Towards RAW Object Detection in Diverse Conditions

Zhong-Yu Li[1] Xin Jin[1] Bo-Yuan Sun[1] Chun-Le Guo[1] Ming-Ming Cheng[1,2*]
[1]VCIP, CS, Nankai University [2]NKIARI, Shenzhen Futian
{lizhongyu,jinxin,boyuansun}@mail.nankai.edu.cn, {guochunle,cmm}@nankai.edu.cn

Abstract

Existing object detection methods often consider sRGB input, which was compressed from RAW data using ISP originally designed for visualization. However, such compression might lose crucial information for detection, especially under complex light and weather conditions. We introduce the AODRaw dataset, which offers 7,785 high-resolution real RAW images with 135,601 annotated instances spanning 62 categories, capturing a broad range of indoor and outdoor scenes under 9 distinct light and weather conditions. Based on AODRaw that supports RAW and sRGB object detection, we provide a comprehensive benchmark for evaluating current detection methods. We find that sRGB pre-training constrains the potential of RAW object detection due to the domain gap between sRGB and RAW, prompting us to directly pre-train on the RAW domain. However, it is harder for RAW pre-training to learn rich representations than sRGB pre-training. To assist RAW pre-training, we distill the knowledge from an off-the-shelf model pre-trained on the sRGB domain. As a result, we achieve substantial improvements under diverse and adverse conditions without relying on extra pre-processing modules. The code and dataset are available at https://github.com/lzyhha/AODRaw.

1. Introduction

Real-world object detection is a fundamental task in computation vision. Significant advancements have been made in this field with public datasets like COCO [27] and VOC [10]. However, these datasets have predominantly focused on sRGB images, which lose some critical information compared to RAW images. The sensor first captures original RAW images with a high bit depth in a typical camera. An image signal processor (ISP) then compresses these RAW images into 8-bit sRGB images. Unlike compressed sRGB images, RAW images retain a higher bit depth and thus preserve more distinguishable information [20, 23], which is crucial for computer vision tasks, particularly in challenging light and weather conditions. Moreover, in real-world applications, working directly with RAW data allows manufacturers to bypass image signal processing, leading to faster processing speeds and reduced computational overhead. Thus, RAW-based object detection has gained attention [40, 42–44] and shows its advantages in adverse conditions. However, exploration in this field remains limited.

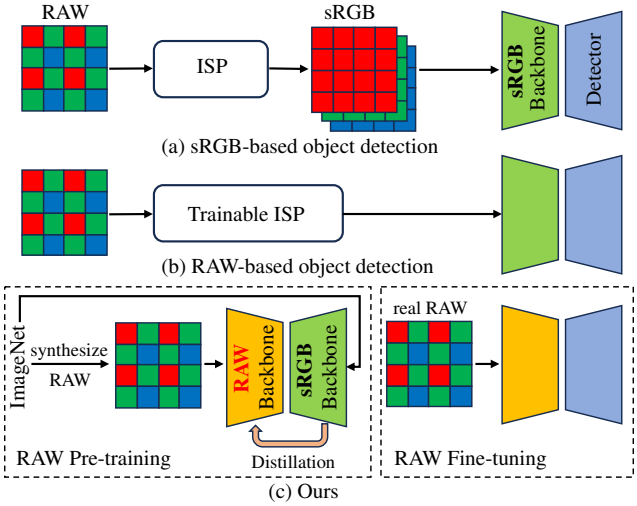


Figure 1. (a) Traditional sRGB-based object detection relies on 8-bit sRGB images, which are compressed from RAW images and lose detailed information. (b) Previous RAW-based methods utilize a trainable image signal processor (ISP) to adapt models pre-trained on the sRGB domain to the RAW domain. (c) We pre-train models on the RAW domain, achieving excellent performance on RAW object detection without requiring ISP modules.

The scarcity of relevant datasets is the key factor limiting the development of RAW-based object detection. However, collecting RAW images for object detection requires much more costs than sRGB images. For example, RAW images cannot be collected from picture websites like the sRGB-based dataset COCO [27]. Thus, taking pictures requires a lot of labor [21, 22], especially in rare weather conditions. Due to the limitations of collecting and annotating images, many RAW object detection methods [8, 32] rely on synthesizing RAW images that lack the authentic noise patterns

*Corresponding Author.

and dynamic range. The real RAW datasets [19, 33, 42] are also limited to the diversity. For example, LOD [19] and RAOD [42] datasets are annotated with only 8 and 6 categories, respectively. Moreover, some datasets focus on outdoor scenes of daylight and low-light while neglecting other adverse conditions. Thus, existing methods have limited applications and cannot fully utilize the advantages of RAW images in handling adverse conditions.

To overcome these limitations, we propose a challenging dataset (**AODRaw**) for **A**dverse condition **O**bject **D**etection with **RAW** images. AODRaw collect real RAW images from various indoor and outdoor scenes, with 2 light conditions, including daylight and low-light, and 3 weather conditions, including clear, rain, and fog. Because multiple light and weather conditions may co-occur, 9 distinct conditions are collected. Across different locations, cities, and scenes, we obtain 7,785 images and 135,601 annotated instances, with 6,504 images captured under adverse conditions. Meanwhile, our AODRaw is annotated with 62 categories, significantly exceeding existing datasets. The diversity of scenes and semantics can further facilitate the development of RAW-based object detection in the real world. Furthermore, we evaluate existing RAW object detection methods [8, 42] based on the AODRaw dataset.

With AODRaw, we aim to design a single model to detect objects across various conditions simultaneously, rather than training separate models for each condition in some previous approaches [8]. For RAW object detection, many methods usually transfer models pre-trained in the sRGB domain to the RAW domain using trainable adapters like neural ISP [8]. However, the domain gap between sRGB and RAW impedes models from understanding the intricate information in RAW images, while adapters also cost more. Some methods [42] train models from scratch, yet limited data availability constrains performance.

