面向多样化条件下的RAW目标检测

李钟毓[1] 靳鑫[1] 孙博远[1] 郭春乐[1] 程明明[1,2*]
[1]VCIP, CS, 南开大学 [2]NKIARI, Shenzhen Futian
{lizhongyu,jinxin,boyuansun}@mail.nankai.edu.cn, {guochunle,cmm}@nankai.edu.cn

Abstract

现有目标检测方法通常考虑*sRGB*输入，这些数据原本是通过专为可视化设计的*ISP*从*RAW*格式压缩而来。然而，这种压缩可能会丢失对检测至关重要的信息，尤其是在复杂光照和天气条件下。我们提出了*AODRaw*数据集，该数据集提供*7,785*张高分辨率真实*RAW*图像，包含*135,601*个标注实例，涵盖*62*个类别，捕捉了*9*种不同光照和天气条件下广泛的室内外场景。基于支持*RAW*和*sRGB*目标检测的*AODRaw*，我们提供了评估当前检测方法的全面基准。我们发现，由于*sRGB*与*RAW*之间的域差异，*sRGB*预训练限制了*RAW*目标检测的潜力，这促使我们直接在*RAW*域上进行预训练。然而，*RAW*预训练比*sRGB*预训练更难学习丰富的表征。为了辅助*RAW*预训练，我们从*sRGB*域预训练的现成模型中蒸馏知识。最终，我们在不依赖额外预处理模块的情况下，实现了在多样化和恶劣条件下的显著性能提升。代码和数据集发布于https://github.com/lzyhha/AODRaw。

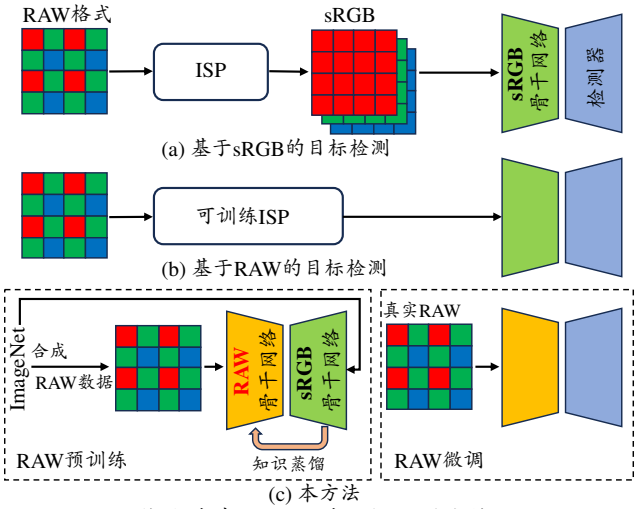


Figure 1. (a) 传统的基于sRGB的目标检测依赖于8位sRGB图像，这些图像由RAW图像压缩而来并丢失了细节信息。(b) 先前基于RAW的方法利用可训练的图像信号处理器(ISP)，将在sRGB域预训练的模型适配到RAW域。(c) 我们直接在RAW域预训练模型，无需ISP模块即可在RAW目标检测任务中取得优异性能。

1. 引言

现实世界中的目标检测是计算视觉领域的一项基础任务。借助COCO [27]和VOC [10]等公开数据集，该领域已取得显著进展。然而这些数据集主要聚焦于sRGB图像，与RAW图像相比会丢失部分关键信息。在典型相机中，传感器首先会捕获高位深的RAW图像。随后图像信号处理器(ISP)将这些RAW图像压缩为8位sRGB图像。与压缩后的sRGB图像不同，RAW图像保持更高位深从而保留了更多可区分信息 [20, 23]，这对计算机视觉任务（尤其在复杂光照和天气条件下）至关重要。更重要的是，在实际应用中直接处理RAW数据可使制造商绕过图像信号处理环节，从而实现更快的处理速度并降低计算开销。因此基于RAW的目标检测已受到关注 [40, 42–44]，并在恶劣条件下展现出优势。但目前该领域的探索仍十分有限。

相关数据集的稀缺性是制约基于RAW格式的目标检测技术发展的关键因素。然而，为目标检测任务采集RAW图像的成本远高于sRGB图像。例如，RAW图像无法像基于sRGB的数据集COCO [27]那样从图片网站获取。因此实地拍摄需要耗费大量人力 [21, 22]，在特殊天气条件下尤为困难。受限于图像采集与标注的困难，现有RAW目标检测方法 [8, 32] 多依赖合成RAW图像，但这些数据缺乏真实的噪声模式与动态范围。现有真实RAW数据集 [19, 33, 42]的类别多样性也严重不足。例如LOD [19]和RAOD [42]数据集分别仅标注了8类和6类目标。此外，部分数据集仅关注日光与弱光下的室外场景，却忽略了其他恶劣天气条件。这导致现有方法应用范围有限，无法充分发挥RAW图像在复杂环境下的优势。

