ADAPTIVE CAMERA SENSOR FOR VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Domain shift remains a persistent challenge in deep-learning-based computer vision, often requiring extensive model modifications or large labeled datasets to address. Inspired by human visual perception, which adjusts input quality through corrective lenses rather than over-training the brain, we propose *Lens*, a novel camera sensor control method that enhances model performance by capturing high-quality images from the model's perspective, rather than relying on traditional human-centric sensor control. *Lens* is lightweight and adapts sensor parameters to specific models and scenes in real-time. At its core, *Lens* utilizes *VisiT*, a training-free, model-specific quality indicator that evaluates individual unlabeled samples at test time using confidence scores, without additional adaptation costs. To validate *Lens*, we introduce *ImageNet-ES Diverse*, a new benchmark dataset capturing natural perturbations from varying sensor and lighting conditions. Extensive experiments on both ImageNet-ES and our new *ImageNet-ES Diverse* show that *Lens* significantly improves model accuracy across various baseline schemes for sensor control and model modification, while maintaining low latency in image captures. *Lens* effectively compensates for large model size differences and integrates synergistically with model improvement techniques.

1 INTRODUCTION

Domain shift, the distribution gap between training and test data, is a well-known challenge that degrades the performance of deep-learning-based computer vision models. Existing solutions mainly focus on either model generalization (Hendrycks et al., 2021; 2019; Sohn et al., 2020; Zhou et al., 2023; Ganin et al., 2016; Cherti et al., 2023; Liu et al., 2021; 2022; Oquab et al., 2023) or model adaptation (French et al., 2017; Sun & Saenko, 2016; Gong et al., 2024; Yuan et al., 2023; Wang et al., 2022b), which require modifying the model itself. However, these approaches typically necessitate significant changes to the model and access to large, labeled target datasets, making them costly, time-consuming, and impractical for real-time applications on resource-constrained devices.

In contrast, human visual perception operates through a fine-tuned interplay between the eyes (sensors) and the brain (model). The eyes function as precise sensors, capturing visual data, while the brain processes and interprets it. When visual input is compromised, whether by blurriness or glare, the typical response is to improve the quality of the input through corrective lenses, sunglasses, or magnifying lenses, rather than retraining the brain to interpret flawed images better. This analogy highlights that **the model is not all you need**; acquiring high-quality images through camera sensors is essential to mitigate covariate shifts and improve visual perception.

Despite existing sensor controls like auto-exposure, which are optimized for human perception, we argue that camera sensor control designed for high-quality image acquisition to improve *model perception* requires a fundamentally different approach. Furthermore, in dynamic environments and on resource-constrained devices, sensor control mechanisms must be able to quickly adapt to varying scenes. To address these issues, we introduce *Lens* (Figure 1), a novel adaptive sensor control system that captures high-quality images robust to real-world perturbations. The core idea of *Lens* is to identify optimal sensor parameters that allow the target neural network to better discriminate between objects, akin to adjusting a pair of glasses for clear vision. *Lens* achieves this by leveraging *VisiT* (Vision Test for neural networks), a training-free, model-specific quality indicator that operates on individual unlabeled samples at test time without additional adaptation costs. *VisiT* assesses data quality based on confidence scores tailored to the target model, ensuring high-quality data without the need for extensive retraining or data collection. By acquiring the most discriminative images for the target model, *Lens* significantly boosts model accuracy without requiring model modification.

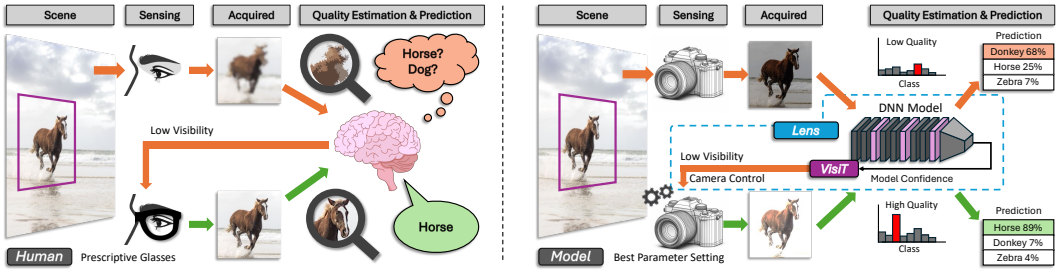


Figure 1: The concept of *Lens*: *Lens* mimics the human vision system, where eyesight quality can be improved through visual sensor control, such as glasses. It leverages sensor parameter adjustments to acquire higher-quality images, thereby enhancing model accuracy.

To demonstrate the effectiveness of *Lens* in realistic sensor control environments, we construct a testbed *ES-Studio Diverse*, where images are captured using a physical camera with varying sensor parameters and light conditions. Using this setup, we create a new dataset called *ImageNet-ES Diverse*, including 288,000 images that capture diverse natural covariate shifts via variations in sensor and light settings, based on 1,000 samples from TinyImageNet (Le & Yang, 2015).

As the first in-depth study on model-centric sensor control, we thoroughly evaluate *Lens* across two benchmarks – ImageNet-ES (Baek et al., 2024) and our newly created *ImageNet-ES Diverse* – using multiple model architectures. We compare *Lens* against various baselines, including human-targeted or random sensor control methods, domain generalization techniques, and lightweight test-time adaptation (TTA) methods. Our results show that *Lens* with *VisiT* significantly outperforms these methods, improving accuracy by up to **51.31%** while effectively reducing image capture time to only **0.6 seconds**, making it fast enough for real-time operation. The effect of sensor control even compensates for a model size difference of **up to 50**$\times$. Additionally, an ablation study on the quality estimator shows that *VisiT* outperforms state-of-the-art out-of-distribution (OOD) scoring methods, validating confidence scores as an effective quality proxy. Our qualitative analysis further supports these findings with visual insights.

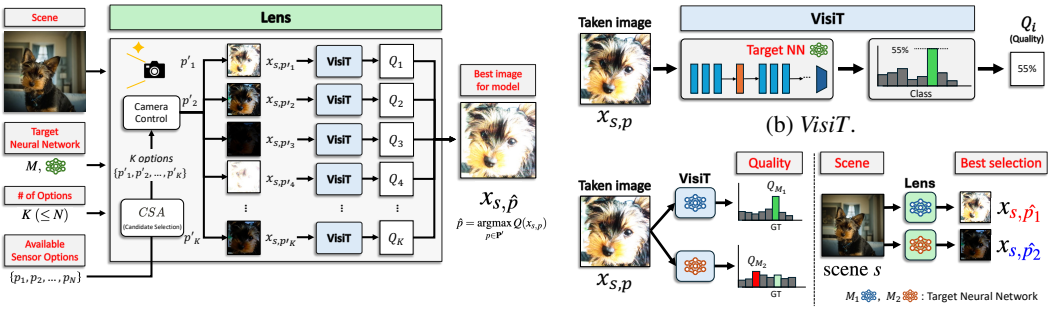
Our key contributions are as follows:

- We introduce *Lens*, a simple yet effective adaptive sensor control method that evaluates data quality from the model's perspective and optimizes camera parameters to improve model accuracy.
- *Lens* adopts VisiT (Vision Test for neural networks), a training-free, model-aware quality indicator that operates on individual unlabeled samples at test time, estimating data quality based on confidence scores.
- We release a new benchmark dataset, *ImageNet-ES Diverse*, containing 288,000 images that capture natural covariate shifts through varying sensor and lighting conditions.
- Our extensive experiments not only highlight the superiority of *Lens* in various scenarios but also reveal valuable insights for future research: (1) Sensor control can significantly improve model accuracy without model modification. (2) Sensor control can synergistically integrate with model improvement techniques. (3) Sensor control must be tailored in a model- and scene-specific manner. (4) High-quality images for model perception differ from those optimized for human vision.

2 Related Work

2.1 Model Improvement: Handling Domain-Shifted Input Data

Frequent domain shifts pose a significant challenge when deploying neural networks in dynamic real-world environments. Although traditional studies have aimed to improve a model's generalizability or adaptability, these methods place a computational burden, particularly for resource-constrained devices operating in real-time applications. **Domain generalization** techniques (Hendrycks et al., 2021; 2019; Sohn et al., 2020; Zhou et al., 2023; Cherti et al., 2023; Liu et al., 2021; 2022; Oquab et al., 2023) aim to train models to handle diverse data distributions, but typically results in significantly larger and more complex models. **Domain adaptation** approaches (Ganin et al., 2016; French et al., 2017; Sun & Saenko, 2016) adapt models to a specific target domain, which necessitates frequent retraining and the collection of substantial amounts of labeled target data.



(a) *Lens*. (b) *VisiT*. (c) Model-specific design of *VisiT* and *Lens*.

Figure 2: Workflow of *Lens*.

To address the need for lightweight, real-time adaptation without the cost of labeling, **test-time adaptation (TTA)** (Nado et al., 2020; Schneider et al., 2020; Wang et al., 2021) methods have been developed, allowing models to adjust to new domains using a small amount of unlabeled target data with unsupervised objectives. However, these lightweight TTA methods can lead to model collapse when faced with rapidly changing environments.

Lastly, a fundamental limitation of these model-centric techniques is their inability to address the data acquisition process itself. They struggle to cope with severe domain shifts that stem from low-quality data, such as images captured in over-exposed or low-light conditions (Baek et al., 2024).

2.2 Input Data Improvement: Mitigating Domain Shifts

To address domain shifts through improved data quality, camera sensor control has recently gained attention. Unlike traditional camera auto-exposure methods designed for human perception (Kuno et al., 1998; Liang et al., 2007), this new research focuses on optimizing sensor inputs specifically for deep-learning models. However, the absence of suitable benchmark datasets led early work to rely on camera sensor simulation (Paul et al., 2023), which falls short in generalizing to real-world domain shifts. Although some research has explored the control of physical camera sensors (Odinaev et al., 2023; Onzon et al., 2021), these efforts have been limited to highly-constrained environments with only a narrow range of exposure options.

To overcome these shortcomings, the ImageNet-ES dataset (Baek et al., 2024) was introduced, capturing domain shifts in real-world conditions by employing a physical camera with varying sensor parameters, such as ISO, shutter speed, and aperture. While the ImageNet-ES dataset demonstrates the potential of sensor control in addressing covariate shifts, identifying the optimal sensor parameters for specific models remains an open challenge. Furthermore, additional benchmark datasets are needed to enhance the generalizability of emerging control mechanisms. To the best of our knowledge, this work offers the first comprehensive exploration on camera sensor control using realistic benchmarks, including ImageNet-ES and our newly introduced *ImageNet-ES Diverse* dataset.

3 *Lens*: Adaptive Glasses for Vision Models

We introduce *Lens*, a post-hoc, adaptive, and camera-agnostic sensor control system for neural networks, designed to adaptively respond to dynamic scene characteristics. The key idea behind *Lens* is to identify the optimal sensor control parameters that capture images in a way that enhances the target model's ability to discriminate features–both in a model-specific and scene-specific manner–akin to adjusting a pair of prescription glasses to provide clear vision tailored to an individual's needs and environment. By **focusing solely on sensor parameter adjustments** and avoiding any modifications to the model itself, *Lens* prevents model collapse and catastrophic forgetting, ensuring reliable performance across varying domains. Moreover, it is lightweight and efficient in terms of both computation and memory. To achieve this, we propose *VisiT* (Vision Test for Neural Networks), a lightweight vision tester integrated into *Lens* that evaluates whether the images captured by the camera sensor are optimally suited for the target model and scene. *VisiT* operates during test time on individual unlabeled samples without modifying the target model.