Differently, we explore pre-training on the RAW domain to reduce the domain gap between pre-training and fine-tuning, achieving notable improvements without any adapters. However, it is more difficult for models to learn high-quality representations from RAW images than sRGB images due to factors such as high dynamic range and camera noise. To mitigate this difficulty, we propose to distill representations from an off-the-shelf model pre-trained on the sRGB domain. Taking ConvNext-T [31] and Cascade RCNN [3], sRGB-based object detection achieves 34.0% AP on AODRaw. RAW object detection improves the performance to 34.8% AP through our RAW pre-training. To summarize, our main contributions are as follows:

- We propose AODRaw, a high-quality dataset for RAW object detection under various light and weather conditions. The dataset comprises diverse and complex images collected from various indoor and outdoor scenes.
- The AODRaw supports research across multiple tasks, including RAW object detection and sRGB object detection under adverse conditions. We evaluate the performance of existing object detection methods on these tasks.
- We pre-train models on RAW images via cross-domain distillation, achieving significant improvements without needing adapters such as neural ISPs.

2. Related Works

2.1. Object Detection

Mainstream object detection methods can be divided into two categories, *i.e.*, multi-stage and one-stage detectors. The multi-stage detectors, *e.g.*, R-CNN series [3, 35, 37], first generate region proposals and then refine them in subsequent stages. Cascade R-CNN [3] further extends this process via multiple refinement stages, progressively improving localization and classification accuracy. Although these methods achieve high accuracy, they have the drawback of slow inference. One-stage methods, such as YOLO [11, 29] and RetinaNet [28], directly predict object locations and categories, enabling faster inference but at a trade-off in precision. In addition, transformer-based approaches [46, 47] have emerged and leverage self-attention mechanisms to model spatial relationships in the image while requiring longer training times. Because these methods are primarily designed for sRGB images, we further evaluates these classic methods in the RAW domain.

2.2. RAW Object Detection

RAW object detection, which leverages unprocessed sensor data, has gained attention due to its potential and advantage in challenging light and weather conditions. However, the field lacks large-scale datasets for pre-training models on RAW images, which is essential for modern object detection methods. Thus, some methods [42] train detectors from scratch using real-time detectors [11]. Due to disadvantages such as camera noise and the limited quantity of RAW images, these methods may converge slowly and face limitations in performance. The other methods [40, 43, 44] adapt models pre-trained on sRGB images to the RAW domain. Among them, some propose a differentiable image signal processor (ISP) [8, 9, 15, 32, 34, 38] for pre-processing RAW images. Fine-tuning models by synthesizing RAW images from COCO [5, 26] also helps mitigate the domain gap. However, the sRGB pre-training still limits the ability to understand RAW images that contain more information than sRGB images, and the ISP modules add extra costs. Thus, we explore pre-training models on RAW images.

2.3. RAW Object Detection Datasets

Existing datasets for object detection, like COCO [27], primarily collect sRGB images. Although these datasets have driven significant progress in object detection, sRGB im-

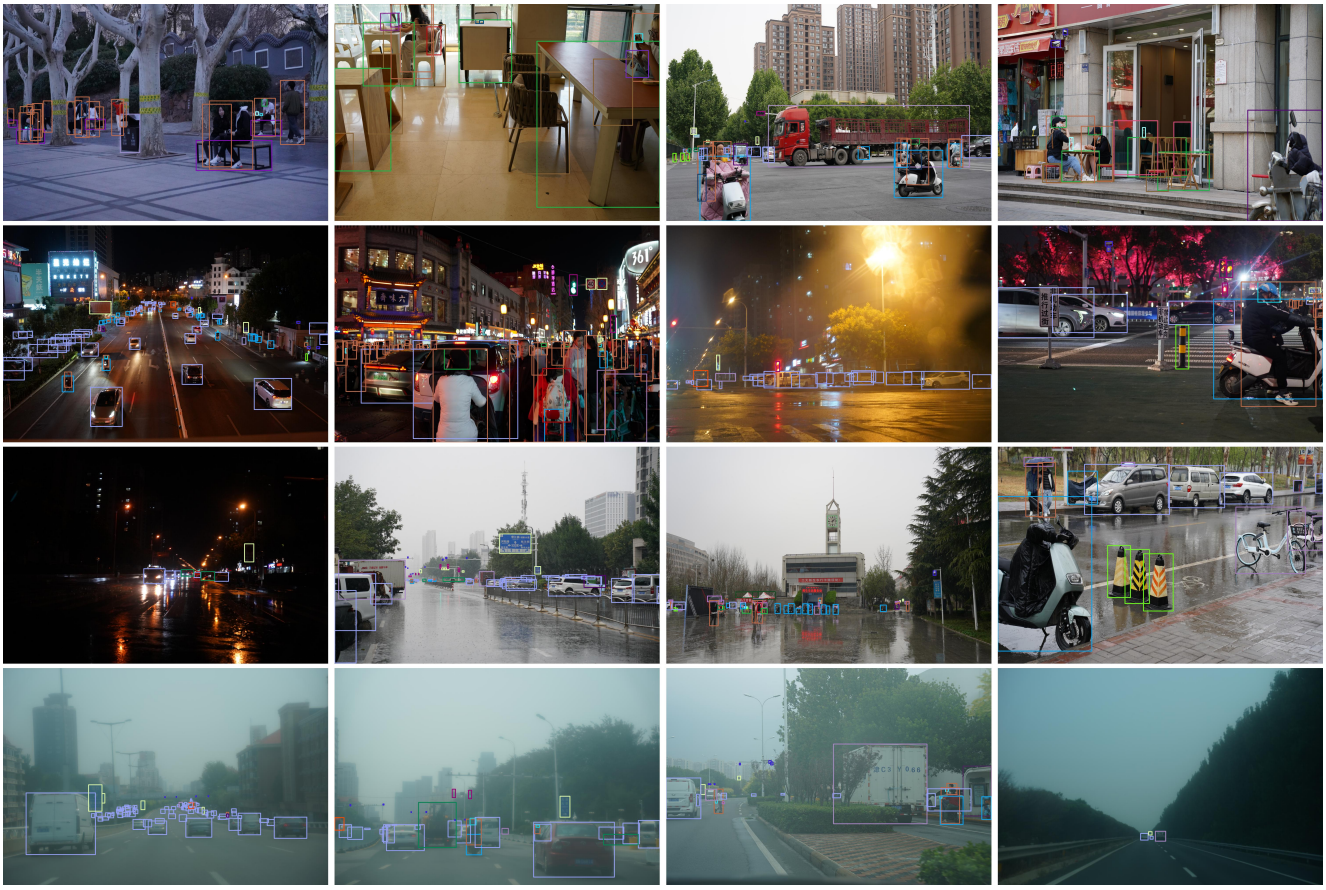


Figure 2. Example of the images in the AODRaw. From top to bottom, we show daylight, low-light, rain, and fog conditions, respectively. A part of the images are taken under multiple conditions. For example, the first one in the third row is taken in low-light and rain conditions. More examples for each condition can be found in the supplementary material.

