setting that provides the best accuracy among the test split.

We employ several models for generalizability. The baseline is ResNet-50 [8], trained with a vanilla training scheme on ImageNet (IN)-1K. To explore whether well-configured sensor parameters could enhance the model performance to the level of domain-generalizated (DG) models, we also evaluate ResNet-50 trained on IN-21K with DeepAugment [13] and AugMix [12]. In addition, we examine whether a larger model demonstrates more robustness by evaluating ResNet-152.

Furthermore, we employ EfficientNet-B0/B3 [32] to test the lightweight model architecture's validity on *ImageNet-ES*. SwinV2-T/B [23] are chosen as representative models from transformer-based architecture, known for its robustness [2, 10]. OpenCLIP-b/h [17] and DINOv2-b/g [26] are selected as domain-generalized versions of SwinV2. All model weights are obtained from PyTorch [27], except for the DG version of ResNet-50, which is released in [13].

Table 3. Evaluation of various models on *ImageNet-ES*. (IN: ImageNet, AE: Auto exposure)

| Model | Num. Params | Pretraining Dataset | DG method | IN | *ImageNet-ES* AE | All params | Best |
|---|---|---|---|---|---|---|---|
| ResNet-50 [8] | 26M | IN-1K | - | 86.3 | 32.2 | 50.2 | 80.1 |
| | | IN-21K | DeepAugment [13] +AugMix [12] | 87.0 | 53.3 | 61.4 | 84.0 |
| ResNet-152 [8] | 60M | IN-1K | - | 87.6 | 41.1 | 54.3 | 83.3 |
| Efficientnet-B0 [32] | 5M | IN-1K | - | 88.1 | 51.4 | 58.1 | 83.8 |
| Efficientnet-B3 [32] | 12M | IN-1K | - | 88.3 | 62.0 | 66.2 | 86.8 |
| SwinV2-T [23] | 28M | IN-1K | - | 90.7 | 54.2 | 63.1 | 86.8 |
| SwinV2-B [23] | 88M | IN-1K | - | 92.0 | 60.1 | 65.6 | 89.0 |
| OpenCLIP-b [17] | 87M | LAION-2B | Text-guided pretrain | 94.3 | 66.3 | 71.0 | 92.7 |
| OpenCLIP-h [17] | 632M | LAION-2B | Text-guided pretrain | 94.7 | 79.1 | 77.6 | 94.7 |
| DINOv2-b [26] | 90M | LVD-142M | Dataset curation | 93.6 | 74.5 | 73.9 | 92.2 |
| DINOv2-g [26] | 1.1B | LVD-142M | Dataset curation | 94.7 | 84.3 | 79.6 | 94.2 |

### 5.3.1 Potential of Sensor Control

Table 3 summarizes the results. Firstly, DG techniques and pretraining on larger datasets consistently improve robustness on *ImageNet-ES*. For instance, in the All params case, DG version of ResNet-50 improves the test accuracy to plain ResNet-50 by 11.2.

In addition, sensor parameter tuning turns out to be as important as domain generalization and model size. First, the auto exposure option degrades performance of all models compared to the accuracy on original images (IN); the current auto exposure does not provide optimal parameters for model predictions. Conversely, the Best case reveals that with well-tuned sensor parameters, performance can be improved remarkably. The Best case improves prediction accuracy over the Auto exposure case by **9.9**$\sim$**47.9** and the All params case by **14.6**$\sim$**29.9**. Surprisingly, EfficientNet-B0 with the Best even outperforms OpenCLIP-h in the Auto exposure and All params cases, despite OpenCLIP-h having around 120$\times$ more parameters, learning from 160$\times$ more training data and employing domain generalization.

These findings suggest that research might have overemphasized model improvement, possibly overlooking the importance of proper input data generation. However, mitigating distribution shifts through sensor control proves to be valuable regardless of model size and DG techniques. The performance gain from sensor control can even surpass

Table 4. Impact of environmental factor. The difference is calculated between the accuracy measured when light is on and off. We provide the difference for auto exposure setting and the maximum of difference amongst all manual parameter settings. More details could be found in Appendix.

| Model | ResNet-50 | ResNet-152 | SwinV2-B | DINOv2 |
|---|---|---|---|---|
| Auto exposure | 4.4 | 3.0 | 4.3 | 2.4 |
| Max. of all params | 11.7 | 14.9 | 15.7 | 18.2 |

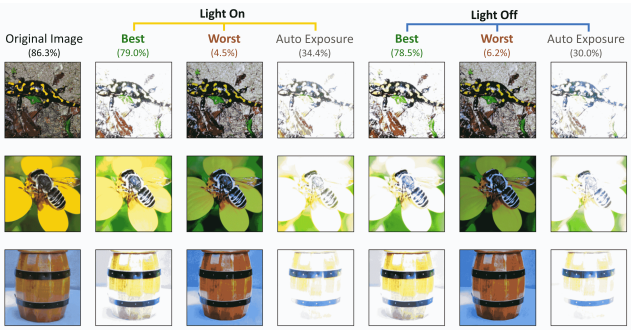



Figure 8. Qualitative results of *ImageNet-ES*: AE: Auto Exposure, Best/Worst: Sampled images from the parameter setting whose accuracy is highest 5 or lowest 5. Accuracy on ResNet-50 is also presented for each group.

that achieved through larger model size, more training data, advanced architectures, and DG techniques. If designed efficiently, sensor control can be an enabling factor of mobile applications where computational resources are limited.