3.1 OVERALL FRAMEWORK

Figure 2a illustrates the overall framework of *Lens*, which operates with a target neural network M that supports batch inference and a camera sensor equipped with a set of N available parameter options, $\mathbf{P} = \{p_1, p_2, \ldots, p_N\}$. Let $x_{s,p}$ represent the image captured by the camera from a target scene s using a sensor parameter option p. The goal of *Lens* is to select the optimal sensor parameter $\hat{p}$ such that the captured image $x_{s,\hat{p}}$ maximizes the accuracy of the target model's interpretation of the scene s. Let $Q(x_{s,p}; M)$ denote the quality estimate for image $x_{s,p}$ in the context of model M. The optimal parameter option $\hat{p}$, as selected by *Lens*, can be represented as:

$$\hat{p} = \arg\max_{p \in \mathbf{P}} Q(x_{s,p}; M)$$

Model- and Scene-Specific Sensor Control. *Lens* adaptively selects the optimal sensor parameter $\hat{p}$ for each model and scene in real-time, rather than relying on a globally fixed parameter determined through offline training for all models and/or scenes. The key insight is that different models have distinct ways of extracting and prioritizing features for scene interpretation. As shown in Figure 2c, two different models can perceive the same captured image differently (left side of the figure), leading to different optimal parameters for each model, even for the same scene (right side of the figure). Similarly, even with a fixed model, each scene contains unique features that are crucial for accurate prediction (further discussed in the Appendix D). As a result, the optimal sensor parameter is likely to vary for each specific combination of model and scene (Baek et al. (2024)).

***VisiT* (Lightweight Vision Test for neural networks).** *Lens* incorporates *VisiT* (Figure 2b) to estimate $Q(x_{s,p}; M)$, which represents the quality of an unlabeled captured image $x_{s,p}$ when interpreted by the target model M. *VisiT* is designed for real-time applications, operating as a lightweight and training-free module at test time, providing model-specific quality estimates for unlabeled images. To achieve the design goal, it is essential to determine an appropriate metric as a proxy for image quality. Specifically, we utilize the model's **confidence score** for its prediction on the image $x_{s,p}$ as a simple yet effective proxy for image quality, which will be further discussed in Section 3.2.

CSA (Candidate Selection Algorithm). The latency of *Lens* in selecting the optimal parameter highly depends on the camera sensor's latency to capture multiple images for different candidate parameter options. While capturing and evaluating images for all N available parameter options would provide the highest accuracy, it introduces significant latency for a single scene prediction, which is undesirable for real-time operation. To address the issue, *Lens* uses CSA to select a subset of the full parameter set $\mathbf{P}$, denoted as $\mathbf{P}' = \{p'_1, \ldots, p'_K\}$, as the candidate options. The number of candidate options, $K(\leq N)$, can be determined based on the system's need to balance computational cost with accuracy. Note that, since *Lens* operates with batch inference, capturing multiple images doesn't incur additional inference costs.

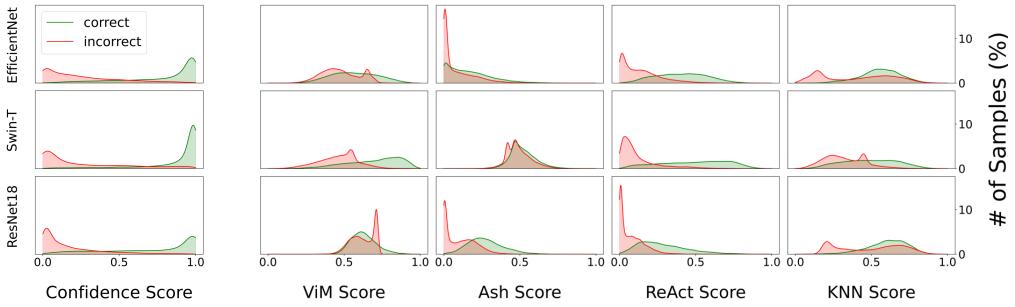
A crucial aspect of CSA is minimizing capture latency without sacrificing accuracy when selecting K candidate options. For example, sensor parameters like shutter speed significantly impact the capture time. Therefore, within the same time budget, it may (or may not) be more beneficial to prioritize multiple high-shutter-speed options over a single low-shutter-speed option, depending on the specific target scene and model. We explore this trade-off by implementing several simple algorithms for CSA and evaluating their performance in Section 5.

3.2 VISIT: LIGHTWEIGHT VISION TEST FOR NEURAL NETWORKS

In this subsection, we provide a detailed description of *VisiT*, the real-time image quality estimator for unlabeled test-time data. The key requirements for *VisiT* design are: (1) Alignment with correctness: The quality estimator must reliably indicate whether the model can accurately predict the sample. (2) Label-free operation: It must function with unlabeled data provided during test time (3) Single-sample assessment: The estimator should be capable of evaluating each image sample independently and immediately. (4) Lightweight operation: It should involve minimal computational overhead, ensuring seamless integration into sensor control pipeline.

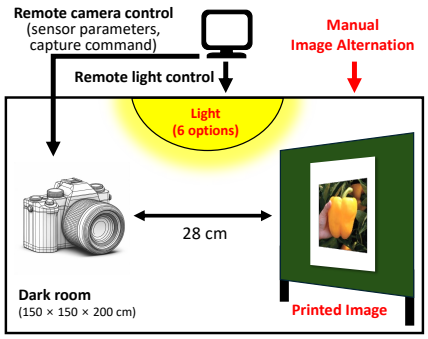
Confidence as a Proxy for Image Quality Assessment. To meet the requirements, we propose using the *confidence score* as a simple yet effective proxy. For a sample image x and target model M, the confidence score is defined as:

$$Confidence(x; M) = \max_{c \in \mathbf{C}} \text{Softmax}(f_M(x))_c$$



(a) Confidence based. (b) OOD (Out-of-Distribution) score based.

Figure 3: Quality indicators as proxies for image quality assessment: Each score is normalized between 0 to 1.



(a) *ES-Studio Diverse*.

(b) Specifics of data collection scheme.

Dataset	Original samples	Light	Camera sensor	ISO	Shutter speed	Aperture	Captured images
Test	1,000 (5 samples /class)	6 options, L1-L7 (w/o L5)	AE (5 shots)	Auto	Auto	Auto	45,000
			M (27 options)	250/2000/ 16000	(1/4")/(1/60")/ (1/1000")	f5.0/f9.0/ f16	243,000

AE: Auto Exposure, M: Manual

(c) Light options.

Light intensity	L1	L2	L3	L4	L6	L7	L5 (excluded)
Left	255	127	255	0	127	0	0
Right	255	127	0	255	0	127	0

Figure 4: Environment and sensor specifics of *ImageNet-ES Diverse*.

where $\mathbf{C}$ is the set of all possible classes, and $f_M(x)$ represents the output logits of the model M before applying the softmax function. The confidence score reflects how certain a model is about its predictions and has been widely used in tasks such as pseudo-labeling, consistency regularization, and high-quality image selection in semi- and self-supervised learning (Oliver et al., 2018; Sajjadi et al., 2016; Sohn et al., 2020; Lee et al., 2013; Cui et al., 2022; Chen et al., 2020; Xie et al., 2020). It is particularly well-suited for real-time applications, as it requires only inference on a sample without incurring additional computational overhead, such as training.

Correlation between Proxies and Image Quality. We conducted an experiment to evaluate the correlation between various proxies and image quality under real-world covariate shifts, using the ImageNet-ES validation dataset (Baek et al., 2024) (details in Appendix E.2). We compared our confidence score with out-of-distribution (OOD) scores, commonly used to identify OOD samples, across three models: EfficientNet (Tan & Le, 2019), Swin-T (Liu et al., 2021), and ResNet18 (He et al., 2016). The OOD scores were sourced from four state-of-the-art methods: ViM (Wang et al., 2022a), ASH (Djurisic et al., 2023), ReAct (Sun et al., 2021), and KNN (Sun et al., 2022)).

As shown in Figure 3, OOD scores tend to overlap between correct and incorrect samples across all OOD techniques and models, suggesting that OOD scores are not always reliable indicators of image quality. This discrepancy arises because OOD scores are primarily designed to detect semantic shifts (high-level features), but are less effective in identifying covariate shifts, which reflect variations in low-level features. In contrast, samples with higher confidence scores have a greater likelihood of being correct, while those with lower confidence scores are more likely to be incorrect. These results underscore the effectiveness of confidence scores as a reliable proxy for image quality.

4 *ImageNet-ES Diverse*: A New Real-World Benchmark

Lens improves image quality by dynamically controlling camera sensor settings, such as ISO, shutter speed, and aperture, to optimize environmental light for each scene. The quality of an image is significantly influenced by the amount and distribution of light within a scene, which depends on both the characteristics of the objects and the surrounding environment. Therefore, it is essential to evaluate the robustness of *Lens* across various scene characteristics. While the recent ImageNet-ES (Baek et al., 2024) dataset captures real-world scenes with varying sensor parameters, it is limited to only

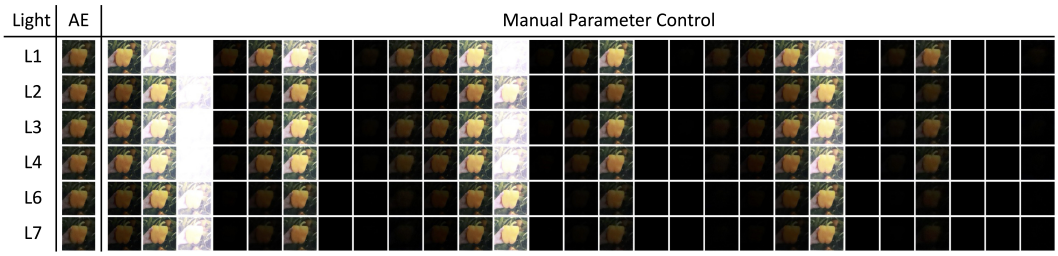


Figure 5: Representative examples of our *ImageNet-ES Diverse* dataset.

two lighting conditions. Furthermore, as it features images displayed on a screen – representing *light-emitting objects* (e.g., traffic lights) – the impact of ambient light conditions can be restricted.

To rigorously evaluate *Lens*, a new benchmark dataset is necessary to complement ImageNet-ES and effectively capture the impact of diverse environmental perturbations. To this end, we developed *ImageNet-ES Diverse*, a more versatile dataset with 288,000 samples of *non-illuminous objects* taken with a physical camera on a customized testbed called *ES-Studio Diverse* (Figure 4a). This dataset includes various sensor parameter settings (Figure 4b) and a broader range of lighting conditions (Figure 4c). As illustrated in Figure 5, *ImageNet-ES Diverse* unveils how sensor control interacts with diverse scene characteristics, valuable not only for *Lens* evaluation but also for future research exploring the effects of sensor settings and light conditions. Further details are in the Appendix C.

5 Experiments

We design experiments to evaluate the impact of *Lens*, which is the first approach to introduce **model- and scene-specific sensor control**, in comparison to traditional model-adjustment solutions that completely overlook image capture pipelines and focus solely on over-training for optimizing prediction accuracy under real-world perturbations. Our experiments are conducted across various model architectures, including widely used methods for domain generalization and test-time adaptation.

Datasets. We utilize the test sets of **ImageNet-ES** (Baek et al., 2024) and our new ***ImageNet-ES Diverse***, both derived from Tiny ImageNet (Le & Yang, 2015) (TIN). These datasets encompass extensive natural perturbations in both environmental and sensor domains including 27 manual controls and 5 auto-exposure shots. ImageNet-ES focuses on luminous objects, while *ImageNet-ES Diverse* features non-luminous objects, allowing them to complement each other effectively. This diversity allows us to validate our approach across a wide range of real-world covariate shifts. More details about each dataset are in the Appendix E.1.

Baselines and Oracles in the Image Acquisition Pipeline. For performance comparison, we consider two baselines and two oracles within the data acquisition pipeline. The first baseline, Auto-Exposure **(AE)**, is a commonly used sensor control designed to optimize images for human perception, though not necessarily for computer vision models. The second baseline, called **Random**, randomly selects parameter settings, and we calculate its performance as the average over all available options. To explore the potential of model- and scene-specific parameter control, we introduce two oracles: Oracle-Specific (**Oracle-S**) and Oracle-Fixed (**Oracle-F**). Oracle-S ideally selects the best sensor parameter for each sample and model, representing the upper bound for *Lens*. Oracle-F, on the other hand, serves as the upper bound for fixed parameter settings, without considering model-scene interactions. The best global parameter option for Oracle-F is selected based on the average accuracy across all models in Table 1 and all scenes in both datasets.

5.1 Generalizability of *Lens*

We investigate the effectiveness of *Lens* across various models, including representative (He et al., 2016; Liu et al., 2022), lightweight (Tan & Le, 2019), and foundation (Cherti et al., 2023; Oquab et al., 2023) models. Furthermore, we examine whether *Lens* can be constructively integrated with domain generalization (DG) techniques. Detailed model setups are in the Appendix E.3.1.