ages lack the detailed information available in RAW images, which can be particularly beneficial in adverse conditions. Due to the scarcity of RAW datasets, many methods [8, 32] rely on synthetic datasets for RAW detection research. Then, a few RAW datasets have been proposed. For example, [45] collects images in low-light conditions, and [1] collects driving images of high dynamic range in fog conditions. The PASCALRAW [33] dataset is collected similarly to the PASCAL VOC [10], providing 4,259 images in daylight. The LOD [19] captures 2,230 paired images in daylight and low-light conditions. RAOD [42] has 25,207 annotated driving images but only covers 6 categories and 2 light conditions. Specific conditions, driving-focused scenes, or a narrow range of annotated categories limit these datasets.

3. AODRaw Dataset

3.1. Data Collection

Diverse conditions. With Sony A7M4, we construct a challenging dataset for RAW-based object detection under adverse and diverse conditions. Specifically, we consider 2 light conditions, *i.e.*, daylight and low-light, and 3 weather conditions, *i.e.*, clear, rain, and fog. For different light conditions, we capture both indoor and outdoor images. Because multiple conditions may co-occur, we finally collect 7,785 real RAW images and the corresponding sRGB images across 9 combined conditions, as shown in Tab. 2. For example, the image in the third row and first column of Fig. 2 is in rain and low-light conditions.

Data diversity. To make it as sufficient as possible for training and evaluation when collecting RAW images requires huge costs, we capture images across various locations, cities, scenes to ensure broad data diversity. Even if a few images are taken at the same location, they are still taken in different positions, directions, and perspectives. As shown in Fig. 2, some images are taken in traffic scenes, while others cover gardens, universities, libraries, streets, and other indoor scenes.

Data annotation. We follow the annotation format of the COCO dataset [27] to annotate bounding boxes in images across 62 categories commonly seen in daily life.

Dataset	Resolution	Images	Categories	Instances	Instances per image	Conditions
OnePlus [44]	4640×3480	141	5	1,228	8.7	1 (low-light)
PASCALRAW [33]	600×400	4,259	3	6,550	1.5	1 (daylight)
LOD [19]	1200×800	2,230	8	9,726	4.4	2 (daylight and low-light)
RAOD [42]	2880×1856	25,207	6	237,379	9.4	2 (daylight and low-light)
AODRaw (Ours)	6000×4000	7,785	62	135,601	17.4	9 (in Tab. 2)

Table 1. Comparison with existing RAW datasets.

Brightness	Indoor		Outdoor							Total
	daylight	low-light	daylight				low-light			
Weather	-	-	clear	fog	rain	fog+rain	clear	fog	rain	
Images	477	1,210	804	1,110	1,252	244	1,842	325	521	7,785
Instances	4,992	10,195	18,575	23,636	24,107	5,381	37,282	4,513	6,920	135,601

Table 2. The number of images per condition.

3.2. Data Analysis

In this section, we analyze the AODRaw dataset and compare it with two previous object detection datasets, *i.e.*, COCO [27] of sRGB object detection and RAOD [42] of RAW object detection. In the supplementary material, we show the detailed analysis about each condition.

Diverse scenes. As summarized in Tab. 2, AODRaw covers 9 conditions, including 2 light conditions, 3 weather conditions, and different combinations. Compared to existing datasets for RAW object detection, which mainly focus on outdoor scenes in daylight or low-light, as shown in Tab. 1, AODRaw has a greater diversity, presenting a more challenging task. Moreover, the scenes in AODRaw are varied and complex. As shown in Fig. 3a and Fig. 3b, images in AODRaw contain varying categories and instances, with up to 19 categories and 327 instances in an image. On average, there are 17.4 instances per image as shown in Tab. 1, exceeding existing datasets.

Increased category diversity. AODRaw includes 62 categories, a significantly higher number than most existing RAW object detection datasets, as shown in Tab. 1. Meanwhile, as shown in Fig. 3d, the distribution of categories exhibits a long-tail pattern, further increasing the challenge of RAW object detection in this dataset.

Object scales. The instances in the AODRaw vary widely in size, as shown in Fig. 3c, with a notably larger proportion of small objects than in previous datasets. This variance requires the detectors to extract multi-scale representations, increasing the complexity of the detection task.

Spatial distribution. The instances in the AODRaw dataset are more uniformly distributed in the images, as shown in Fig. 4. This uniform spatial distribution helps reduce spatial bias. Additionally, there is a slight bias towards the bottom of images, as most images are captured in outdoor scenes.

Light distribution. The distribution of lightness, calculated as the average gray value of sRGB images, is shown in Fig. 3e and Fig. 3f. The distribution reveals a broad range of light conditions in AODRaw.

3.3. Data Split

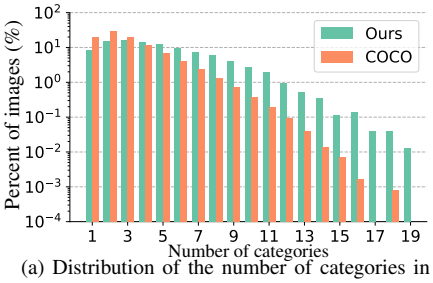
The dataset is randomly split for training and testing sets with a 7:3 ratio. To ensure that each split contains sufficient images of each condition, we split each condition individually and then merge the splitting results. As a result, we obtained 5,445 training images and 2,340 testing images, which contain 94,949 and 40,652 instances, respectively.

