Towards Reflected Object Detection: A Benchmark

Yiquan Wu
Nanjing University of Aeronautics and Astronautics
Nanjing, China, 211106
nuaaimage@163.com

Zhongtian Wang
Nanjing University of Aeronautics and Astronautics
Nanjing, China, 211106

You Wu
Guilin University of Technology
Guilin, China, 541006

Ling Huang
Guilin University of Technology
Guilin, China, 541006

Hui Zhou
Guilin University of Technology
Guilin, China, 541006

Shuiwang Li(✉)
Guilin University of Technology
Guilin, China, 541006
lishuiwang0721@163.com

Abstract

Object detection has greatly improved over the past decade, thanks to advances in deep learning and large-scale datasets. However, detecting objects reflected on surfaces remains an underexplored area. Reflective surfaces are ubiquitous in daily life, appearing in homes, offices, public spaces, and natural environments. Accurate detection and interpretation of reflected objects are essential for various applications. This paper addresses this gap by introducing an extensive benchmark specifically designed for Reflected Object Detection. Our Reflected Object Detection (ROD) dataset features a diverse collection of images showcasing reflected objects in various contexts, providing standard annotations for both real and reflected objects. This distinguishes it from traditional object detection benchmarks. The ROD dataset encompasses 10 categories and 6 reflective surfaces, including 23,520 images of real and reflected objects on different backgrounds, complete with standard bounding-box annotations and the classification of objects as real or reflected. In addition, we present baseline results by adapting five state-of-the-art object detection models to address this challenging task. The experimental results underscore the limitations of existing methods when applied to reflected object detection, highlighting the need for specialized approaches. By releasing the ROD dataset, we aim to support and advance future research on detecting reflected objects. The dataset and code are available at: https://github.com/jirouvan/ROD.

***Keywords:** Reflected object detection Benchmark Object detection ROD dataset.*

1. Introduction

The field of object detection has seen remarkable advancements over the past decade, driven by the development of deep learning techniques and the availability of large-scale datasets [51, 54, 2]. These advancements have significantly improved the accuracy and robustness of object detection systems in various applications [29]. However, one area that remains underexplored is the detection of objects reflected in surfaces, such as glass, metal, water, plastic, polishing and tile. See Fig. 1 for an illustration of the difference between conventional object detection and reflected object detection.

Reflective surfaces are ubiquitous in our daily lives, appearing in a wide array of environments and applications [35, 49, 24]. Mirrors, glass windows, water surfaces, and

polished metals are just a few examples of materials that produce reflections. These reflective surfaces are prevalent in various settings, including homes, offices, public spaces, and natural environments, making the ability to detect and interpret reflected objects a crucial aspect of many technological applications [35, 25, 19, 28]. For instance, in surveillance, security systems can more effectively identify real intrusions or threats by differentiating reflections from genuine objects [40, 41, 38]. For autonomous driving, accurate identification of real objects versus reflections enables vehicles to navigate more safely and avoid accidents caused by misinterpretation [34, 21, 50, 9]. For service robots, robots can perform tasks with greater accuracy, such as picking and placing items, by correctly identifying real objects instead of their reflections [36, 24]. This improved object detection also facilitates better navigation in environments with reflective surfaces, such as warehouses [31, 7]. In smart homes, systems can provide more tailored responses by recognizing when a person is truly present rather than reacting to their reflection [33, 42, 6]. In medical applications, imaging and diagnostic tools yield more accurate results when they accurately interpret reflections, leading to better patient outcomes and more precise medical interventions.

Given the widespread presence of reflective surfaces in daily life, developing technologies that can effectively detect and interpret reflected objects is essential. This capability can enhance the performance and reliability of various applications, including smart home systems, surveillance, autonomous driving, and medical devices. However, to the best of our knowledge, there is currently no public benchmark for reflected object detection. This paper aims to address this gap by introducing a benchmark specifically designed for this purpose. We propose a comprehensive benchmark that includes a diverse set of images featuring reflected objects in various contexts. Our benchmark is designed to test the limits of current object detection methods and provide a standardized evaluation framework for developing and comparing new algorithms tailored to reflected object detection. The benchmark provides standard annotations used in object detection for identifying both actual (real) objects and their reflections. Additionally, it offers extra details that indicate whether an object is real or a reflection. This feature distinguishes it from traditional object detection benchmarks, which typically do not provide information about whether an object is a reflection. In addition to introducing the benchmark, this paper also presents baseline results by adapting several state-of-the-art object detection models. These results highlight the limitations of existing methods when applied to reflected object detection and underscore the need for specialized approaches. We analyze the performance of these models across different reflection scenarios and provide insights into the specific challenges posed by reflections.