Table 1 summarizes the results. While Oracle-F selects the best fixed parameter to maximize average accuracy, it still suffers performance drops in many cases, revealing the limitations of using fixed parameters – *no single parameter optimally supports all scenarios*. In contrast, Oracle-S consistently

Table 1: Accuracy comparison among the baselines and *Lens* with various models.

Model	Num. Params	Pretraining Dataset	DG method	IN	ImageNet-ES (Baek et al., 2024) Oracle S	Oracle F	Naive control AE	Naive control Random	*Lens* (Ours)	*ImageNet-ES Diverse* (new) Oracle S	Oracle F	Naive control AE	Naive control Random	*Lens* (Ours)
ResNet-50 (He et al., 2016)	26M	IN-1K	-	86.3	92.1	49.2	32.2	50.2	**78.3**	63.5	38.5	17.7	12.1	**43.4**
		IN-21K	DeepAugment* +AugMix†	85.2	92.9	66.9	53.3	61.4	**83.2**	80.5	64.4	36.5	23.8	**65.5**
ResNet-152 (He et al., 2016)	60M	IN-1K	-	87.6	93.6	81.9	41.1	54.3	**81.0**	69.0	44.8	22.2	14.2	**49.2**
EfficientNet-B0 (Tan & Le, 2019)	5M	IN-1K	-	88.1	94.0	83.7	51.4	58.1	**80.9**	66.8	43.4	22.0	14.0	**46.2**
EfficientNet-B3 (Tan & Le, 2019)	12M	IN-1K	-	88.3	94.8	86.6	62.0	66.2	**83.2**	76.0	57.5	33.9	21.4	**55.8**
SwinV2-T (Liu et al., 2022)	28M	IN-1K	-	90.7	95.0	86.5	54.3	63.1	**82.4**	71.9	50.9	26.6	17.0	**50.8**
SwinV2-S (Liu et al., 2022)	50M	IN-1K	-	91.7	95.4	87.8	60.0	65.5	**84.7**	54.1	54.1	31.0	19.0	**55.8**
SwinV2-B (Liu et al., 2022)	88M	IN-1K	-	92.0	95.4	88.3	60.2	65.6	**85.7**	74.5	54.0	31.0	18.6	**55.4**
OpenCLIP-b (Cherti et al., 2023)	87M	LAION-2B	Text-guided pretrain	94.3	97.5	92.4	66.3	71.0	**90.7**	83.0	66.5	38.8	24.5	**67.6**
OpenCLIP-h (Cherti et al., 2023)	632M	LAION-2B		94.7	98.4	94.3	79.1	77.6	**93.0**	88.1	74.6	45.5	29.4	**74.5**
DINOv2-b (Oquab et al., 2023)	90M	LVD-142M	Dataset curation	93.6	97.5	85.2	74.5	73.9	**90.6**	87.8	72.4	44.7	28.3	**72.8**
DINOv2-g (Oquab et al., 2023)	1.1B	LVD-142M		94.7	98.0	90.7	84.3	79.6	**92.9**	92.9	82.5	62.7	35.3	**82.8**
All models				90.6	95.4	82.8	59.9	65.6	**85.5**	77.4	58.6	34.4	21.5	**60.0**

∗: (Hendrycks et al., 2021), †: (Hendrycks et al., 2019), IN: ImageNet (Le & Yang, 2015), S: Specific, F: Fixed, AE: Auto exposure, Random: Random Selection

outperforms Oracle-F by large margins and even matches or exceeds performance on ImageNet (IN), the training domain. This highlights the potential of scene- and model-specific sensor control. More importantly, *Lens* consistently boosts model performance compared to AE and Random across both benchmarks and all models, by large margins ranging from **8.52% to 47.46%**. *Lens* also delivers significantly better worst-case performance than Oracle-F, with gains of 29.1% in ImageNet-ES and 4.9% in *ImageNet-ES Diverse*, demonstrating the robustness of adaptive sensor control. These results show the importance of targeting sensor control to the model, rather than human perception, and demonstrate that *Lens* effectively *unlocks* the potential of model-specific adaptive sensor control.

Moreover, *Lens*, without requiring additional pretraining or extra data collection, outperforms the baseline methods even when they are combined with complex DG techniques like DeepAugment (Hendrycks et al., 2021) and AugMix (Hendrycks et al., 2019)), and applied to significantly larger models. For instance, *Lens* on ResNet-50 (He et al., 2016) achieves superior performance compared to baseline controls on DG-applied ResNet-50, and even outperforms those on the larger ResNet-152 (He et al., 2016), with gains ranging from **6.87% to 51.31%**. Furthermore, *Lens* on EfficientNet-B3 (Tan & Le, 2019), with only 12M parameters, surpasses the DG-enhanced OpenCLIP-h (Cherti et al., 2023), a model with 632M parameters, delivering **4.2%** higher accuracy; *Lens* can compensate for a **50× model size difference** through real-time sensor control. Lastly, when combined with DG techniques and larger models, *Lens*'s performance improves further, highlighting its synergistic nature. These findings emphasize the importance of optimizing data acquisition process, rather than focusing solely on model improvements.

5.2 Real-Time Adaptation Performance

To assess the real-time adaptability of *Lens*, we compare its performance with lightweight Test-Time Adaptation (TTA) methods, which are designed for real-time model adaptation. Additionally, we analyze the adaptation cost of *Lens*, focusing on the image capturing overhead associated with selected sensor parameter candidates, demonstrating its efficiency in real-time scenarios.

TTA Baselines and Target Models. We establish three representative TTA baselines: **BN1** (Prediction-time batch normalization, Nado et al. (2020)), **BN2** (Batch Normalization Adaption, Schneider et al. (2020)), and **TENT** (Wang et al., 2021). These methods are applied to the batch normalization layer and offer minimal computational and memory overhead. We apply these TTA baselines to two lightweight models – **ResNet-18** (He et al., 2016) and **EfficientNet-B0** (Tan & Le, 2019) – using data acquired via the traditional auto-exposure (AE) method. We then compare their performance against the same models when used with data acquired through *Lens*. Detailed explanations of each TTA method and the deployed models are provided in the Appendix E.4.

Table 2: Real-time adaptation performance analysis of *Lens* against TTA methods.

Model	TIN	Environments	Oracle		Naive control		Test-Time Adaptation			***Lens* (Ours)**			
			S	F	AE	Random	BN1	BN2	TENT	**Full (k=27)**	**CSA1 (k=6)**	**CSA2 (k=6)**	**CSA3 (k=18)**
ResNet-18 (He et al., 2016)	80.4	ImageNet-ES (Baek et al., 2024)	87.9	54.1	39.2	46.6	30.7	34.2	32.0	**73.8** (2.4sec)	**73.4** (0.5sec)	**72.6** (0.5sec)	**73.7** (0.2sec)
		ImageNet-ES Diverse	52.6	32.5	13.1	9.2	15.8	20.0	16.0	**34.6** (2.4sec)	**26.9** (0.5sec)	**27.4** (0.5sec)	**25.1** (0.2sec)
EfficientNet-B0 (Tan & Le, 2019)	84.9	ImageNet-ES (Baek et al., 2024)	92.3	61.2	42.6	51.2	31.9	41.7	42.6	**77.8** (2.4sec)	**76.2** (0.5sec)	**77.4** (0.5sec)	**78.6** (0.2sec)
		ImageNet-ES Diverse	60.5	38.5	19.9	11.7	15.0	17.5	15.6	**39.6** (2.4sec)	**32.4** (0.5sec)	**32.9** (0.5sec)	**31.4** (0.2sec)

TIN: Tiny-ImageNet (Le & Yang, 2015), AE: Auto exposure, Random: Random Selection, S: Specific, F: Fixed
BN1: (Nado et al., 2020), BN2: (Schneider et al., 2020), TENT: (Wang et al., 2021), CSA1: Random Selection, CSA2: Grid Random Selection, CSA3: Cost-Based Selection

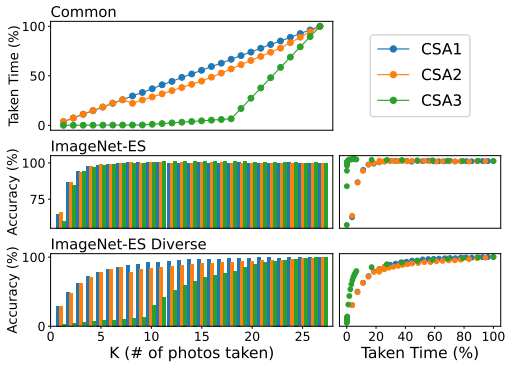


Figure 6: Cost analysis of CSAs (EfficientNet-B0) on ImageNet-ES & *ImageNet-ES Diverse*.

Table 3: Ablations on the Quality Estimator

Models			C1	C2	V1	V2
Tiny ImageNet			80.4	84.9	93.7	89.3
ImageNet ES	Oracle	S	87.9	92.3	96.3	95.0
		F	54.1	61.2	81.5	74.8
	Naive control	AE	39.2	51.2	70.4	62.0
		Random	46.6	42.6	71.4	66.5
	***Lens* with OOD techniques**	ViM	53.4	60.6	81.3	74.8
		ReAct	53.2	60.1	81.5	74.0
		ASH	47.2	55.9	61.2	3.2
		KNN	53.2	60.6	81.5	74.8
	***Lens with VisiT* (ours)**		**73.8**	**77.8**	**89.7**	**85.6**

Tiny ImageNet (Le & Yang, 2015), ImageNet-ES (Baek et al., 2024)
S: Specific, F: Fixed
AE: Auto exposure, Random: Random Selection
C1: ResNet18 (He et al., 2016), C2: EfficientNet-B0 (Tan & Le, 2019)
V1: Swin-B (Liu et al., 2021), V2: DeiT (Touvron et al., 2022)

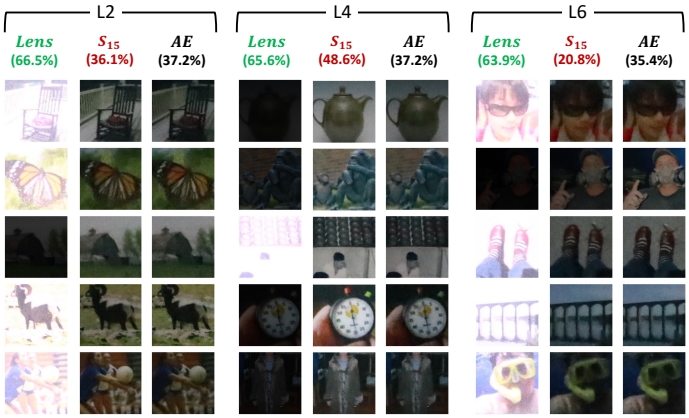
Candidate Selection Algorithms (CSAs) for *Lens*. Capturing images for all available parameter options for a scene introduces high latency, so we develop three candidate selection algorithms (CSAs) for *Lens* to enable lightweight, real-time operation. These CSA algorithms consider two key factors: **the number of image captures** (K) and **the overall capture time** per scene.

- **CSA1:** A simple method that randomly selects K options from the available options.
- **CSA2:** A grid-based random selection leveraging spatial locality. Observing that parameter settings closer in parameter space often yield similar image qualities, CSA2 divides the parameter space into grids and randomly selects K options from these grids. With 27 available options in our benchmarks (i.e., three options per each of the three parameters), the number of grids becomes 1^3 for $K = 1$–7, 2^3 for $K = 8$–26, and 3^3 for $K = 27$.
- **CSA3:** This method selects K options with the lowest capture costs, prioritizing settings with shorter shutter speeds, which are the primary contributors to capture latency. If multiple options share the same capture cost, the selection is made randomly.