为突破这些限制，我们提出了面向恶劣条件下**RAW**图像目标检测的挑战性数据集(**AODRaw**)。AODRaw采集了真实室内外场景的RAW图像，涵盖2种光照条件（日光/弱光）与3种天气条件（晴天/雨天/雾天）。由于多种光照和天气条件可能同时出现，我们采集了9种不同的环境条件。跨越不同地理位置、城市和场景，我们获得了7,785张图像和135,601个标注实例，其中6,504张图像是在恶劣条件

*Corresponding Author.

下拍摄的。同时，我们的AODRaw数据集标注了62个类别，显著超过现有数据集。场景和语义的多样性可以进一步促进现实世界中基于RAW格式的目标检测技术发展。此外，我们基于AODRaw数据集评估了现有RAW目标检测方法 [8, 42]。

通过AODRaw，我们的目标是设计一个单一模型来同时检测各种环境条件下的目标，而不是像先前某些方法 [8]那样为每种条件单独训练模型。对于RAW目标检测，许多方法通常使用可训练适配器（如神经ISP [8]）将sRGB域预训练的模型迁移到RAW域。然而，sRGB与RAW之间的域差距阻碍了模型理解RAW图像中的复杂信息，同时适配器也会增加计算成本。部分方法 [42]采用从头训练的方式，但有限的数据量制约了模型性能。

不同的是，我们探索在RAW域上进行预训练，以减少预训练与微调之间的领域差距，在不使用任何适配器的情况下实现了显著提升。然而，由于高动态范围和相机噪声等因素，模型从RAW图像中学习高质量表征比从sRGB图像中学习更为困难。为了缓解这一困难，我们提出从现成的sRGB域预训练模型中蒸馏表征。以ConvNext-T [31]和Cascade RCNN [3]为例，基于sRGB的目标检测在AODRaw上实现了34.0%的平均精度（AP）。通过我们的RAW预训练，RAW目标检测将性能提升至34.8% AP。综上所述，我们的主要贡献如下：

- 我们提出了AODRaw数据集，这是一个用于多种光照和天气条件下RAW格式目标检测的高质量数据集。该数据集包含从各类室内外场景采集的多样化复杂图像。
- AODRaw支持多项任务研究，包括恶劣条件下的RAW目标检测和sRGB目标检测。我们评估了现有目标检测方法在这些任务上的表现。
- 我们通过跨域蒸馏在RAW图像上进行模型预训练，在不依赖神经ISP等适配器的情况下取得了显著性能提升。

2. 相关工作

2.1. 目标检测

主流目标检测方法可分为两大类，即多阶段检测器与单阶段检测器。多阶段检测器（如R-CNN系列 [3, 35, 37]）首先生成候选区域，再通过后续阶段进行精细化调整。Cascade R-CNN [3]通过多级精调机制进一步扩展该流程，逐步提升定位与分类精度。尽管这类方法具有较高准确率，但其推理速度较慢。单阶段方法（如YOLO [11, 29]和RetinaNet [28]）直接预测目标位置与类别，虽能实现更快推理速度，但需以精度为代价。此外，基于Transformer的方法 [46, 47]利用自注意力机制建模图像空间关系，但需要更长的训练时间。由于这些方法主要针对sRGB图像设计，我们进一步评估这些经典方法在RAW域的表现。

2.2. RAW目标检测

RAW目标检测利用未经处理的传感器数据，因其在复杂光照和天气条件下的潜力与优势而受到关注。然而，该领域缺乏用于RAW图像预训练模型的大规模数据集，而这对于现代目标检测方法至关重要。因此，部分方法 [42]采用实时检测器 [11]从头开始训练检测模型。由于相机噪声和RAW图像数量有限等劣势，这些方法可能收敛缓慢且性能受限。其他方法 [40, 43, 44]则通过调整在sRGB图像上预训练的模型来适配RAW数据域。其中，一些研究提出了可微分图像信号处理器（ISP）[8, 9, 15, 32, 34, 38]用于RAW图像预处理。通过从COCO数据集 [5, 26]合成RAW图像来微调模型，也有助于缩小域间差距。然而，sRGB预训练仍限制了模型理解RAW图像（比sRGB包含更多信息）的能力，且ISP模块会引入额外成本。为此，我们探索直接在RAW图像上进行模型预训练。