4. Benchmark

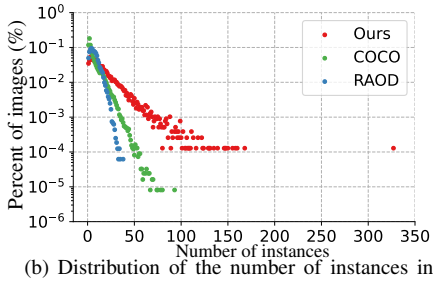
4.1. Implementation Details

Model training. We implement all object detection methods using a popular code base, mmdetection [4]. The models are trained for 48 epochs with a batch size of 16, except for Deformable DETR [47], which is trained for 100 epochs. Please refer to the supplementary material for more hyper-parameters. In addition, the RAW images are originally saved in the bayer pattern with the shape of $1 \times H \times W$, where H and W mean the height and width of images. To be compatible with existing models, we transform the RAW images into $3 \times H \times W$ using demosaicing following [42]. Following [26], the RAW images are further processed through gamma correction for faster convergence.

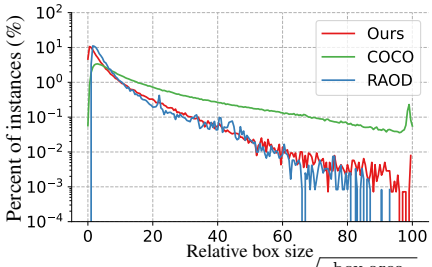
Image resolution. The images in the AODRaw dataset are recorded at a resolution of 6000×4000. It is unrealistic to feed such huge images into the detectors. Thus, we adopt two experiment settings: 1) down-sampling the images into a lower resolution of 2000×1333 following the approach in [42], and 2) slicing the images into a collection of 1280×1280 patches with a patch overlap of 300 and ignoring the objects whose IoU with the sliced images is lower than 0.4,



(a) Distribution of the number of categories in images.

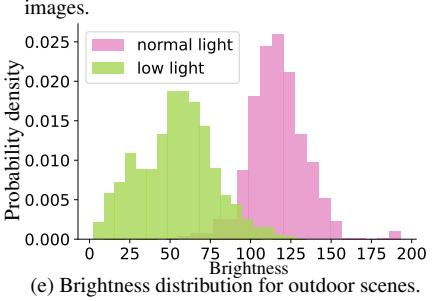


(b) Distribution of the number of instances in images.

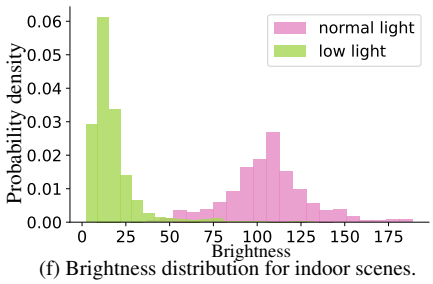


(c) Relative bounding box size $\sqrt{\frac{\text{box area}}{\text{image area}}}$.

(d) The number of instances per category.

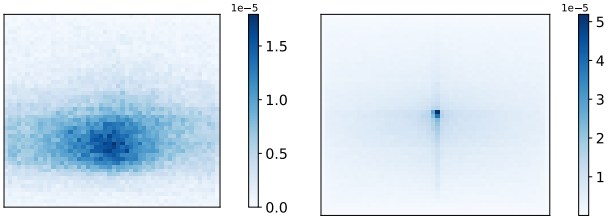


(e) Brightness distribution for outdoor scenes.



(f) Brightness distribution for indoor scenes.

Figure 3. Statistics indicate that our AODRaw dataset contains increased category and instance diversity.



(a) Our AODRaw. (b) COCO.

Figure 4. The distribution of object centers.

resulting in 71,782 images and 417,781 instances. The first setting supports faster training, but too tiny objects with an area of less than 32^2 are ignored because they will disappear after down-sampling. The second requires more time for training, but it can fully use high-quality annotations and support tiny object detection [7, 41]. In the following, we adopt down-sampling by default.

Evaluation protocol. We evaluate the models using the popular metric Average Precision (AP) [4, 27], along with AP_{75} and AP_{50} at the IoU threshold of 0.75 and 0.50. About AP_s, AP_m, and AP_l for small, medium, and large objects, we set object area ranges as $[0, 128^2)$, $[128^2, 320^2)$, and $[320^2, +\infty)$, respectively, when using the setting of down-sampling images. When slicing images, the ranges are set as $[0, 64^2)$, $[64^2, 160^2)$, and $[160^2, +\infty)$, respectively. To facilitate object detection in adverse conditions, we also report AP_{low}, AP_{rain}, and AP_{fog} for low-light, rain, and fog conditions, apart from AP_{normal} for the normal condition (the combination of daylight and clear weather).

4.2. Analysis

With AODRaw, we analyze the performances of various detectors for object detection with both sRGB and RAW images, as shown in Tab. 3. We evaluate some popular and milestone works, including multi-stage detectors (Faster RCNN [35], Sparse RCNN [37], and Cascade RCNN [3]), one-stage detectors (RetinaNet [28] and GFocal [25]), and transformer-based detectors (Deformable DETR [47]).

sRGB object detection in adverse conditions. Cascade RCNN achieves superior performance among the evaluated methods, with 25.6% AP and 27.3% AP_{normal}. However, the AP_{low}, AP_{rain}, and AP_{fog} are only 23.8%, 24.7%, and 20.4%, showing that the adverse conditions bring more challenges. More advanced backbones like ConvNeXt and Swin-T can improve performance. For example, ConvNeXt-T outperforms ResNet by 9.7% in AP_{normal}, but with lower improvements on adverse conditions, *i.e.*, 7.7% AP_{low}, 6.2% AP_{rain}, and 6.8% AP_{fog}. Such a gap shows the drawback of sRGB images in adverse conditions.