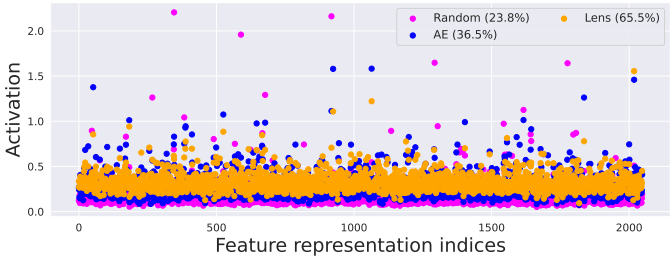
Results. Table 2 presents the results. *Lens* with full options ($K = 27$) significantly outperforms all TTA baselines across all models and both benchmarks, with gains ranging from **14.6% to 45.9%**. This underscores the superiority of sensor adaptation to model adaptation. Furthermore, the three CSAs for *Lens* drastically reduce capturing time by 91.7% (to only **0.2 seconds**) or require as few as **6 image captures** while maintaining accuracy. Figure 6 shows detailed interactions between capture time, K, and accuracy for EfficientNet-B0 (Tan & Le, 2019) across both benchmarks, using five random seeds. Note that the correlation between capture time and K is consistent across both benchmarks (marked as "common") because the CSAs are neither model- nor scene-specific, relying instead on non-deterministic selection at the time of capture. The results show that while each CSA has a different trade-off between K and taken time, all CSAs maintain high accuracy until taken time significantly decreases. These results verify *Lens*'s ability to balance accuracy and efficiency in real-time adaptation scenarios.

5.3 Ablation Study on the Quality Estimator

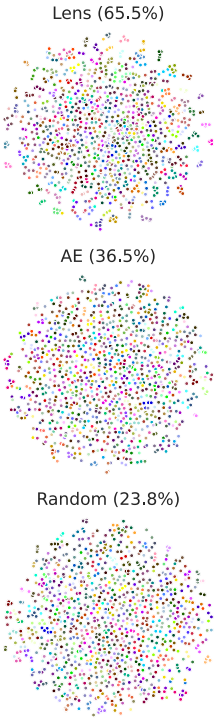
We investigate the effectiveness of *VisiT*, the proposed quality estimator for *Lens* that leverages confidence scores. For comparison, we replace *VisiT* with four state-of-the-art out-of-distribution (OOD) scoring methods, as introduced in Section 3.2. We evaluate these approaches on the ImageNet-



(a) Image quality comparisons based on human-level perception. (s_{15}: one of 27 sensor parameter options which can be taken by 'Random').



(b) Feature (model-level perception) activations on 'L2' samples.



(c) Feature embeddings.

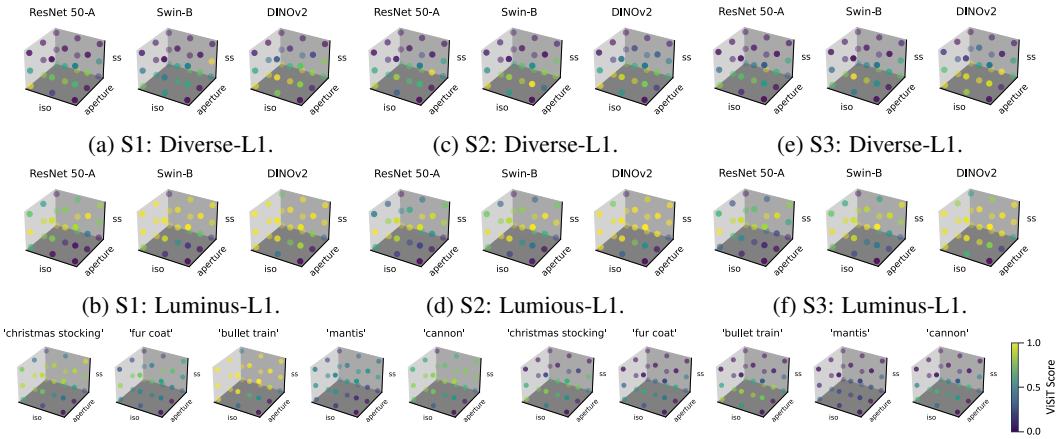
Figure 7: Sensing for human vs. sensing for DNN (ResNet-50 (He et al., 2016) augmented with AugMix (Hendrycks et al., 2019) and DeepAug (Hendrycks et al., 2021)) on *ImageNet-ES Diverse*.

ES dataset across four models: ResNet-18 (He et al., 2016), EfficientNet (Tan & Le, 2019), Swin-T (Liu et al., 2021), and DeiT (Touvron et al., 2022). As shown in Table 3, *Lens* integrated with *VisiT* consistently outperforms *Lens* paired with all OOD scoring baselines across every model, achieving an average gain of **20.7%**. This demonstrates that confidence scores are more reliable than OOD scores for evaluating the image quality from the model's perspective, which in turn enables *Lens* to identify optimal sensor parameters in the face of real-world perturbations.

5.4 Qualitative Analysis

Sensing for Human Vision vs. Model Vision. Figure 7a highlights the fundamental difference in how humans and neural networks perceive images, using examples from *ImageNet-ES Diverse*. While humans may struggle to discern details in dark or bright images (those selected by *Lens* in L2, L4, and L6), these images lead to better model accuracy (63.9-66.5%). In contrast, models perform poorly (20-48.6%) on images captured using auto-exposure (AE) settings or human-centered settings (S15 in L2, L4, and L6). Figure 7b further emphasizes this perceptual mismatch by showcasing distinct feature activation distributions for sample images under different sensor control methods. Specifically, the images provided by AE and Random settings cause the model to heavily activate certain features (those far from the average) that are treated as marginal for images acquired by *Lens*, which can degrade prediction performance. Moreover, Figure 7c demonstrates that, although *Lens*-acquired images may seem unintuitive from a human perspective, they enable the model to generate feature embeddings that are more clearly distinguishable between classes compared to those captured with AE and Random settings. These findings highlight the critical need to understand perception differences between humans and neural networks when designing effective sensor control strategies.

Solution Space Analysis on Camera Sensor Controls. Figure 8 illustrates the necessity of model- and scene-specific sensor control to effectively handle real-world perturbations. Each grid point represents one of the 27 parameter options from ImageNet-ES and *ImageNet-ES Diverse*, color-coded by the *VisiT* score of the image captured with that option. Each subfigure shows the results from three different models, illustrating that the same parameter setting for an identical sample can yield



(a) S1: Diverse-L1. (c) S2: Diverse-L1. (e) S3: Diverse-L1.

(b) S1: Luminus-L1. (d) S2: Lumious-L1. (f) S3: Luminus-L1.

(g) Class-wise solution space for ResNet50-A on Luminus. (h) Class-wise solution space for ResNet50-A on *Diverse*.

Figure 8: Solution spaces of parameter control in real perturbations. (ResNet 50-A: ResNet50 (He et al., 2016) + Augmix (Hendrycks et al., 2019) + DeepAugment (Hendrycks et al., 2021), Swin-B (Liu et al., 2022), and DINOv2 (Oquab et al., 2023))

significantly different quality scores when the model is changed. For example, an optimal parameter for Swin-B (Liu et al., 2022)) may perform poorly for DINOv2 (Oquab et al., 2023) or ResNet18 He et al. (2016), demonstrating the need for model-specific control.

The figure pairs (8a and 8b), (8c and 8d), and (8e and 8f) represent the same class sample captured under an identical lighting condition but with different object characteristics from "Diverse" scenes in *ImageNet-ES Diverse* and "Luminus" scenes in ImageNet-ES. The column-wise differences between the two datasets emphasize the importance of scene-specific control. With the same sample and L1 setting, fast shutter speeds yield low-quality images in "Diverse" scenes but high-quality images in "Luminus" scenes. Finally, Figures 8g and 8h show that under the same model, lighting conditions, and object characteristics, optimal sensor parameters can vary across different classes. Overall, sensor parameters must be dynamically adjusted based on both model and scene characteristics.

6 Conclusion

This paper presents *Lens*, the first method that introduces **model- and scene-specific camera sensor control** inspired by human visual perception; by capturing high-quality images from the model's perspective, *Lens* improves neural network performance. *Lens* employs *VisiT*, a lightweight, training-free, model-specific quality indicator based on model confidence, which operates on individual unlabeled samples at test time. Evaluations on two benchmarks of real perturbations, including our new dataset *ImageNet-ES Diverse* collected to address previously missing but notable perturbations, demonstrate that *Lens* with *VisiT* improves model accuracy by up to **51.31%**, outperforming representative test-time adaptation (TTA) baselines and domain generalization (DG) techniques based on naive control. Furthermore, *Lens* shows generalizability across various architectures and can be synergistically combined with all DG methods. By ensuring efficiency in adaptation costs while maintaining performance, *Lens* has the potential for real-time applications. Our qualitative analysis of sensor controls validates the importance of model- and scene-specific control. These findings underscore the significant impact of sensor control over domain generalization and test-time adaptation, offering a promising approach for enhancing AI systems' adaptability in real-world situations.

Limitations and Future Work. While *Lens* presents a novel paradigm of sensing for deep neural networks with significant potential for adoption in challenging scenarios across various tasks, such as autonomous driving, surveillance, and real-time 3D vision applications, it also opens avenues for further exploration. In this work, model confidence serves as a simple yet effective proxy for image quality assessment, but this can lead to overconfidence, especially in poorly calibrated models. Future work could investigate more robust quality estimators to mitigate overconfidence or explore synergies with TTA. Additionally, enhancing our Candidate Selection Algorithms (CSAs) by incorporating model- and scene-specific factors, along with optimizing resource scheduling through methods like reinforcement learning, would further strengthen the system's performance.

REPRODUCIBILITY STATEMENT

We include the source code in the supplementary material, along with instructions. Detailed information on the experiments, including datasets, scenarios, and hyperparameters, are in the Appendix.

REFERENCES

Eunsu Baek, Keondo Park, Jiyoon Kim, and Hyung-Sin Kim. Unexplored faces of robustness and out-of-distribution: Covariate shifts in environment and sensor domains. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. arxiv. *arXiv preprint arXiv:1706.05587*, 5, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Peng Cui, Yang Yue, Zhijie Deng, and Jun Zhu. Confidence-based reliable learning under dual noises. *Advances in Neural Information Processing Systems*, 35:35116–35129, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Geoffrey French, Michal Mackiewicz, and Mark Fisher. Self-ensembling for visual domain adaptation. *arXiv preprint arXiv:1706.05208*, 2017.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. Sotta: Robust test-time adaptation on noisy data streams. *Advances in Neural Information Processing Systems*, 36, 2024.

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*, pp. 770–778, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

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. *arXiv preprint arXiv:1912.02781*, 2019.

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*, pp. 8340–8349, 2021.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

Tetsuya Kuno, Hiroaki Sugiura, and Narihiro Matoba. A new automatic exposure system for digital still cameras. *IEEE Transactions on Consumer Electronics*, 44(1):192–199, 1998.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

JiaYi Liang, YaJie Qin, and ZhiLiang Hong. An auto-exposure algorithm for detecting high contrast lighting conditions. In *2007 7th International Conference on ASIC*, pp. 725–728. IEEE, 2007.

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*, pp. 740–755. Springer, 2014.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pp. 21–37. Springer, 2016.

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

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*, pp. 12009–12019, 2022.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

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

Ismoil Odinaev, Jing Wei Chin, Kin Ho Luo, Zhang Ke, Richard H.Y. SO, and Kwan Long Wong. Optimizing camera exposure control settings for remote vital sign measurements in low-light environments. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 6086–6093, 2023.

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. Realistic evaluation of deep semi-supervised learning algorithms. *Advances in neural information processing systems*, 31, 2018.

Emmanuel Onzon, Fahim Mannan, and Felix Heide. Neural auto-exposure for high-dynamic range object detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7706–7716, 2021.

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.

Sibendu Paul, Kunal Rao, Giuseppe Coviello, Murugan Sankaradas, Oliver Po, Y. Charlie Hu, and Srimat Chakradhar. Enhancing video analytics accuracy via real-time automated camera parameter tuning. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, pp. 291–304, 2023.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part III 14*, pp. 443–450. Springer, 2016.

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.

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

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4921–4930, 2022a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022b.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

Kaiyang Zhou, Chen Change Loy, and Ziwei Liu. Semi-supervised domain generalization with stochastic stylematch. *International Journal of Computer Vision*, 131(9):2377–2387, 2023.

Appendix

Adaptive Camera Sensor for Vision Models

A Further Discussion

In this section, we discuss future directions of this work, as outlined in Section 6.