2.3. RAW格式目标检测数据集

现有目标检测数据集（如COCO [27]）主要采集sRGB图像。尽管这些数据集推动了目标检测领域的重大进展，但sRGB图像缺乏RAW格式所包含的细节信息，而这些信息在恶劣条件下尤为重要。由于RAW数据集的稀缺性，许多方法 [8, 32]依赖合成数据集进行RAW检测研究。目前已有少量RAW数据集被提出：例如[45]采集了弱光条件下的图像，[1]采集了雾天环境下的高动态范围驾驶图像。PASCALRAW [33]数据集采用与PASCAL VOC [10]相同的采集方式，提供了4,259张日光环境下的图像。LOD [19]采集了2,230组日光与弱光条件下的配对图像。RAOD [42]包含25,207张标注驾驶图像，但仅涵盖6个类别和2种光照条件。这些数据集普遍存在适用场景受限（如专注驾驶场景）、标注类别范围狭窄或仅针对特定环境条件等问题。

3. AODRaw 数据集

3.1. 数据采集

多样化场景。使用索尼A7M4相机，我们构建了一个具有挑战性的RAW格式目标检测数据集，涵盖各种恶劣及多样化环境条件。具体而言，我们考虑了2种光照条件（即日光与低照度），以及3种天气条件（即晴天、雨天和雾天）。针对不同光照条件，我们同时采集了室内与室外场景图像。由于多种环境条件可能同时出现，我们最终收集了7,785张真实RAW格式图像及其对应的sRGB图像，覆盖9种组合环境条件，如Tab. 2所示。例如Fig. 2中第三行第一列的图像即为雨天结合低照度条件下的样本。

数据多样性。为了在采集RAW图像时尽可能满足训练和评估的需求，同时避免高昂的成本，我们在不同地点、城市和场景下拍摄图像，以确保数据的广泛多样性。即使在同一地点拍摄少量图像，也会选择不同的位置、方向和视角进行采集。如Fig. 2所示，部分图像拍摄于交通场景，其他则涵盖花园、大学、图书馆、