#### 5.3.2 Direction of Sensor Control

Table 4 summarizes the impact of lighting conditions (on and off) on model accuracy. Specifically, the table shows the difference in the model accuracy caused by changes in lighting when using the same sensor parameter option. Notably, the auto-exposure option fails to ensure consistent model performance, showing accuracy differences of 2.4∼4.4 across diverse models. Moreover, manual configuration of sensor parameters is also susceptible to the impact of environmental variables, resulting in an accuracy variance of up to 18.2. Our results pose a challenge for future research on sensor control. It should focus on adaptively controlling parameters based on given environments, instead of attempting to find universally optimal parameters.

To obtain further insights into sensor control, we perform a qualitative analysis on *ImageNet-ES*, as described in Figure 8. Interestingly, these examples show that images visually appealing to humans do not necessarily yield the best prediction outcomes. Images captured with the Best options sometimes appear significantly lighter, posing a challenge for human observers to discern the underlying semantic information. Conversely, some images captured with the worst-performing options are more conducive to semantic interpretation by humans. Lastly, images captured by controlled parameters are occasionally predicted more accurately than the original samples. These results imply that sensor control research should prioritize features that the model can leverage effectively, rather than relying solely on human intuition. A more detailed examination of the camera parameters is in Appendix.

## 6. Conclusion and Future Work

In this study, we investigated distribution shifts resulting from perturbations in both environmental and sensor domains. To achieve this goal, we established a controllable testbed, named *ES-Studio*, for image acquisition across diverse environmental and camera sensor configurations. We curated a new dataset of 202k images, called *ImageNet-ES*.

Employing *ImageNet-ES*, we have conducted comprehensive studies in OOD detection, domain generalization and camera sensor control. With respect to OOD detection, our findings indicated limitations in the widely used semantics-centric framework. We proposed that OOD detection should extend its scope to incorporate both S-OOD and C-OOD. Regarding domain generalization, we demonstrated that environmental and sensor-related augmentations offer additional useful information to the model, improving robustness in both conventional and *ImageNet-ES* benchmarks. Finally, we discovered that well-tuned sensor parameters can enable a lightweight, basic model to perform comparably to or better than more advanced models. Sensor control necessitates a model-centric design instead of relying solely on human aesthetics. We hope that these insights will inspire future research and *ImageNet-ES* will be utilized in addressing real-world distribution shifts.

**Limitations and Future Work.** Taking photos of displays in *ES-Studio* is reproducible, controllable, and scalable, but it might not fully consider the interaction between real 3D objects and light sources, and the non-luminous properties of real objects. It would be more realistic to replace displays with real objects or printed photos. Sensor control can support applications like autonomous driving and surveillance cameras, which require image capture in various environments. But since it primarily handles physical light, it needs to be combined with digital post-processing. For practical training of a neural net for sensor control, gradients need to be computed without extra photos.

## Acknowledgements

This research was supported in part by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00223530), and in part by the National Research Foundation (NRF) of Korea grants funded by the Korea government (MSIT) (No. RS-2023-00212780, No. RS-2023-00222663 and No. RS-2023-00265147). Hyung-Sin Kim is the corresponding author.

## References


[1] Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. *International Conference on Computer Vision (ICCV)*. 1, 2, 3, 5, 6

[2] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021. 7

[3] Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*, 2023. 6

[4] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5, 6

[5] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[6] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. 2023. 1, 3, 5, 6

[7] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[10] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*, 2019. 1, 3, 7

[11] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. 2017. 1, 3, 5, 6

[12] Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 3, 7

[13] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021. 1, 3, 7

[14] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 3

[15] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. pages 8759–8773, 2022. 1, 3

[16] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951–10960, 2020. 3

[17] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 7

[18] Bor-Chun Chen Ashish Shah Philip H.S. Torr Puneet K. Dokania Ser-Nam Lim Jishnu Mukhoti, Tsung-Yu Lin. Raising the bar on the evaluation of out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 4365–4375, 2023. 3

[19] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018. 1, 3

[20] Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, and Hui Xue. Imagenet-e: Benchmarking neural network robustness via attribute editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20371–20381, 2023. 1, 3

[21] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. 2018. 1, 3, 5, 6

[22] 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. 1, 3

[23] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 7

[24] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 3

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3, 7

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 7

[28] Yonggang Zhang-Jing Zhang Chen Gong Tongliang Liu Bo Han Qizhou Wang, Feng Liu. Watermarking for out-of-distribution detection. 2022. 3

[29] Evgenia Rusak, Steffen Schneider, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Imagenet-d: A new challenging robustness dataset inspired by domain adaptation. In *ICML 2022 Shift Happens Workshop*, 2022. 3

[30] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021. 1, 3, 5, 6

[31] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1, 3

[32] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 5, 7

[33] Junjiao Tian, Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Exploring covariate and concept shift for out-of-distribution detection. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 3

[34] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[35] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022. 6

[36] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. 3

[37] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching supplementary material. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930, 2022. 1, 3, 5, 6

[38] Jiayu Wu, Qixiang Zhang, and Guoxi Xu. Tiny imagenet challenge. *Technical report*, 2017. 2, 4, 5, 6

[39] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021. 3

[40] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu. Full-spectrum out-of-distribution detection. *International Journal of Computer Vision*, pages 1–16, 2023. 3

[41] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[42] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *arXiv preprint arXiv:2306.09301*, 2023. 1, 3, 5

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 7

[44] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. *International Conference on Learning Representations (ICLR)*, 2022. 3