RAW object detection in adverse conditions. It is inappropriate to adopt models pre-trained on sRGB images when fine-tuning on RAW images. For example, RAW-based Cascade RCNN only achieves 33.7% AP when using sRGB pre-training, which is lower than 34.0% of the sRGB-based method. This phenomenon is partially caused by the domain gap between sRGB and RAW. As shown in Tab. 4, the detector trained on one domain will be significantly degraded when testing on another, showing that the RAW and sRGB domain models cannot generalize well to each other.

To overcome the domain gap, previous object detection methods in the RAW domain usually connect a neural image signal processor (ISP) with detectors, where the ISP projects the images from RAW to the sRGB domain. In Tab. 5, we evaluate two recently proposed methods designed for RAW object detection, RAOD [42] and RAW-Adapter [8]. Results show that neural ISP can stimulate the potential of RAW images. For example, RAOD achieves 34.4% AP and outperforms the 34.0% AP of the sRGB-

Method	Backbone	Pre-Train	Fine-Tune	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP_{normal}	AP_{low}	AP_{rain}	AP_{fog}
Faster RCNN [35]	ResNet-50 [16]	sRGB	sRGB	23.3	41.3	23.7	13.1	30.8	36.4	26.0	22.0	24.4	19.6
Retinanet [28]	ResNet-50 [16]			19.1	33.6	19.2	10.1	26.6	29.5	21.5	17.8	19.2	16.5
GFocal [25]	ResNet-50 [16]			24.2	40.3	24.7	13.3	31.9	37.0	26.5	22.3	24.1	21.1
Sparse RCNN [37]	ResNet-50 [16]			15.6	28.3	15.0	7.2	22.1	28.9	17.9	15.0	14.6	12.6
Deformable DETR [47]	ResNet-50 [16]			16.6	31.9	15.6	7.7	23.9	30.1	18.3	15.2	16.4	13.1
Cascade RCNN [3]	ResNet-50 [16]			25.6	41.4	26.4	13.7	32.4	38.3	27.3	23.8	24.7	20.4
Faster RCNN [35]	Swin-T [30]	sRGB	sRGB	28.4	50.1	28.8	15.6	35.9	42.6	32.0	26.0	27.2	23.3
Faster RCNN [35]	ConvNeXt-T [31]			29.7	51.7	30.1	17.1	37.3	45.4	33.1	28.3	27.1	24.4
GFocal [25]	Swin-T [30]			30.1	48.9	30.6	16.3	38.0	44.4	32.7	28.1	28.2	24.5
GFocal [25]	ConvNeXt-T [31]			32.1	49.9	33.6	18.7	39.9	49.5	35.2	30.3	31.8	26.0
Cascade RCNN [3]	Swin-T [30]			32.0	50.2	34.0	17.5	40.1	46.3	35.4	30.0	28.2	25.0
Cascade RCNN [3]	ConvNeXt-T [31]			34.0	52.7	36.3	19.3	40.8	52.1	37.0	31.5	32.9	27.2
Faster RCNN [35]	Swin-T [30]	sRGB	RAW	28.1	50.0	28.2	16.0	35.7	42.6	30.7	26.5	26.2	22.0
Faster RCNN [35]	ConvNeXt-T [31]			29.4	51.3	29.6	16.3	37.6	44.4	32.7	27.3	29.2	24.6
GFocal [25]	Swin-T [30]			29.9	48.2	30.6	16.3	38.3	45.0	33.1	27.6	29.0	23.8
GFocal [25]	ConvNeXt-T [31]			31.5	50.0	32.9	17.9	39.5	48.4	34.9	29.4	32.2	26.7
Cascade RCNN [3]	Swin-T [30]			31.7	49.8	32.8	17.7	39.7	47.8	35.3	29.8	28.6	23.9
Cascade RCNN [3]	ConvNeXt-T [31]			33.7	52.0	35.9	18.6	41.7	51.3	36.8	31.3	31.3	27.2
Faster RCNN [35]	Swin-T [30]	RAW	RAW	28.6	50.2	28.5	15.6	36.9	43.1	32.1	26.7	27.6	23.2
Faster RCNN [35]	ConvNeXt-T [31]			30.2	52.3	31.0	17.0	39.1	46.9	33.8	27.7	30.2	26.6
GFocal [25]	Swin-T [30]			30.7	49.7	31.8	17.2	39.4	47.4	33.7	28.6	28.5	25.3
GFocal [25]	ConvNeXt-T [31]			32.1	50.4	33.4	17.7	40.6	49.6	35.8	29.9	32.8	27.1
Cascade RCNN [3]	Swin-T [30]			32.2	50.5	33.8	17.9	40.5	49.7	35.5	30.0	29.5	25.1
Cascade RCNN [3]	ConvNeXt-T [31]			34.8	53.3	36.7	20.6	42.8	52.5	37.7	32.1	36.1	28.4

Table 3. Evaluation of object detection using RGB images, with different pre-training and fine-tuning settings.

Training	Evaluation	AP	AP_{50}	AP_{75}
sRGB	sRGB	34.0	52.7	36.3
	RAW	28.0	43.2	29.6
RAW	sRGB	21.2	33.1	22.5
	RAW	34.8	53.3	36.7

Table 4. The domain gap between sRGB and RAW.

based method. In particular, RAOD achieves greater improvements in adverse conditions than the normal condition (0.9% AP_{low}, 4.8% AP_{rain}, and 2.2% AP_{fog} vs 0.3% AP_{normal}). However, the neural ISP incurs extra computational costs and cannot fill the gap between RAW and sRGB domains, preventing the knowledge learned by the pre-trained model from being fully utilized.

5. RAW Pre-training

5.1. Method

Based on the above analysis, we aim to overcome the gap between sRGB pre-training and RAW fine-tuning by directly pre-training the models on RAW images, enabling us to achieve superior performances without requiring pre-processing modules like neural ISP.