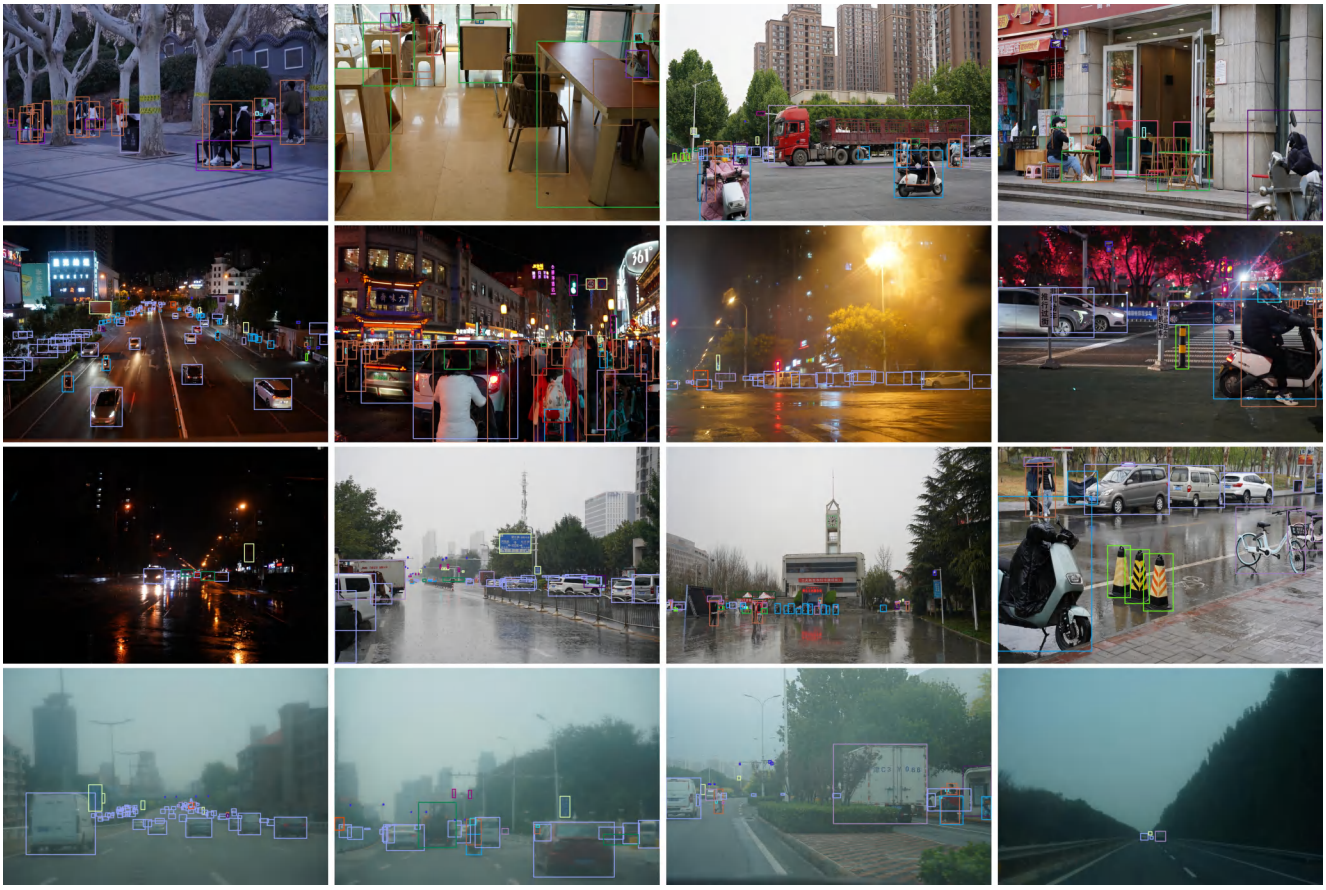


Figure 2. AODRaw中的图像示例。从上至下分别展示了日光、低光照、雨天和雾天条件下的场景。部分图像是在多重条件下拍摄的。例如，第三行首张图像即是在低光照与雨天复合条件下拍摄的。各条件下的更多示例可参阅补充材料。

数据集	分辨率	图像数量	类别数	实例数	每图实例数	拍摄条件
OnePlus [44]	4640 × 3480	141	5	1,228	8.7	1种(低光照)
PASCALRAW [33]	600 × 400	4,259	3	6,550	1.5	1种(日光)
LOD [19]	1200 × 800	2,230	8	9,726	4.4	2种(日光与低光照)
RAOD [42]	2880 × 1856	25,207	6	237,379	9.4	2种(日光与低光照)
AODRaw (本工作)	6000 × 4000	7,785	62	135,601	17.4	9种(见Tab. 2)

Table 1. 与现有RAW数据集的对比分析。

街道以及其他室内场景。

数据标注。我们遵循COCO数据集 [27]的标注格式，对日常生活中常见的62个类别图像中的边界框进行标注。

3.2. 数据分析

本节我们分析AODRaw数据集，并将其与两个先前的目标检测数据集进行比较，即sRGB目标检测的COCO [27]和RAW目标检测的RAOD [42]。在补充材料中，我们展示了针对每种场景的详细分析。

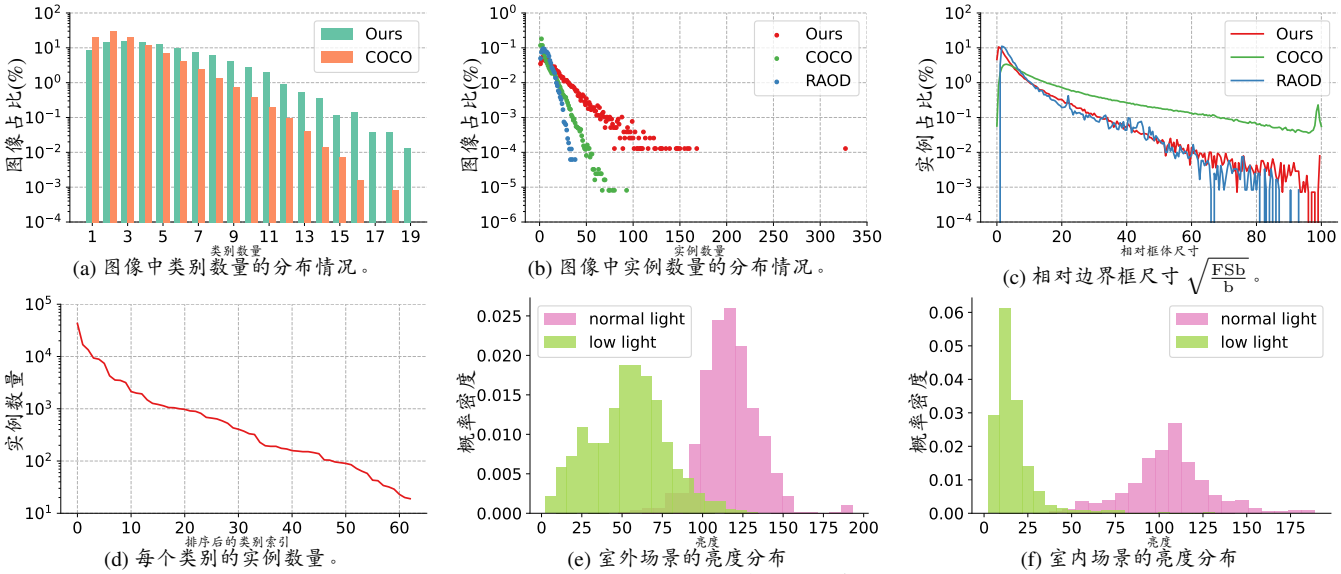
多样化场景。如Tab. 2所示，AODRaw涵盖了9种场景条件，包括2种光照条件、3种天气条件及其不同组合。与现有的RAW目标检测数据集相比（如Tab. 1所示，这些数据集主要关注日光或低光下的户外场景），AODRaw具有更高的多样性，提出了更具挑战性的任务。此外，AODRaw中的场景变化丰富且复杂。如Fig. 3a和Fig. 3b所示，AODRaw中的图像包含不同类别和实例数量，单张图像最多包含19个类别和327个实例。平均而言，如Tab. 1所示，每张图像包含17.4个实例，超过了现有数据集。