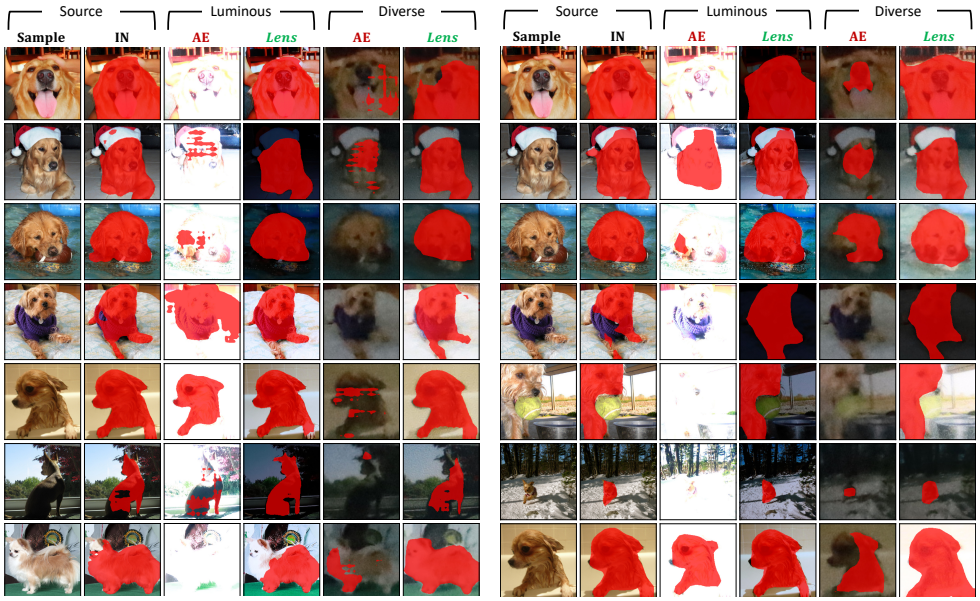
A.1 Towards More Realistic Scenarios

In this study, we utilized ImageNet-ES and *ImageNet-ES Diverse* as real-world perturbations, state-of-the-art Environmental and Sensor (ES) perturbation datasets. These datasets are pioneering in enabling effective evaluation of the impact of sensor control on environmental changes. By leveraging these resources, our work lays a robust foundation for addressing domain shift challenges in more complex and realistic scenarios through sensor control.

More Realistic Datasets. Extending *Lens* from classification tasks to advanced vision tasks such as semantic segmentation and object detection, and further into applications like autonomous driving or surveillance systems, presents a promising research direction. However, existing datasets lack both sensor control information and the labeled data necessary for these tasks. While ImageNet-ES and *ImageNet-ES Diverse* have facilitated the evaluation of *Lens* for classification, similar datasets tailored to other vision tasks are required. Therefore, the creation and implementation of sensor-controlled datasets for these advanced tasks are crucial for future research on *Lens*. Additionally, to encompass a broader range of realistic scenarios, we intend to collect and integrate more dynamic datasets, including multiple objects and dynamically changing scenes, as well as advanced tasks that incorporate ES perturbations similar to those in ImageNet-ES and *ImageNet-ES Diverse*. This will enable us to validate and enhance the robustness of our methodology against various domain shifts encountered in real-world applications, thereby providing a comprehensive evaluation of our method's resilience and effectiveness across diverse environments.

Potential to Adaptation on Advanced Vision Tasks. To showcase *Lens*'s versatility in various vision tasks and its value in collecting dataset containing sensor control factors, we performed a qualitative analysis focusing on two key applications: **Semantic Segmentation** and **Object Detection**. We compared *Lens* with AE (Auto Exposure), a baseline camera sensor control method described in Section 5. The evaluation involved two semantic segmentation models (FCN Long et al. (2015) and DeepLab v3 Chen et al. (2017)) and two object detection models (Faster R-CNN Ren et al. (2016) and SSDLite300 (Liu et al., 2016; Sandler et al., 2018; Howard et al., 2019)), representing standard or lightweight architectures. The analysis focused on the 'dog' class, a commonly used category in the training datasets of target models and a superclass in the evaluation datasets. Since these tasks generate multiple outputs, unlike the classification tasks for which *Lens* was initially designed, we adapted *Lens* by modifying the *VisiT* score for each specific task. Detailed experimental setups including the *VisiT* adaptations, are provided in Table 4. As shown in Figures 9 and 10, *Lens* achieves results that closely approximate, and sometimes outperform, those of the original sample (source domain) in most cases for both tasks and all targeted models. In contrast, AE failed to recognize the target class ('dog') in corresponding results. This suggests that *Lens* has significant potential for adaptation to other vision tasks using similar approaches. Since the performance of *Lens* varies depending on the customization of the *VisiT* Score for each target task, and further elaboration on this represents a promising avenue for future research.

System Considerations to Handle More Challenging Scenarios Rapid environmental shifts present significant challenges in real-world applications such as autonomous driving and surveillance systems. Successfully adapting sensing systems to time-constrained scenarios requires careful consideration of several factors, which can provide potential avenues for future research in this field. In these contexts, it is essential to account for limited available resources and ensure effective scheduling within specified timeframes. This involves balancing trade-offs between accuracy, the number of images captured, and system latency. Furthermore, the latency of each module—such as model inference and image capture—can vary depending on the deployed system architecture and must be meticulously managed to maintain overall system performance. Considering these factors, optimizing CSA algorithms emerges as a promising direction for *Lens*.



(a) FCN Long et al. (2015) (b) DeepLabV3 Chen et al. (2017)

Figure 9: Qualitative Analysis on both Benchmark (Semantic Segmentation)

Generalizability in Heterogeneous Camera Devices As outlined in the methodology section, our approach is inherently camera-agnostic. This is achieved through the Candidate Selection Algorithm (CSA), which selects camera parameter options independently of the specific range of supported camera settings. By tuning these parameters, the *Lens* can optimize image quality irrespective of the underlying camera hardware. Consequently, the strategy for selecting sensor options that yield the highest image quality remains effective even when the camera equipment varies. However, we acknowledge that the performance may differ based on the specific camera hardware and varying environmental conditions. To address this variability, future research could focus on customizing or adapting the CSA to better accommodate a wider array of camera devices and usage scenarios. This would further enhance the robustness and versatility of our system across diverse mobile platforms.

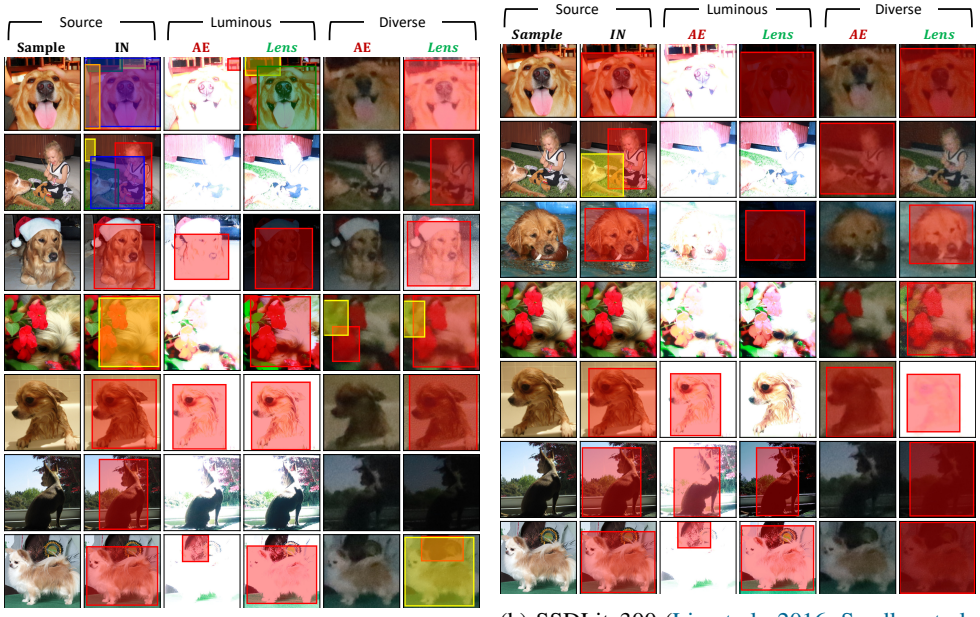
A.2 Potential of *Lens* on New Faces

Addressing Radical Distortion Problems In our current study, we did not evaluate radial distortion because it arises independently from light changes caused by environmental factors and sensor control. These factors posed critical issues in real domain shifts, but existing works related to robustness couldn't handle them effectively, making them the primary focus of our investigation. Despite not evaluating radial distortion directly, our methodology has the potential to address it by controlling framing parameters such as PTZ (pan, tilt, and zoom). Given two key points, 1) Adjusting pan, tilt, and zoom can minimize radial distortion effects. 2) Our policy algorithm selects the highest-quality images based on camera parameters. Therefore, incorporating framing parameters as control options is expected to effectively manage radial distortion. As a result, jointly applying sensor and framing control could enable the handling of a broader spectrum of domain shifts more effectively. Future

Table 4: Detailed Settings of Experiments on Other Vision Tasks

Tasks	**Semantic Segmentation**	**Object Detection**
Description	Identify and highlight pixels corresponding to 'dog'. (if the maximum confidence score indicates 'dog')	Detect objects and draw valid bounding boxes. (only for confidence scores >0.6)
***Lens* Adaptation (*VisiT* Score)**	Average of the confidence scores of the highlighted pixels.	Average of the confidence scores of the valid bounding boxes.
Models (backbone)	FCN Long et al. (2015) (ResNet50), DeepLab v3 Chen et al. (2017) (MobileNet v3)	Faster RCNN Ren et al. (2016) (ResNet50), SSDLite300 (MobileNet v3)
Datasets	[Training] COCO v1 Lin et al. (2014) (task), ImageNet-1k Deng et al. (2009) (backbone) [Evaluation] Luminus (ImageNet-ES Baek et al. (2024), Diverse (*ImageNet-ES Diverse*)	

ResNet50 He et al. (2016), MobileNet v3 Howard et al. (2019), SSDLite300 (Liu et al., 2016; Sandler et al., 2018; Howard et al., 2019)



(a) FasterRCNN Ren et al. (2016)

(b) SSDLite300 (Liu et al., 2016; Sandler et al., 2018; Howard et al., 2019)

Figure 10: Qualitative Analysis on both Benchmark (Object Detection)

research will explore integrating advanced PTZ control algorithms and real-time image quality assessments to further enhance our methodology's robustness against diverse domain shifts.

***Lens* for Representation Learning** Our method was specifically designed to capture high-quality images in scenarios that utilize model inference results and did not initially consider the high-quality image acquisition processes required for the training stages of representation learning, as suggested in the review. Given that most representation learning pipelines predominantly rely on fine-tuning pre-trained models for downstream tasks, we recognize the possibility of integrating *Lens* during the training stage. This integration could generate customized high-quality images tailored for both pre-trained models and target tasks, potentially reducing data collection costs and enhancing model performance.

A.3 Overconfidence Concern

In our ablation study, we compared our method with the OOD (Out-of-Distribution) score as an alternative proxy for assessing image quality and confirmed the effectiveness of our approach. However, we acknowledge that the confidence score may not be optimal in all scenarios, and the issue of overconfidence remains a concern, as mentioned in the Future Work section. This concern is evident in the performance gap between Oracle-S and *Lens*, highlighting the potential for improving *Lens* by addressing overconfidence.

Despite these challenges, *Lens* was intentionally designed for generalizability and simplicity, requiring low computational costs and no additional training. It has demonstrated significant potential for real-time applications. Moving forward, our future research will focus on mitigating the overconfidence problem by enhancing our current work and considering the adaptability of our approach to various real-world environments. This will further improve the reliability and performance of *Lens*.