Synthetic ImageNet-RAW. Visual pre-training [14, 16, 17] has made huge progress with the help of large-scale datasets like ImageNet-1K [36] that has over one million images. However, it is unrealistic to collect real RAW datasets of comparable size to large-scale sRGB datasets. Thus, we synthesize a 16-bit RAW dataset from the ImageNet-1K for RAW pre-training, using the unprocessing method [2]. We refer to the generated dataset as ImageNet-RAW. The unprocessed operation is inserted into the pipeline of data augmentations. Thus, we randomly adjust the average brightness and simulated noise in each iteration so that models can generalize well across different conditions.

RAW pre-training with cross-domain distillation. By replacing the sRGB input with synthetic RAW images, we can pre-train models using the classification targets provided by the original ImageNet-1K dataset and keep the hyper-parameters consistent with sRGB pre-training. However, pre-training in the RAW domain presents additional challenges compared to the sRGB domain due to factors such as the noise and high dynamic range (HDR) inherent in RAW images. Following prior works [5] that reveal that HDR has a negative impact on training, we also find that applying gamma correction during pre-training improves the Top-1 accuracy on ImageNet-RAW. As a result, we apply

Method	Backbone	Neural ISP	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP_{normal}	AP_{low}	AP_{rain}	AP_{fog}
Baseline	ConvNeXt-T [31]	✗	33.4	51.8	35.3	19.4	41.1	50.7	36.8	31.0	30.2	27.0
	ResNet-18 [16]	✗	18.1	30.9	18.3	8.2	24.8	33.3	20.6	16.7	17.3	14.6
Gamma correction [13]	ConvNeXt-T [31]	✗	33.7	52.0	35.9	18.6	41.7	51.3	36.8	31.3	31.3	27.2
RAOD [42]	ConvNeXt-T [31]	✓	34.4	52.9	35.9	19.5	42.9	52.2	37.1	31.9	35.0	29.2
RAW-Adapter [8]	ResNet-18 [16]	✓	19.9	33.2	20.1	9.8	27.3	34.4	22.3	18.1	20.8	16.9
Ours	ConvNeXt-T [31]	✗	34.8	53.3	36.7	20.6	42.8	52.5	37.7	32.1	36.1	28.4
	ResNet-18 [16]	✗	22.3	36.6	23.5	11.3	29.3	36.3	25.4	20.1	22.4	18.6

Table 5. Comparison with methods that adapt models pre-trained on sRGB domain to RAW images.

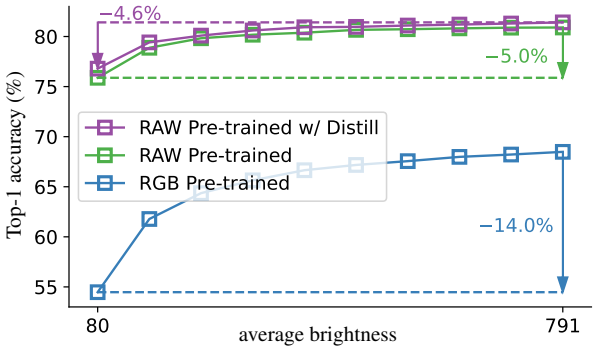


Figure 5. Top-1 accuracy on ImageNet-RAW when synthesizing RAW images under different average brightness. The maximum average brightness for an image is 2^{16}.

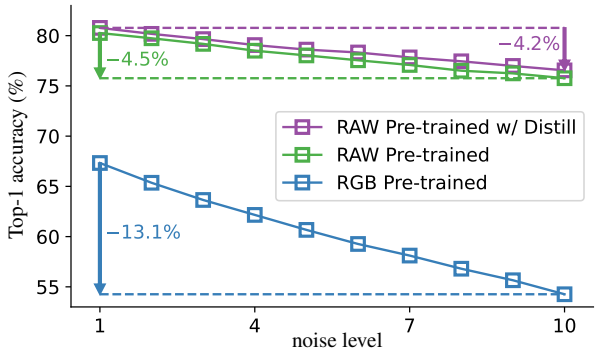


Figure 6. Top-1 accuracy on ImageNet-RAW when adding different noise levels to synthesized RAW images. Here, the noise level represents the standard deviation of the shot noises.

Distillation	AP	AP_{50}	AP_{75}
✗	34.1	52.4	35.9
Logit	34.3	52.4	36.6
Logit + Feature	34.8	53.3	36.7

Table 6. Ablation for the knowledge distillation when using Cascade RCNN and ConvNeXt-T.

gamma correction by default during pre-training. Regarding for the noise, when no noise is added during the synthesis of ImageNet-RAW, the model achieves a Top-1 accuracy of 74.8% after pre-training for 50 epochs. However, when noise is introduced, the accuracy drops to 74.4%, indicating that noise negatively impacts model performance. To alleviate this problem, we propose using the knowledge distillation [12, 18, 39]. As cross-domain distillation, we take an off-the-shelf model pre-trained on the sRGB domain as the teacher to assist the pre-training on the RAW domain. The student shares the same architecture as the teacher for a fair comparison. Specifically, a logit distillation with the Kullback-Leibler divergence loss and a feature distillation with the L1 loss are combined.

Enhanced robustness to adverse conditions. RAW pre-training and cross-domain distillation enhance the robustness to different conditions. For one sRGB image, when converted to RAW at different epochs of pre-training, its brightness is randomly adjusted, and random noise is added. While distillation provides consistent targets regardless of synthesized noise and brightness, the models are prompted to learn representations invariant to those conditions.

Fig. 5 and Fig. 6 verify the effects on robustness. By synthesizing the ImageNet-RAW validation set under different brightness and noise levels and evaluating the pre-trained models, we can observe that models pre-trained using distillation exhibit greater robustness to adverse conditions. For example, when reducing the brightness from 791 to 80, the performance degradation is 4.6% Top-1 accuracy when distillation, lower than 5.0% without distillation. Similarly, when increasing the noise level from 1 to 10, distillation exhibits a lower performance gap, *i.e.*, 4.2% vs 4.5%. In addition, we show that sRGB pre-training has a significantly higher performance degradation of 14.0% and 13.1% when adjusting the brightness and noise, respectively, showing that RAW pre-training effectively enhances the robustness.