类别多样性提升。AODRaw包含62个类别，其数量显著超过现有大多数RAW格式目标检测数据集（如Tab. 1所示）。如Fig. 3d所示，该数据集的类别分布呈现长尾特性，这进一步增加了RAW目标检测任务的挑战性。

目标尺度特征。如Fig. 3c所示，AODRaw中的实例在

光照条件	室内		室外							总计
	日光	低照度	日光				低照度			
天气状况	-	-	晴天	雾天	雨天	雾雨	晴天	雾天	雨天	
图像数量	477	1,210	804	1,110	1,252	244	1,842	325	521	7,785
实例数量	4,992	10,195	18,575	23,636	24,107	5,381	37,282	4,513	6,920	135,601

Table 2. 各场景条件下的图像数量统计。



(a) 图像中类别数量的分布情况。 (b) 图像中实例数量的分布情况。 (c) 相对边界框尺寸 $\sqrt{\frac{FSb}{b}}$。

(d) 每个类别的实例数量。 (e) 室外场景的亮度分布 (f) 室内场景的亮度分布

Figure 3. 统计数据显示我们的AODRaw数据集具有更高的类别和实例多样性

尺寸上具有显著差异，其中小目标占比明显高于既往数据集。这种尺度差异要求检测器必须提取多尺度表征，从而增加了检测任务的复杂度。

空间分布特性。Fig. 4显示，AODRaw数据集中的实例在图像空间分布上更为均匀。这种均匀的空间分布有助于降低空间偏差。由于多数图像采集于室外场景，数据略微呈现向图像底部集中的趋势。

光照分布特性。通过计算sRGB图像平均灰度值得出的亮度分布如Fig. 3e和Fig. 3f所示。分析表明，AODRaw涵盖了广泛的光照条件变化范围。

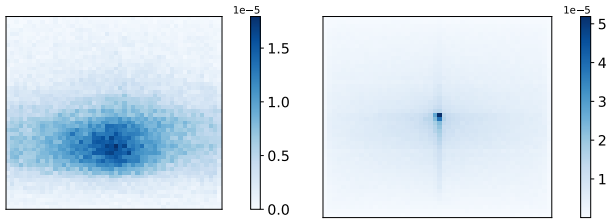
3.3. 数据划分

该数据集按7:3的比例随机划分为训练集和测试集。为确保每个划分都包含足够数量的各类别图像，我们分别对每个类别进行划分后再合并划分结果。最终获得5,445张训练图像和2,340张测试图像，分别包含94,949个和40,652个实例。

4. 基准测试

4.1. 实现细节

模型训练。我们使用流行的代码库mmdetection [4]实现了所有目标检测方法。除Deformable DETR [47]训练100轮次外，其他模型均以16的批量大小训练48轮次。更多超参数请参阅补充材料。此外，RAW图像以拜耳模式存储，形状为$1 \times H \times W$，其中H和W表



(a) 我们的数据集AODRaw。 (b) COCO数据集。

Figure 4. 目标中心点分布对比。

示图像的高度和宽度。为兼容现有模型，我们按照[42]的方法通过去马赛克将RAW图像转换为$3 \times H \times W$格式。遵循[26]的研究，RAW图像还经过伽马校正处理以加速收敛。

图像分辨率。AODRaw数据集中的图像记录分辨率为6000×4000。直接将如此大尺寸的图像输入检测器是不现实的。因此我们采用两种实验方案：1）按照[42]的方法将图像降采样至2000×1333的较低分辨率；2）将图像切割为1280×1280的图块（重叠区域300像素），并忽略与切割图像IoU低于0.4的物体，最终得到71,782张图像和417,781个实例。第一种方案训练速度更快，但面积小于32^2的微小物体在降采样后会消失因而被忽略。第二种方案需要更长的训练时间，但能充分利用高质量标注并支持微小物体检测[7, 41]。下文默认采用降采样方案。