B More Analysis

B.1 Label-wise Analysis

To validate the performance of *Lens* for individual labels in the source domain (ImageNet), we assessed the label-wise accuracy of the target models in Section 5.1 (Experiment 1) for both the

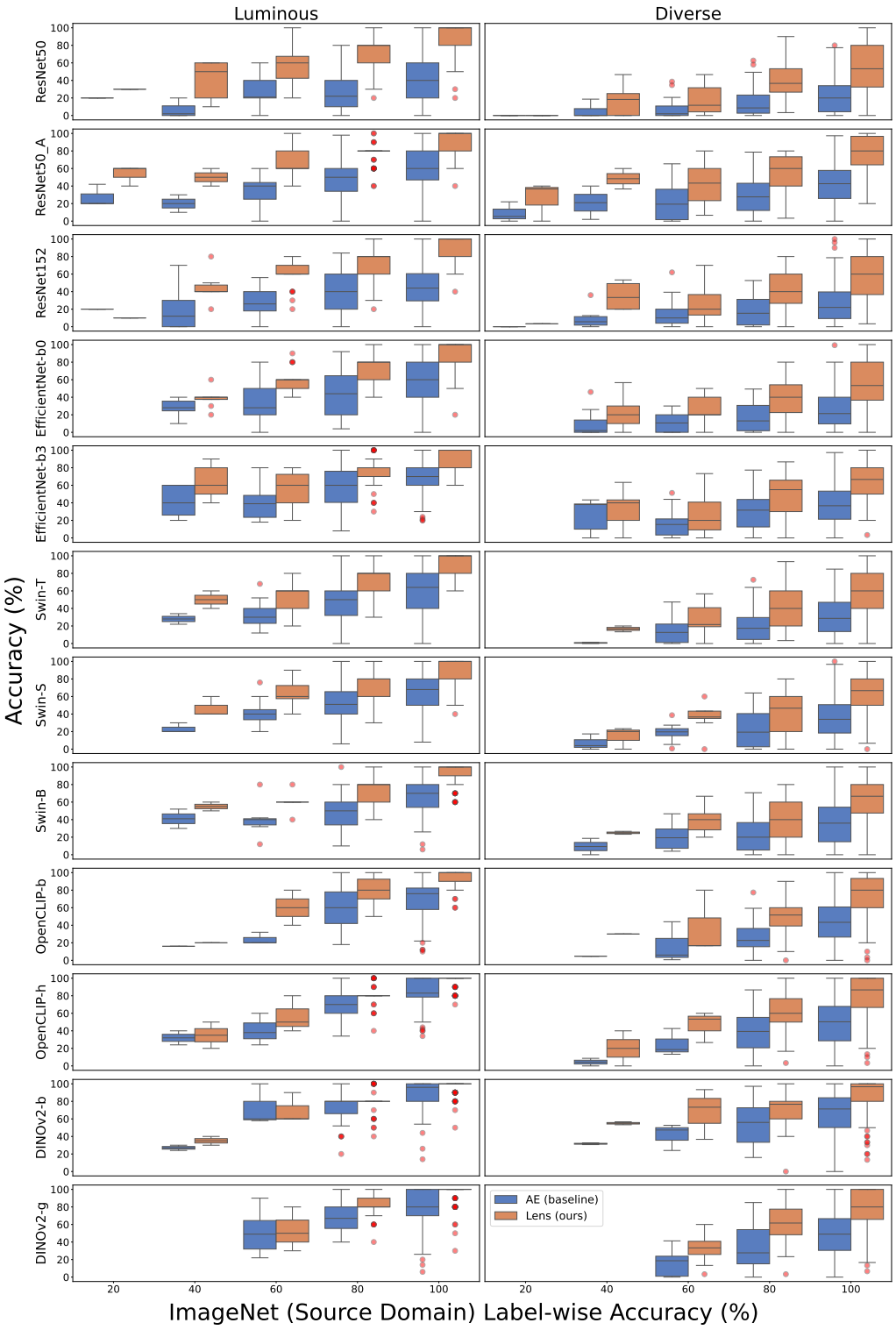


Figure 11: Generalizability of *Lens* Based on Label-wise Performance in the Source Domain.

representative baseline (AE: Auto Exposure setting) and *Lens*. As illustrated in Figure 11, *Lens* consistently outperforms the baseline, regardless of the performance of individual labels in the source domain.

Table 5: Ablation study on *VisiT*: Confidence (***C***) vs. Entropy (***E***) of Logits.

Model	Num. Params	Pretraining Dataset	DG method	IN	ImageNet-ES Luminous				*ImageNet-ES Diverse*			
					Naive control		***Lens***		Naive control		***Lens***	
					AE	Random	***C***	***E***	AE	Random	***C***	***E***
ResNet-50 (He et al., 2016)	26M	IN-1K	-	86.3	32.2	50.2	**78.3**	**78.5**	17.7	12.1	**43.4**	**43.2**
		IN-21K	DeepAugment* +AugMix†	85.2	53.3	61.4	**83.2**	**83.5**	36.5	23.8	**65.5**	**65.5**
ResNet-152 (He et al., 2016)	60M	IN-1K	-	87.6	41.1	54.3	**81.0**	**81.4**	22.2	14.2	**49.2**	**49.5**
EfficientNet-B0 (Tan & Le, 2019)	5M	IN-1K	-	88.1	51.4	58.1	**80.9**	**80.4**	22.0	14.0	**46.2**	**46.8**
EfficientNet-B3 (Tan & Le, 2019)	12M	IN-1K	-	88.3	62.0	66.2	**83.2**	**82.8**	33.9	21.4	**55.8**	**55.8**
SwinV2-T (Liu et al., 2022)	28M	IN-1K	-	90.7	54.3	63.1	**82.4**	**82.2**	26.6	17.0	**50.8**	**50.6**
SwinV2-S (Liu et al., 2022)	50M	IN-1K	-	91.7	60.0	65.5	**84.7**	**84.9**	31.0	19.0	**55.8**	**55.7**
SwinV2-B (Liu et al., 2022)	88M	IN-1K	-	92.0	60.2	65.6	**85.7**	**84.9**	31.0	18.6	**55.4**	**55.3**
OpenCLIP-b (Cherti et al., 2023)	87M	LAION-2B	Text-guided pretrain	94.3	66.3	71.0	**90.7**	**90.2**	38.8	24.5	**67.6**	**67.1**
OpenCLIP-h (Cherti et al., 2023)	632M	LAION-2B		94.7	79.1	77.6	**93.0**	**92.8**	88.1	29.4	**74.5**	**74.7**
DINOv2-b (Oquab et al., 2023)	90M	LVD-142M	Dataset curation	93.6	74.5	73.9	**90.6**	**91.0**	44.7	28.3	**72.8**	**72.8**
DINOv2-g (Oquab et al., 2023)	1.1B	LVD-142M		94.7	84.3	79.6	**92.9**	**93.1**	62.7	35.3	**82.8**	**83.3**
All models				90.6	59.9	65.6	**85.5**	**85.5**	34.4	21.5	**60.0**	**60.1**

Luminous Baek et al. (2024), ∗:(Hendrycks et al., 2021), †: (Hendrycks et al., 2019), IN: ImageNet (Le & Yang, 2015), AE: Auto exposure

While there are limitations to the improvements when the accuracy in the original sample is excessively low, in most cases, the accuracy enhancements approach those observed in the sampled data (ImageNet: source domain). This pattern is consistent across all datasets and models utilized in our experiments.

B.2 Ablation Study on the Quality Estimator: Confidence vs. Entropy of Logits

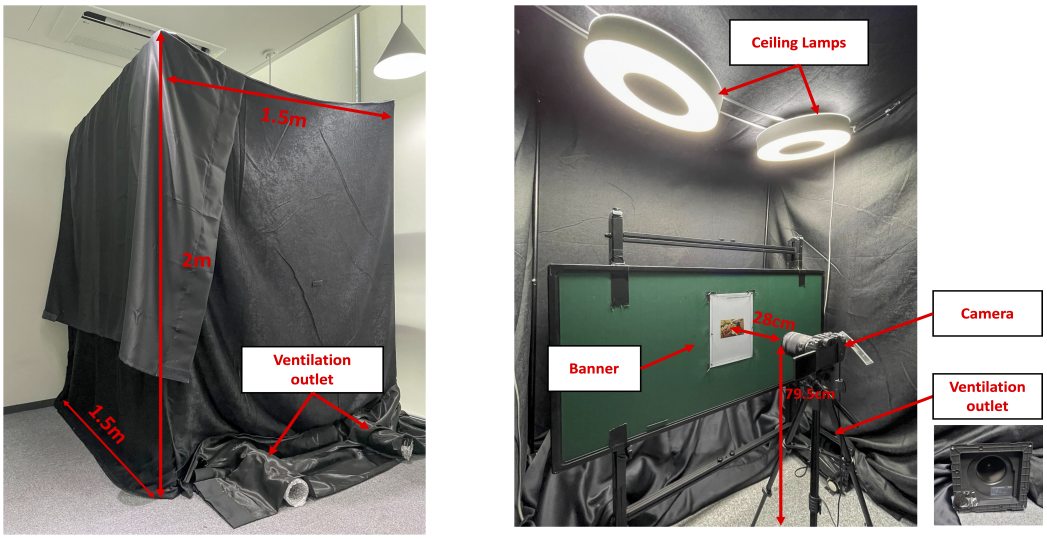
The confidence score and the entropy of logits are interchangeable approaches, as both metrics are based on logits. As shown in Table 5, replacing the *VisiT* score with the entropy of logits yields performance comparable to that of *VisiT* using the confidence score; however, it does not exceed this performance. Therefore, we opted to introduce confidence as a simpler and more representative metric for use in *VisiT* for *Lens*.

C Details on ImageNet-ES Diverse and ImageNet-ES Diverse Studio Implementations

This section provides details on how *ES-Studio Diverse* is built and *ImageNet-ES Diverse* is collected in *ES-Studio Diverse*. While the whole dataset will be open to the public after acceptance, the test set is provided for review at the following link: `https://shorturl.at/z3OoL`

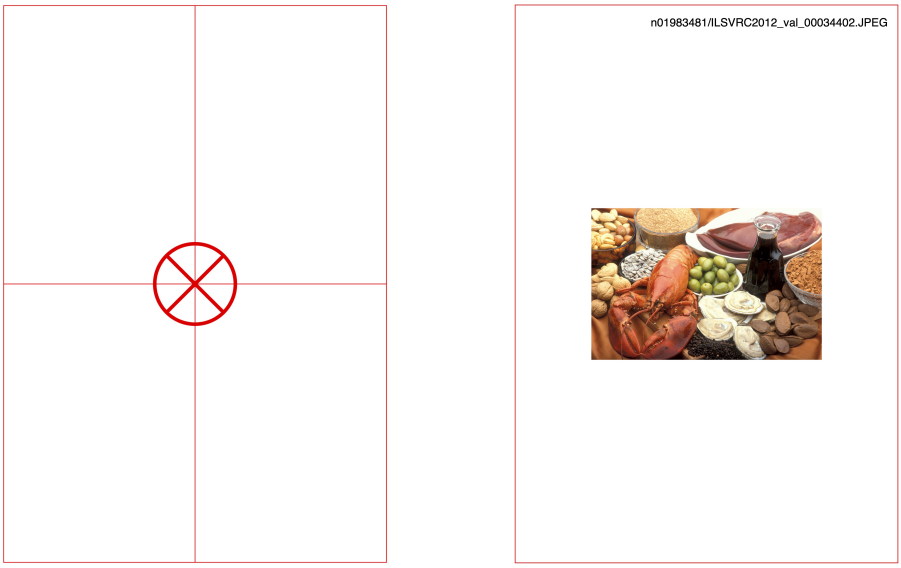
C.1 ES-Studio Diverse Setup

ES-Studio Diverse is established with the primary objective of ensuring the reproducibility of our proposed dataset while minimizing external factors, focusing specifically on light conditions and camera sensors. As illustrated in Figure 12, *ES-Studio Diverse* is designed as a completely dark room with dimensions of (1.5 m × 1.5 m × 2 m), equipped with four key components (blackboard, camera, ceiling lamps and desktop). In terms of the dark room setup, all sides are covered with blackout fabric to effectively block out any external light. Within the dark room, Banner (Component 1) is located in fixed position. To prevent light reflection from the desk, it is covered with blackout fabric that extends to the floor. To avoid any image distortion, careful attention is given to the height of the camera (Component 3), ensuring it is positioned at a distance of 28 cm from the banner (Component



(a) External. (b) Internal.

Figure 12: Actual appearance of *ES-Studio Diverse*.



(a) Alignment banner. (b) Sample banner.

Figure 13: Banner examples of *ImageNet-ES Diverse*.