Tab. 6 further shows the advantages of distillation on real-world RAW object detection. Logit-based distillation improves the AP by 0.2%, and feature-based distillation further extends the improvement by 0.5%.

5.2. Experiments Results.

Through RAW pre-training and distillation, ConvNeXt-T achieves 81.8% Top-1 accuracy on synthetic ImageNet-RAW. Although pre-training accuracy is still lower than sRGB pre-training (82.1%), the performance on real RAW object detection has significantly improved across various architectures, as shown in Tab. 3. For instance, RAW

Method	Backbone	Pre-Train	Fine-Tune	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP_{normal}	AP_{low}	AP_{rain}	AP_{fog}
Cascade RCNN	Swin-T	sRGB	sRGB	28.1	44.6	29.0	11.1	20.1	33.9	30.5	26.4	32.3	23.2
Cascade RCNN	ConvNeXt-T			29.9	46.5	31.0	12.7	24.0	35.5	33.1	28.0	33.0	27.8
Cascade RCNN	Swin-T	sRGB	RAW	29.2	46.2	30.2	10.9	19.8	35.1	31.0	27.8	32.3	24.6
Cascade RCNN	ConvNeXt-T			29.7	46.9	30.6	11.5	22.2	35.4	32.3	27.8	33.1	27.0
Cascade RCNN	Swin-T	RAW	RAW	29.8	47.0	30.9	11.4	21.7	35.4	31.4	28.1	32.9	27.3
Cascade RCNN	ConvNeXt-T			30.7	48.0	32.4	11.7	23.9	36.8	33.6	28.9	34.1	29.3

Table 7. Results of all-weather object detection using sliced RGB images.

Method	Params (M)	FPS	Epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l	AP_{normal}	AP_{low}	AP_{rain}	AP_{fog}
YOLOX-Tiny [11]	5.1	222.6	300	16.4	32.1	14.9	6.8	23.2	29.4	18.0	15.2	15.1	12.3
YOLOv6-n [24]	4.3	170.7	400	18.0	30.0	18.0	7.6	24.4	32.8	19.3	16.0	16.5	14.0
YOLOv8-n [24]	3.0	188.1	500	18.9	32.0	18.8	8.9	26.5	33.2	21.4	16.3	16.8	15.4
YOLOv8-n [24]†	3.1	57.6	500	19.7	32.8	19.9	9.4	27.0	32.9	21.8	16.9	18.9	14.9
YOLO-MS-XS [6]	4.5	113.0	300	24.7	40.0	25.1	12.1	33.4	41.4	28.2	22.4	21.2	19.7

Table 8. Evaluation of real-time object detection using down-sampled images. The models are trained and evaluated with an input size of 1280×1280. † means using the trainable pre-processing module proposed by [42]. Meanwhile, we measure the frames per second (FPS) of all models using an NVIDIA 3090 GPU.

pre-training improves Cascade RCNN and ConvNeXt-T by 1.1% AP than sRGB pre-training. In particular, the models achieve more significant improvements in adverse conditions, *i.e.*, 0.8% AP_{low}, 4.8% AP_{rain}, and 1.2% AP_{fog} vs 0.9% AP_{normal}. Compared to sRGB-based object detection, we improve by 0.8% AP. These results demonstrate the advantages of our RAW pre-training.

6. Experiments on Sliced AODRaw

The experiments in Section 4 and Section 5 use the down-sampling setting. Tab. 7 further lists the results of the slicing setting, where the hyper-parameters follow the down-sampling experiments, except that we fine-tune models for 12 epochs. The experiments with ConvNeXt show the same trends as those in Tab. 3. Adapting models pre-trained on the sRGB domain to RAW object detection degrades the AP by 0.2% compared to sRGB object detection, and RAW pre-training improves the performance by 0.8%. Differently, with the Swin transformer, RAW object detection outperforms sRGB object detection even when using sRGB pre-training. It may be because the sliced images contain more visual information, allowing the model to converge well compared to down-sampled images. Meanwhile, RAW pre-training further improves performance by 0.6% AP, especially in adverse conditions. Overall, these results further illustrate the effectiveness of RAW pre-training.

7. Experiments on Real-Time Object Detection

We also evaluate real-time object detection of YOLO-Series. As shown in Tab. 8, YOLO-MS-XS [6] and YOLO-v8-n [24] achieve 24.7% and 18.9% AP with a high FPS and parameters less than 5M. However, adverse conditions have lower performance, such as 16.3% AP_{low}, 16.8% AP_{rain}, and 15.4% AP_{low} of YOLO-v8-n. Some methods have tried to boost the performance via a trainable ISP. Here, we take the recently proposed method [42] as an example for analysis. In Tab. 8, integrating [42] with YOLO-v8-n improves the performance to 16.9% AP_{low} and 18.9% AP_{rain}. However, the FPS is significantly reduced, destroying the real-time property of the detector. In summary, the proposed AODRaw provides a new foundation to push real-time RAW object detection development.

8. Conclusion

This paper introduces AODRaw, a challenging dataset for RAW-based object detection across diverse and adverse conditions. Compared to traditional sRGB datasets, AODRaw offers diverse RAW images that retain essential visual information for object detection in complex light and weather conditions. Based on AODRaw, we evaluate existing methods of RAW object detection. Meanwhile, we utilize a cross-domain knowledge distillation to directly pre-train models on the RAW domain, solving the domain gap between sRGB pre-training and RAW fine-tuning. In this way, we improve the performance, particularly in adverse conditions, without relying on extra pre-processing modules. Our dataset is a benchmark for evaluating detection methods and a foundation for developing detection methods that generalize well across varied conditions. Our insights highlight the potential of RAW pre-training to advance real-world object detection and encourage further research on exploiting RAW images for challenging environments.

Acknowledgments. This research was supported by NSFC (62225604) Shenzhen Science and Technology Program (JCYJ20240813114237048). Computation is supported by the Supercomputing Center of Nankai University (NKSC).