评估方法。我们采用通用指标平均精度（AP）[4, 27]进行评估，同时报告IoU阈值

为0.75和0.50时的AP_{75}与AP_{50}。对于小、中、大物体的AP_s、AP_m和AP_l，降采样方案下的物体面积范围分别设为$[0, 128^2)$、$[128^2, 320^2)$和$[320^2, +\infty)$。图像切割方案下则分别设为$[0, 64^2)$、$[64^2, 160^2)$和$[160^2, +\infty)$。为评估恶劣条件下的检测性能，除正常条件（日光与晴朗天气组合）的AP_{normal}外，我们还报告低光照、雨天和雾天条件下的AP_{low}、AP_{rain}和AP_{fog}。

4.2. 分析

基于AODRaw，我们评估了多种检测器在sRGB和RAW图像上的目标检测性能，如Tab. 3所示。我们测试了若干具有里程碑意义的主流方法，包括多阶段检测器（Faster RCNN [35]、Sparse RCNN [37]和Cascade RCNN [3]）、单阶段检测器（RetinaNet [28]和GFocal [25]）以及基于Transformer的检测器（Deformable DETR [47]）。

恶劣条件下的**sRGB**目标检测。在评估中，Cascade RCNN表现出最优性能，其AP达到25.6%，AP_{normal}为27.3%。然而，AP_{low}、AP_{rain}和AP_{fog}分别仅为23.8%、24.7%和20.4%，表明恶劣条件带来了更大挑战。采用更先进的骨干网络（如ConvNeXt和Swin-T）可提升性能。例如，ConvNeXt-T在AP_{normal}上比ResNet高出9.7%，但在恶劣条件下的提升幅度较小，分别为AP_{low}提升7.7%、AP_{rain}提升6.2%以及AP_{fog}提升6.8%。这种差距揭示了sRGB图像在恶劣条件下的固有缺陷。

恶劣条件下的**RAW**格式目标检测。在RAW图像上进行微调时，直接采用基于sRGB图像预训练的模型是不合适的。例如，采用sRGB预训练的RAW版Cascade RCNN仅获得33.7%的平均精度（AP），这甚至低于基于sRGB方法的34.0%。该现象部分源于sRGB与RAW之间的域差异。如Tab. 4所示，在一个域上训练的检测器在另一个域测试时性能会显著下降，这表明RAW域与sRGB域模型无法实现良好的相互泛化。

为克服域差异，现有RAW域目标检测方法通常将神经图像信号处理器（ISP）与检测器串联，通过ISP将图像从RAW域映射至sRGB域。在Tab. 5中，我们评估了两种最新提出的RAW目标检测方法：RAOD [42]和RAW-Adapter [8]。结果表明神经ISP能有效释放RAW图像的潜力。例如RAOD取得34.4%的平均精度，优于基于sRGB方法的34.0%。特别是在恶劣条件下，RAOD的改进幅度更大（低照度提升0.9%AP_{low}，雨天4.8%AP_{rain}，雾天2.2%AP_{fog}，而正常条件仅提升0.3%AP_{normal}）。但神经ISP会带来额外计算开销，且无法完全弥合RAW与sRGB域的鸿沟，导致预训练模型的知识难以被充分迁移利用。

5. RAW预训练

5.1. 方法

基于上述分析，我们旨在通过直接在RAW图像上进行模型预训练，克服sRGB预训练与RAW微调之间的差距，从而无需神经ISP等预处理模块即可实现卓越性能。

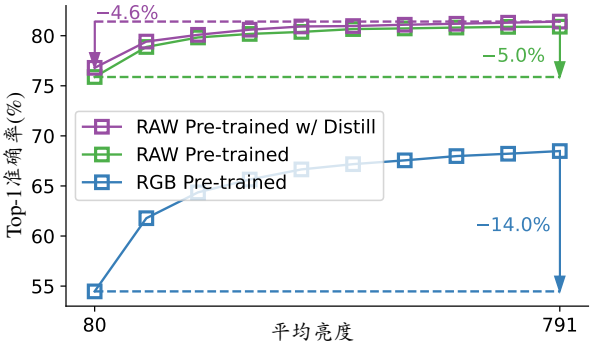


Figure 5. 在不同平均亮度下合成RAW图像时，ImageNet-RAW数据集上的Top-1准确率。图像的最大平均亮度为2^{16}。

合成**ImageNet-RAW**数据集。视觉预训练[14, 16, 17]在ImageNet-1K [36]等百万级规模数据集的助力下取得了巨大进展。然而，要收集与大规模sRGB数据集体量相当的真实RAW数据集并不现实。因此，我们采用逆处理方法[2]从ImageNet-1K合成了一个16位RAW数据集用于RAW预训练。我们将生成的数据集称为ImageNet-RAW。该逆处理操作被嵌入数据增强流程中。通过这种方式，我们在每次迭代中随机调整平均亮度和模拟噪声，使模型能够良好适应不同成像条件。