1) in a straight line. To guarantee the consistent positioning of the banner, we use Figure 13a to adjust the camera angle. Light is controlled by two ceiling lamps (Component 2), strategically positioned at the midpoint between the magnetic blackboard and the camera lens. The entire setup aims to maintain consistency and accuracy in the captured images. Additionally, to address thermal issues and minimize errors and delays during data collection, ventilation outlets are installed. Finally, the detailed specification of each component is as follows:

- **Banner (Component 1):** The banner used in *ES-Studio Diverse* is A4-sized (210 mm x 297 mm) and is securely attached to a magnetic blackboard using six magnets. During dataset collection, as shown in Figure 13b, printed subset images from Tiny-ImageNet (Le & Yang, 2015) were placed precisely at the center of the banner while maintaining the original image's aspect ratio. Additionally, to prevent image quality degradation caused by the printing process, light reflection, humidity, and the properties of the paper material, we carefully tested multiple DPIs (72, 300, 600)

and various paper types. Ultimately, we selected a DPI of **300**, as higher DPIs result in smaller image sizes despite improving image quality. We also chose a **PVC banner**, which is resistant to light reflection, humidity, and creasing.

- **Ceiling Lamps (Component 2):** We have installed two 'Philips Hue White & Color Ambiance Infuse' lights, each with a maximum lumen1 output of 3700 lm. We choose this model for its ability to provide sufficient brightness even in dark room, allowing for an appropriate depiction of a light-on scenario. Additionally, these ceiling lamps offer the advantage of automating dataset collection through remote control APIs. To prevent the issue of light reflecting on the banner, the banner and the camera are positioned at a sufficient distance from the ceiling lights.
- **Camera (Component 3):** The camera selected for *ES-Studio Diverse* is 'Canon EOS-RP' body paired with 'RF 24-105mm F4-7.1 IS STM' lens. When combining this lens and body configuration, ISO can be implemented in the range of 100 to 40000, shutter speed from 1/4000 to 30 seconds, and aperture from f4.0 to f22. We opted for a full-frame CMOS sensor model rather than a crop one to achieve a broader field of view and higher resolution. We acknowledge that a change in the camera, even with the same parameter settings (both manual and AE), can lead to variations in the captured image. In other words, a change in the camera's hardware, even with identical software settings, can result in differences in the final output.
- **Desktop Computer (Component 4):** We automate the data collection system using the 'Apple Mac Studio M2 Max' desktop model, which communicates with the three aforementioned components via WIFI network. The desktop utilizes the Phillips Hue API for lighting control and the Canon camera control (CC) API for wireless camera control. The automation not only minimizes errors that could occur with human intervention, such as changes in camera position and external light interference, but also ensures consistency and accuracy, enabling faster and more efficient capturing and preprocessing processes.

This comprehensive configuration ensures a controlled environment within *ES-Studio Diverse*, limiting external influences to only light factors and camera sensors.

C.2 Data Collection Module Implementation

In this section, we revisit the key points discussed in Section 4 of the main text and subsequently delve into the finer details. In terms of reference dataset, a total of 1000 image samples were selected from validation set of Tiny-ImageNet (Le & Yang, 2015), a 200-class subset derived from ImageNet-1K.

During the data collection phase, each original reference image was taken with manual camera sensor parameter settings (M) and auto exposure (AE) settings. For the camera sensor parameter settings (M), we use 27 options provided in Table 6. For the auto exposure (AE) settings, the images were repeatedly captured five times, respectively. This procedure was reiterated for all light options provided in Figure 4 in Section 4.

While determining manual parameter options, we aimed to evenly cover the ranges of each camera sensor parameter (i.e. ISO, shutter speed, aperture). However, scenarios involving shutter speeds exceeding 1 second were excluded, considering their infrequent occurrence in real-world situations.

Additionally, the data collection process involved meticulous efforts to minimize distortion through precise camera angle adjustments and thorough attention to diverse camera settings. Regarding the focus, it has been set to AF (auto focus) mode, and the metering is set to evaluative metering mode, allowing the camera to assess the entire frame for metering before determining the exposure. We have set the recording resolution during shooting to the maximum supported by the camera, which is approximately 26 million (6240×4160) pixels.

C.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, the printed version of a reference image is fixed on the magnetic blackboard and captured by the camera. We extracted crucial information from the captured image, including the left top point, width and height of the reference image. Then, one of the captured images under auto exposure (AE) settings are selected to adjust the coordinates of the region of interest (ROI). For other images, we determined the valid area of each image by using

Table 6: Manual camera sensor parameter setting in test set.

Parameter No.	ISO	Shutter speed	Aperture
1	250	1/4’	f5.0
2	2000	1/4’	f5.0
3	16000	1/4’	f5.0
4	250	1/60’	f5.0
5	2000	1/60’	f5.0
6	16000	1/60’	f5.0
7	250	1/1000’	f5.0
8	2000	1/1000’	f5.0
9	16000	1/1000’	f5.0
10	250	1/4’	f9.0
11	2000	1/4’	f9.0
12	16000	1/4’	f9.0
13	250	1/60’	f9.0
14	2000	1/60’	f9.0
15	16000	1/60’	f9.0
16	250	1/1000’	f9.0
17	2000	1/1000’	f9.0
18	16000	1/1000’	f9.0
19	250	1/4’	f16
20	2000	1/4’	f16
21	16000	1/4’	f16
22	250	1/60’	f16
23	2000	1/60’	f16
24	16000	1/60’	f16
25	250	1/1000’	f16
26	2000	1/1000’	f16
27	16000	1/1000’	f16

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 Diverse*, 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 five individuals to ensure that all images are captured consistently by documenting identified issues such as 1) crop errors, 2) missed images, and 3) label mismatches. In cases where even one reviewer detected minor issues, the appropriate measures and re-evaluation were applied to the corresponding images. And if necessary, reshoot or reprocessing was carried out. This validation process ensured that the collected images align accurately with the original image.

D Details on Scene specific camera control concepts

This section provides details on Model-Scene specific camera control concepts. Figure 11 contains the overview of Scene-specific design of VisiT.

E Details on Experimental Setups

E.1 Another benchmark: ImageNet-ES

Table 7: Environment and Sensor specifics of ImageNet-ES (Baek et al., 2024).

Dataset	Original samples	Light	Camera sensor	ISO	Shutter speed	Aperture	Captured images
Test	1,000 (5 samples/class)	On/Off	Auto exposure (5 shots) Manual (27 options)	Auto 250/2000/16000	Auto (1/4’)/(1/60’)/(1/1000’)	Auto f5.0/f9.0/f16	10,000 54,000

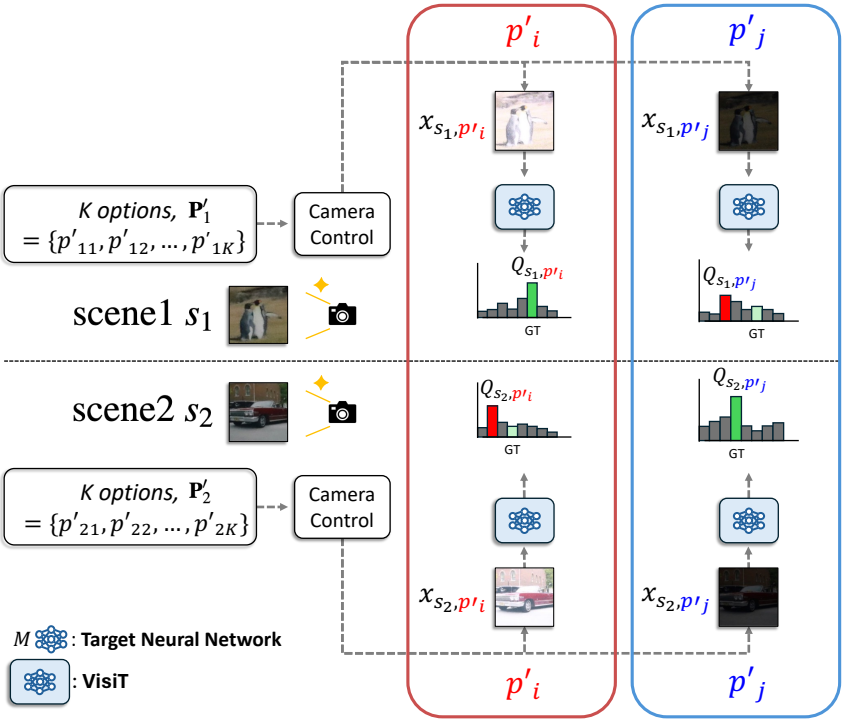


Figure 14: Scene-specific design of *VisiT*.

In this paper, two test datasets were used, both of which include extensive natural perturbations in environmental and sensor domains, incorporating 27 manual controls and 5 auto-exposure shots.

For **ImageNet-ES** (Baek et al., 2024), a subset of Tiny ImageNet (Le & Yang, 2015) was displayed on a monitor and captured using a camera. During the process, lighting conditions were varied by turning the lights on and off, and camera parameters were adjusted. Details of the environment and sensor specifications are provided in Table 7. Five images were randomly selected from each of the 200 classes in the Tiny ImageNet validation set (Le & Yang, 2015). To ensure visual fidelity, each sampled image had a resolution greater than 375 × 500 pixels, avoiding distortion when displayed on the screen. In total, 1,000 samples were collected for the test dataset.

E.2 Experiment Setups of Out-of-Distribution (OOD) Detection.

E.2.1 Models

We use ResNet-18 (He et al., 2016), EfficientNet (Tan & Le, 2019), Swin-T (Liu et al., 2021) and DeiT Touvron et al. (2022). All model weights used in the OOD detection experiments are sourced from the timm library. Since the pretrained model weights are designed to generate predictions for 1,000 classes (as in ImageNet), we fine-tune the classifier of each model to align with the 200 classes present in Tiny-ImageNet. To achieve this, non-resized images from the Tiny-ImageNet training set are used during the fine-tuning process, while the feature extractor of each model remains frozen. We provide additional results for DeiT (Touvron et al., 2022) in Figure 15.

E.2.2 OOD (Out-of-Distribution) techniques

Table 8: Datasets used in OOD detection experiments

Experiment Setting	Train		Validation			Test			
	ID	OOD	ID	OOD	C-OOD	Near S-OOD	Far S-OOD	C-OOD	ID
Semantics-centric	S3	OpenImage-O (train)	S1	Textures (test)	$val_{\text{ImageNet-ES}}$ (128 options)	n/a	n/a	n/a	n/a
MS-OOD	S3+	S3−	S1+	S1−		SSB-hard NINCO	iNaturalist Textures (test), OpenImage-O	$test_{\text{ImageNet-ES}}$ (54 options)	S2+

Table 9: Description of partitions of Tiny-ImageNet validation set (10K samples)

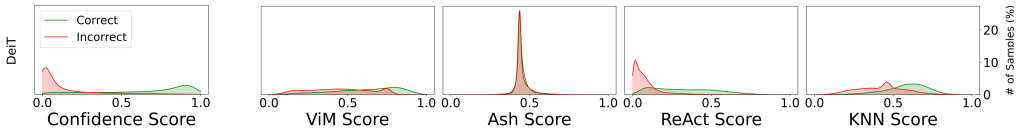
Partition	S1 Reference of $val_{\text{ImageNet-ES}}$	S2 Reference of $test_{\text{ImageNet-ES}}$	S3 $(val_{\text{Tiny-ImageNet}} \setminus (S1 \cup S2))$
# of samples	1,000	1,000	8,000

The datasets used for semantics-centric and Model-Specific OOD (MS-OOD) frameworks are outlined in Table 8. Other public datasets are used in their entirety, but we split the validation set of Tiny-ImageNet into three segments: S1, S2, and S3. We assign the same images to the validation and test splits of ImageNet-ES as S1 and S2, respectively. The remainder is designated as S3, which includes 40 images per class. Since the images in Tiny-ImageNet are provided in a resized version (64×64), corresponding images from ImageNet are used to preserve the original resolution. This partitioning scheme of Tiny-ImageNet is described in Table 9.

To train OOD detection methods within the semantics-centric framework, we use S3 and the training set of OpenImage-O as the ID and OOD datasets, respectively. Within the MS-OOD framework, S3+ and S3– are employed as the ID and OOD datasets, respectively. To validate the semantics-centric framework on ImageNet-ES, we use S1 and the test set of Textures as the ID and OOD datasets, respectively. Within the MS-OOD framework, we use S1+ and S1– as the ID and OOD datasets, respectively. For both frameworks, the validation set of ImageNet-ES is used as the C-OOD dataset.