References

[1] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *CVPR*, 2020. 3

[2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 6

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE TPAMI*, 2019. 2, 5, 6

[4] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 4, 5

[5] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Felix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2, 6

[6] Yuming Chen, Xinbin Yuan, Ruiqi Wu, Jiabao Wang, Qibin Hou, and Ming-Ming Cheng. Yolo-ms: Rethinking multi-scale representation learning for real-time object detection. *arXiv preprint arXiv:2308.05480*, 2023. 8

[7] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *IEEE TPAMI*, 45(11):13467–13488, 2023. 5

[8] Ziteng Cui and Tatsuya Harada. Raw-adapter: Adapting pre-trained visual model to camera raw images. In *ECCV*, 2024. 1, 2, 3, 5, 7

[9] Steven Diamond, Vincent Sitzmann, Frank Julca-Aguilar, Stephen Boyd, Gordon Wetzstein, and Felix Heide. Dirty pixels: Towards end-to-end image processing and perception. *ACM Trans. Graph.*, 40(3), 2021. 2

[10] M. Everingham, L. Gool, Christopher K. I. Williams, J. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88:303–338, 2009. 1, 3

[11] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 8

[12] Guangyu Guo, Longfei Han, Le Wang, Dingwen Zhang, and Junwei Han. Semantic-aware knowledge distillation with parameter-free feature uniformization. *Visual Intelligence*, 2023. 7

[13] Hongwei Guo, Haitao He, and Mingyi Chen. Gamma correction for digital fringe projection profilometry. *Appl. Opt.*, 43(14):2906–2914, 2004. 7

[14] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *Computational visual media*, 2023. 6

[15] Yanhui Guo, Fangzhou Luo, and Xiaolin Wu. Learning degradation-independent representations for camera isp pipelines. In *CVPR*, 2024. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 7

[17] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 6

[18] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 7

[19] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *BMVC*, 2021. 2, 3, 4

[20] Xin Jin, Linghao Han, Zhen Li, Zhi Chai, Chunle Guo, and Chongyi Li. Dnf: Decouple and feedback network for seeing in the dark. In *CVPR*, 2023. 1

[21] Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Xialei Liu, Chongyi Li, and Ming-Ming Cheng. Make explict calibration implicit: "calibrate" denoiser instead of the noise model. In *arxiv:2308.03448v2*, 2023. 1

[22] Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Ruixun Zhang, Xialei Liu, and Chongyi Li. Lighting every darkness in two pairs: A calibration-free pipeline for raw denoising. In *ICCV*, 2023. 1

[23] Xin Jin, Pengyi Jiao, Zheng-Peng Duan, Xingchao Yang, Chun-Le Guo, Bo Ren, and Chong-Yi Li. Lighting every darkness with 3dgs: Fast training and real-time rendering for hdr view synthesis. In *NeurIPS*, 2024. 1

[24] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 8

[25] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NeurIPS*, 2020. 5, 6

[26] Zhihao Li, Ming Lu, Xu Zhang, Xin Feng, M. Salman Asif, and Zhan Ma. Efficient visual computing with camera raw snapshots. *IEEE TPAMI*, 46(7):4684–4701, 2024. 2, 4

[27] 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 *ECCV*, 2014. 1, 2, 3, 4, 5

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2, 5, 6

[29] Haoliang Liu, Wei Xiong, and Yu Zhang. Yolo-core: contour regression for efficient instance segmentation. *Machine Intelligence Research*, 2023. 2

[30] 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 *ICCV*, 2021. 6

[31] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 2, 6, 7

[32] Ali Mosleh, Avinash Sharma, Emmanuel Onzon, Fahim Mannan, Nicolas Robidoux, and Felix Heide. Hardware-in-the-loop end-to-end optimization of camera image processing pipelines. In *CVPR*, 2020. 1, 2, 3

[33] Alex Omid-Zohoor, David Ta, and Boris Murmann. Pascalraw: raw image database for object detection. *Stanford Digital Repository*, 2014. 2, 3, 4

[34] Haina Qin, Longfei Han, Juan Wang, Congxuan Zhang, Yanwei Li, Bing Li, and Weiming Hu. Attention-aware learning for hyperparameter prediction in image processing pipelines. In *ECCV*, 2022. 2

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE TPAMI*, 2017. 2, 5, 6

[36] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 6

[37] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 2, 5, 6

[38] Yujin Wang, Tianyi Xu, Fan Zhang, Tianfan Xue, and Jinwei Gu. Adaptiveisp: Learning an adaptive image signal processor for object detection. In *NeurIPS*, 2024. 2

[39] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 7

[40] Chyuan-Tyng Wu, Leo F. Isikdogan, Sushma Rao, Bhavin Nayak, Timo Gerasimow, Aleksandar Sutic, Liron Ainkedem, and Gilad Michael. Visionisp: Repurposing the image signal processor for computer vision applications. In *ICIP*, 2019. 1, 2

[41] Xingxing Xie, Gong Cheng, Qingyang Li, Shicheng Miao, Ke Li, and Junwei Han. Fewer is more: Efficient object detection in large aerial images. *Science China Information Sciences*, 2024. 5

[42] Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, Youliang Yan, and Zhiwei Xiong. Toward raw object detection: A new benchmark and a new model. In *CVPR*, 2023. 1, 2, 3, 4, 5, 7, 8

[43] Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, and Takeshi Ohashi. Dynamicisp: Dynamically controlled image signal processor for image recognition. In *ICCV*, 2023. 2

[44] Ke Yu, Zexian Li, Yue Peng, Chen Change Loy, and Jinwei Gu. Reconfigisp: Reconfigurable camera image processing pipeline. In *ICCV*, 2021. 1, 2, 4

[45] Bo Zhang, Yuchen Guo, Runzhao Yang, Zhihong Zhang, Jiayi Xie, Jinli Suo, and Qionghai Dai. Darkvision: a benchmark for low-light image/video perception. *arXiv preprint arXiv:2301.06269*, 2023. 3

[46] Chang-Bin Zhang, Yujie Zhong, and Kai Han. Mr. detr: Instructive multi-route training for detection transformers. *arXiv preprint arXiv:2412.10028*, 2024. 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 2, 4, 5, 6