跨领域蒸馏的**RAW**数据预训练。通过将sRGB输入替换为合成的RAW图像，我们可以利用RAW ImageNet-1K数据集提供的分类目标进行模型预训练，同时保持与sRGB预训练一致的超参数设置。然而，与sRGB域相比，RAW域预训练由于RAW图像固有的噪声和高动态范围（HDR）等因素而面临额外挑战。根据先前研究[5]表明HDR会对训练产生负面影响，我们也发现预训练期间应用伽马校正能提升ImageNet-RAW上的Top-1准确率。因此，我们在预训练中默认采用伽马校正。关于噪声问题，当ImageNet-RAW合成过程中不添加噪声时，模型经过50轮预训练后Top-1准确率达到74.8%。而引入噪声后准确率降至74.4%，表明噪声会对模型性能产生负面影响。为解决这个问题，我们提出采用知识蒸馏[12, 18, 39]技术。作为跨域蒸馏方案，我们选用sRGB域预训练的现成模型作为教师网络，来辅助RAW域的预训练过程。学生网络与教师网络采用相同架构，以确保公平比较。具体而言，我们结合了采用Kullback-Leibler散度损失的逻辑蒸馏与采用L1损失的特征蒸馏方法。

增强对恶劣条件的鲁棒性。RAW预训练和跨域蒸馏增强了模型对不同条件的适应能力。对于单张sRGB图像，在预训练的不同阶段转换为RAW格式时，其亮度会被随机调整，并添加随机噪声。而蒸馏过程则提供了不受合成噪声和亮度影响的稳定目标，促使模型学习对这些条件保持不变的表示。

Fig. 5和Fig. 6验证了该方法对鲁棒性的提升效果。

方法	主干网络	预训练	微调	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. 使用RGB图像进行物体检测的评估，包含不同的预训练和微调设置。

训练	评估	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. sRGB与RAW之间的域间差异。

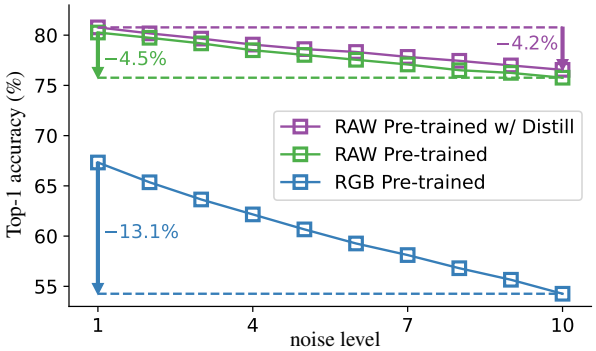


Figure 6. 在合成RAW图像上添加不同噪声水平时的ImageNet-RAW Top-1准确率。此处噪声水平表示散粒噪声的标准差。

通过在ImageNet-RAW验证集上模拟不同亮度和噪声水平，并对预训练模型进行评估，我们可以观察到采用蒸馏预训练的模型在恶劣条件下表现出更强的鲁棒性。例如，当亮度从791降至80时，采用蒸馏的模型性能下降仅为4.6% Top-1准确率，低于未蒸馏时的5.0%。同样地，当噪声水平从1增加到10时，蒸馏模型的性能差距更小，即4.2%对比4.5%。此外，实验表明sRGB预训练模型在调整亮度和噪声时性能下降显著更高（分别为14.0%和13.1%），这证明RAW预训练有效提升了鲁棒性。

Tab. 6进一步展示了蒸馏方法在真实世界RAW目标检测任务中的优势。基于logit的蒸馏使AP提升了0.2%，而基于特征的蒸馏则进一步将改进幅度扩大到0.5%。

5.2. 实验结果

通过RAW数据预训练与蒸馏，ConvNeXt-T在合成的ImageNet-RAW数据集上实现了81.8%的Top-1准确率。尽管该预训练准确率仍低于sRGB预训练（82.1%），但在真实RAW目标检测任务中，各类架构模型均展现出显著性能提升，如Tab. 3所示。例如，相较于sRGB预训练，RAW预训练使Cascade RCNN与ConvNeXt-T的AP指标提升了1.1%。特别值得注意的是，模型在恶劣环境条件下获得更显著的改进：AP_{low}提升0.8%、AP_{rain}提升4.8%、AP_{fog}提升1.2%，而正常条件下仅提升0.9%。与基于sRGB的目