We test five S-OOD datasets (SSB-hard, NINCO, iNaturalist, Textures, OpenImage-O) and categorize them into near-OOD and far-OOD categories following prior work. We use S2+ as the ID in the test set, assign samples from the test set of ImageNet-ES to labels following each framework's policy, and conduct the tests.

We validate OOD detection techniques on validation set of ImageNet-ES, including ViM (Wang et al., 2022a), ReAct (Sun et al., 2021), ASH (Djurisic et al., 2023) and MSP (Hendrycks & Gimpel, 2017). These methods demonstrate state-of-the-art performance and serve as baselines in recent OOD research. To validate current OOD detection methods, we leverage the results and APIs provided by OpenOOD (Yang et al., 2022). All implementations are based on the OpenOOD package.



(a) Confidence based. (b) OOD (Out-of-Distribution) score based.

Figure 15: Proxy for image quality assessment on DeiT: Each score is normalized between 0 to 1.

E.3 Experiment Setups of Generalizability

E.3.1 Target Models

The baseline model is ResNet-50 (He et al., 2016), trained using a standard training scheme on ImageNet (IN)-1K. To investigate whether well-configured sensor parameters could improve model performance to the level of domain-generalized (DG) models, we also evaluate ResNet-50 trained on IN-21K with DeepAugment (Hendrycks et al., 2021) and AugMix (Hendrycks et al., 2019). Additionally, we assess whether a larger model demonstrates increased robustness by evaluating ResNet-152. Furthermore, we use EfficientNet-B0/B3 (Tan & Le, 2019) to test the validity of lightweight model architectures on *ImageNet-ES Diverse*. SwinV2-T/S/B (Liu et al., 2022) are selected as representative models from transformer-based architectures, known for their robustness. OpenCLIP-b/h (Cherti et al., 2023) and DINOv2-b/g (Oquab et al., 2023) are included as domain-generalized versions of SwinV2.

E.4 Experiment setups of real-time adaptations

E.4.1 Test-Time Adaptation (TTA)

Deep learning has recently achieved significant improvements in performance. However, deploying models in real-world scenarios remains challenging, particularly when domain shifts occur, which are caused by variations between the training and testing environments. These domain shifts can significantly degrade model performance, as models are typically trained on static datasets that do not account for such variability. In order to tackle this issue without requiring access to the source domain, several Test-time adaptation (TTA) methods have been suggested.
Prediction-time Batch Normalization **(BN1)**, (Nado et al., 2020) aims to improve robustness of the model under covariate shifts, by updating Batch normalization statistics during the prediction-time. BN1 is simple and computationally efficient, because it doesn't require backward propagation. Batch normalization adaptation **(BN2)**, (Schneider et al., 2020) goes a step further by dynamically updating the Batch Normalization running statistics over the course of inference. Rather than using test batch statistics in isolation, BN2 updates the running mean and variance continuously based on test data as it is encountered, which makes the adaptation more flexible under varying test batches. Fully Test-time entropy minimization **(Tent)**, (Wang et al., 2021) enables models to adapt to domain shifts during test time by minimizing the output entropy of the predictions. Tent updates not only the BN statistics but also the entire model's parameters during test time, which provides greater robustness across a broader range of test samples but also increases computational costs.

E.4.2 Models

Of the models used in Section 3, we selected two representative lightweight architectures, ResNet-18 (He et al., 2016) and EfficientNet-B0 (Tan & Le, 2019), for this experiment. These models were chosen due to their compatibility with our three TTA methods, as they include Batch Normalization (BN) layers required for the application of these techniques.

E.5 Detailed results for experiments on real-time adaptations.

The results are shown in Tables 10, 11, and 12.

F Additional Representative Examples from *ImageNet-ES Diverse*.

More *ImageNet-ES Diverse* examples are provided in Figure 16.

Table 10: Comparison to test-time adaptation methods on CSA1(Random Based).

Model		ResNet-18 (He et al., 2016)		EfficientNet-B0 (Tan & Le, 2019)	
TIN		80.4		84.9	
Environments		ImageNet-ES (Baek et al., 2024)	*ImageNet-ES Diverse*	ImageNet-ES (Baek et al., 2024)	textitImageNet-ES *Diverse*
Oracle	Specific	87.9	52.6	92.3	60.5
	BF	60.6	23.9	67.7	30.5
Naive control	AE	39.2	13.1	42.6	19.9
	Random	46.6	9.2	51.2	11.7
TTA	BN1	30.7	15.8	31.9	15.0
	BN2	34.2	20.0	41.7	17.5
	TENT	32.0	16.0	42.6	15.6
Lens(Ours)	k=1	46.8	9.3	49.8	11.3
	k=2	61.8	15.3	68.4	19.7
	k=3	68.7	19.7	73.8	24.5
	k=4	70.2	22.7	77.0	28.4
	k=5	71.9	24.9	76.6	30.8
	k=6	73.4	26.9	76.2	32.4
	k=7	73.3	28.1	77.2	33.9
	k=8	73.1	29.8	78.3	34.9
	k=9	73.2	30.0	77.2	36.0
	k=10	73.6	31.0	77.9	36.2
	k=11	74.1	31.2	78.2	36.4
	k=12	73.6	31.6	77.2	37.5
	k=13	73.8	32.4	78.2	37.6
	k=14	74.0	32.5	77.8	38.4
	k=15	73.9	33.2	78.4	38.4
	k=16	73.9	33.1	77.6	38.6
	k=17	73.2	33.2	77.8	38.8
	k=18	74.2	33.3	77.5	39.2
	k=19	73.9	33.7	78.1	39.1
	k=20	73.9	33.7	78.1	39.2
	k=21	73.9	34.0	77.9	39.6
	k=22	73.8	34.4	77.8	39.2
	k=23	74.1	34.0	77.6	39.2
	k=24	73.7	34.5	77.8	39.2
	k=25	73.9	34.3	77.8	39.5
	k=26	73.8	34.4	77.8	39.8
	k=27	73.8	77.8	34.6	39.6

TIN: Tiny-ImageNet (Le & Yang, 2015), AE: Auto exposure, Naive: Random Selection BF: Best Fixed
BN1: (Nado et al., 2020), BN2: (Schneider et al., 2020), TENT: (Wang et al., 2021)

Table 11: Comparison to test-time adaptation methods on CSA2(Grid Random Based).

Model		ResNet-18 (He et al., 2016)		EfficientNet-B0 (Tan & Le, 2019)	
TIN		80.4		84.9	
Environments		ImageNet-ES (Baek et al., 2024)	*ImageNet-ES Diverse*	ImageNet-ES (Baek et al., 2024)	textitImageNet-ES *Diverse*
Oracle	Specific	87.9	52.6	92.3	60.5
	BF	60.6	23.9	67.7	30.5
Naive control	AE	39.2	13.1	42.6	19.9
	Random	46.6	9.2	51.2	11.7
TTA	BN1	30.7	15.8	31.9	15.0
	BN2	34.2	20.0	41.7	17.5
	TENT	32.0	16.0	42.6	15.6
Lens(Ours)	k=1	47.0	8.8	50.0	11.2
	k=2	61.8	15.3	67.2	19.4
	k=3	68.3	19.5	72.8	24.8
	k=4	70.9	23.2	76.1	28.7
	k=5	72.2	25.2	76.2	30.7
	k=6	72.6	27.4	77.4	32.9
	k=7	72.6	28.8	77.5	33.2
	k=8	72.9	25.9	78.3	31.1
	k=9	72.6	26.9	77.2	32.0
	k=10	73.1	28.0	78.1	33.1
	k=11	72.9	28.3	77.9	33.5
	k=12	73.2	28.9	77.4	34.3
	k=13	73.1	29.1	77.6	35.0
	k=14	73.6	29.7	77.7	35.3
	k=15	73.6	30.2	77.6	35.8
	k=16	73.4	30.6	77.5	36.1
	k=17	73.4	31.1	78.1	36.2
	k=18	73.6	31.3	77.9	36.5
	k=19	73.7	32.1	77.7	37.0
	k=20	73.9	33.7	78.1	39.2
	k=21	73.9	31.8	77.7	37.8
	k=22	73.5	32.2	77.6	38.0
	k=23	73.8	32.9	77.5	37.8
	k=24	73.9	33.0	77.6	38.7
	k=25	73.6	34.0	77.8	38.8
	k=26	73.7	34.1	77.8	39.2
	k=27	73.8	77.8	34.6	39.6

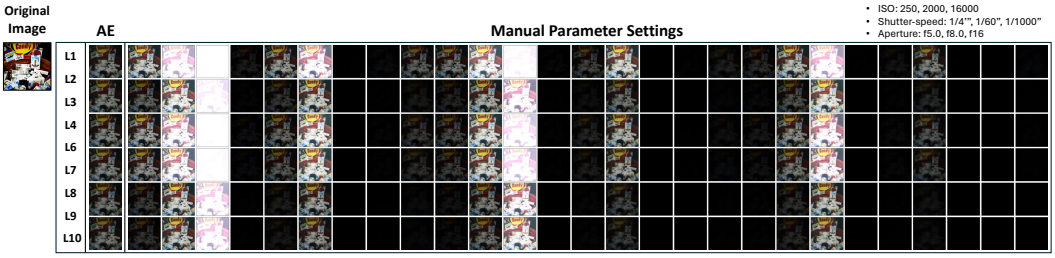
TIN: Tiny-ImageNet (Le & Yang, 2015), AE: Auto exposure, Naive: Random Selection BF: Best Fixed
BN1: (Nado et al., 2020), BN2: (Schneider et al., 2020), TENT: (Wang et al., 2021)

Table 12: Comparison to test-time adaptation methods on CSA2(Cost-Based).

Model		ResNet-18 (He et al., 2016)		EfficientNet-B0 (Tan & Le, 2019)	
TIN		80.4		84.9	
Environments		ImageNet-ES (Baek et al., 2024)	*ImageNet-ES Diverse*	ImageNet-ES (Baek et al., 2024)	textitImageNet-ES *Diverse*
Oracle	Specific	87.9	52.6	92.3	60.5
	BF	60.6	23.9	67.7	30.5
Naive control	AE	39.2	13.1	42.6	19.9
	Random	46.6	9.2	51.2	11.7
TTA	BN1	30.7	15.8	31.9	15.0
	BN2	34.2	20.0	41.7	17.5
	TENT	32.0	16.0	42.6	15.6
Lens(Ours)	k=1	42.5	0.6	47.3	1.1
	k=2	60.1	1.1	66.1	1.6
	k=3	67.2	1.4	72.7	2.2
	k=4	69.9	1.8	75.9	2.8
	k=5	71.0	1.9	76.4	3.3
	k=6	71.2	2.1	77.0	3.6
	k=7	72.4	2.5	77.5	4.0
	k=8	72.3	2.6	77.5	4.5
	k=9	72.6	2.9	77.6	4.9
	k=10	73.1	8.4	77.8	11.7
	k=11	73.1	12.4	78.6	16.7
	k=12	73.2	16.4	78.4	20.7
	k=13	73.2	17.7	78.6	23.5
	k=14	73.3	20.1	78.4	26.0
	k=15	73.4	22.1	78.8	27.6
	k=16	73.5	23.3	78.6	29.2
	k=17	73.8	24.2	78.6	30.3
	k=18	73.7	25.1	78.6	31.4
	k=19	73.9	27.9	78.4	33.7
	k=20	73.8	29.2	78.3	35.6
	k=21	73.7	30.6	78.5	36.5
	k=22	73.8	31.9	77.9	37.1
	k=23	73.6	33.1	78.1	38.1
	k=24	73.7	33.3	77.9	38.8
	k=25	73.8	34.0	77.8	39.0
	k=26	73.8	34.3	77.8	39.4
	k=27	73.8	77.8	34.6	39.6

TIN: Tiny-ImageNet (Le & Yang, 2015), AE: Auto exposure, Naive: Random Selection BF: Best Fixed
BN1: (Nado et al., 2020), BN2: (Schneider et al., 2020), TENT: (Wang et al., 2021)

(a)



(b)

(c)

Figure 16: *ImageNet-ES Diverse* samples (a), (b), (c) : In all subfigures, a broad range of variations can be observed based on each parameter option. Under the same camera parameter settings (Auto Exposure and Manual parameter settings), it is apparent that the captured image undergoes substantial changes depending on environmental variations (l1-l7) and camera parameter settings. This indicates that alterations in camera sensor and environmental settings can bring significant variations.