方法	主干网络	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. 与将sRGB域预训练模型适配到RAW图像的方法的比较。

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. 使用Cascade RCNN和ConvNeXt-T时知识蒸馏的消融实验。

标检测方法相比，我们的方法实现了0.8%的AP提升。这些结果充分证明了RAW预训练方案的优势。

6. 切片AODRaw 实验

Section 4和Section 5节的实验采用降采样设置。Tab. 7进一步列出了切片设置下的实验结果，其中超参数遵循降采样实验配置，但我们将模型微调周期调整为12轮。ConvNeXt的实验结果与Tab. 3呈现相同趋势。将在sRGB域预训练的模型迁移至RAW目标检测时，其AP值较sRGB目标检测下降0.2%，而RAW预训练可使性能提升0.8%。值得注意的是，当使用Swin Transformer时，即使采用sRGB预训练，RAW目标检测性能仍优于sRGB目标检测。这可能是因为切片图像保留了更丰富的视觉信息，相较于降采样图像更有利于模型收敛。同时，RAW预训练可额外带来0.6%的AP提升，在恶劣条件下表现尤为显著。总体而言，这些结果进一步验证了RAW预训练的有效性。

7. 实时目标检测实验

我们还评估了YOLO系列的实时目标检测性能。如Tab. 8所示，YOLO-MS-XS [6]和YOLO-v8-n [24]在参数量小于5M的情况下，分别以24.7%和18.9%的平均精度（AP）实现了高帧率。但在恶劣条件下性能显著下降，例如YOLO-v8-n在低照度、雨天和雾天场景下的AP分别为16.3% AP_{low}、16.8% AP_{rain}和15.4% AP_{low}。现有方法尝试通过可训练图像信号处理器（ISP）来提升性能。本文以最新提出的[42]方法为例进行分析。在Tab. 8中，将[42]与YOLO-v8-n结合后，其低照度和雨天场景下的AP分别提升至16.9% AP_{low}和18.9% AP_{rain}。但该方法导致帧率大幅下降，破坏了检测器的实时性。综上所述，本文提出的AODRaw为推进RAW格式实时目标检测研究提供了新的基础。

8. 结论

本文提出了AODRaw，这是一个针对多样化恶劣条件下基于RAW格式的目标检测的挑战性数据集。与传统sRGB数据集相比，AODRaw提供了保留关键视觉信息的多样化RAW图像，适用于复杂光照和天气条件下的目标检测任务。基于AODRaw，我们对现有RAW目标检测方法进行了系统评估。同时，我们采用跨域知识蒸馏技术直接在RAW域上预训练模型，解决了sRGB预训练与RAW微调之间的域差异问题。通过这种方法，我们显著提升了模型性能（特别是在恶劣条件下），且无需依赖额外的预处理模块。本数据集既是评估检测方法的基准平台，也为开发具有跨条件泛化能力的检测方法奠定了基础。我们的研究揭示了RAW预训练在推动现实世界目标检测方面的潜力，并鼓励进一步探索利用RAW图像应对挑战性环境的研究方向。

致谢 本研究得到国家自然科学基金（62225604）和深圳市科技计划（JCYJ20240813114237048）的资助。计算资源由南开大学超级计算中心（NKSC）提供支持。

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. 2

[2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 5

[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] Linwei Chen, Ying Fu, Kaixuan Wei, Dezhi Zheng, and Fe-

方法	主干网络	预训练	微调	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. 基于切片RGB图像的全天候目标检测结果。

方法	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. 使用降采样图像进行实时目标检测的评估。模型训练和评估的输入尺寸为1280×1280。†表示采用[42]提出的可训练预处理模块。同时，我们使用NVIDIA 3090 GPU测量了所有模型的每秒帧数(FPS)。

lix Heide. Instance segmentation in the dark. *International Journal of Computer Vision*, 131(8):2198–2218, 2023. 2, 5

[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. 7, 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. 4

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

[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. 5

[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. 5

[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. 5, 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. 5

[18] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[19] Yang Hong, Kaixuan Wei, Linwei Chen, and Ying Fu. Crafting object detection in very low light. In *BMVC*, 2021. 1, 2, 3

[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. 7, 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

[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

[33] Alex Omid-Zohoor, David Ta, and Boris Murmann. Pascalraw: raw image database for object detection. *Stanford Digital Repository*, 2014. 1, 2, 3

[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. 5

[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. 5

[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. 4

[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, 3

[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. 2

[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