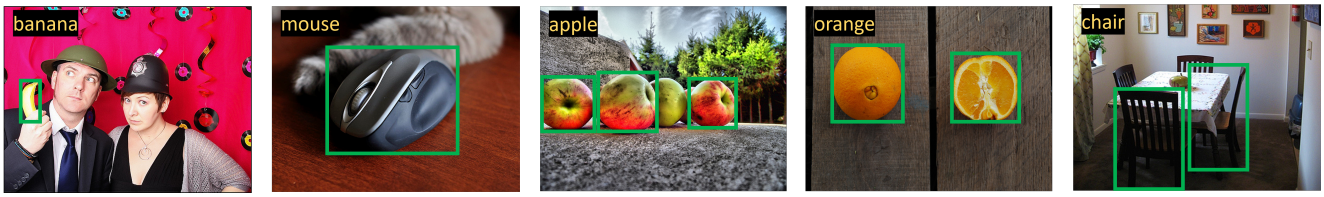
The deployment of reflective object detection (ROD) systems in surveillance scenarios raises critical privacy concerns. Security systems utilizing ROD may inadvertently capture individuals' private activities behind reflective surfaces (e.g., windows, mirrors). Improper handling or leakage of such data could compromise individual privacy rights and entail legal liabilities. To mitigate these risks, developing ROD technologies necessitates the establishment of comprehensive ethical guidelines and regulatory frameworks to ensure lawful, transparent, and privacy-preserving implementations. Concurrent technical safeguards—such as encrypted data transmission protocols, role-based access controls, and edge-computing architectures for localized data processing—should be prioritized to enhance security while maintaining functional efficacy.

1.1. Contribution

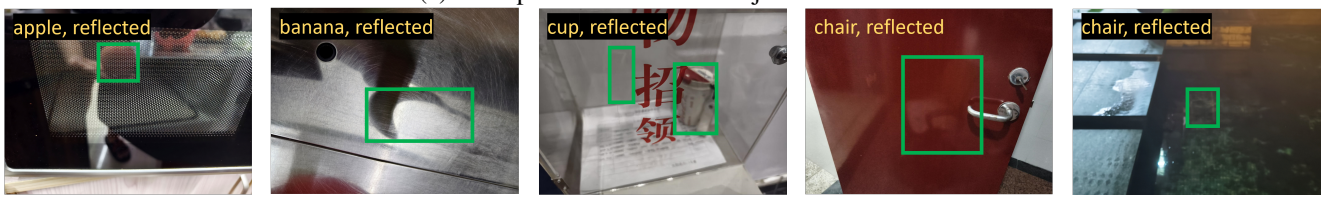
In this work, we make the first attempt to explore reflected object detection by introducing the ROD benchmark, which is specifically designed for detecting reflected objects. This benchmark provides a well-annotated dataset and robust evaluation metrics to facilitate research in this challenging area. The ROD benchmark fills a crucial gap in current object detection methods by focusing on reflected objects. It aims to provide researchers with a valuable resource to develop and test algorithms that handle the complexities of reflected objects. ROD dataset comprises a diverse set of 10 classes and 6 reflective surfaces of generic objects, totaling 23,520 images annotated with axis-aligned bounding boxes, category labels, and object nature (real or reflected). Sample images from the ROD dataset are illustrated in Fig. 2. In addition, we developed five baseline detectors based on five state-of-the-art algorithms, namely RO-YOLOv8, RO-YOLOv10, RO-RTMDet, RO-YOLOX, and RO-PPYOLOE. These baselines serve to evaluate detectors' performance and provide benchmarks for future research on ROD dataset. In summary, our contributions include:

- We make the first attempt to explore detecting reflected objects, a previously underexplored area in object detection. By focusing on this unique challenge, we hope to inspire further research and innovation in the detecting reflected objects.

- We introduce ROD dataset, the first benchmark dedicated to detecting reflected objects, consists of 10 classes and 6 reflective surfaces of generic objects, with 23,520 images annotated with bounding boxes, object categories, and the characteristics of the objects. This dataset will enable detailed analysis and evaluation of algorithms developed for detecting reflected objects.



(a) Example of conventional object detection.



(b) Example of detecting reflected objects.

Figure 1: While previous object detection focused on the identification and localization of objects, this work focuses on information beyond that and concerns about the nature of objects in addition, as shown in (a) and (b), respectively. Note the nature of the objects (i.e., real or reflected) are marked in (b) additionally.

- To support further research on ROD dataset, we develop five baseline detectors based on state-of-the-art models: RO-YOLOv8, RO-YOLOv10, RO-RTMDet, RO-YOLOX, and RO-PPYOLOE. These baseline models will provide initial performance metrics and serve as reference points for future studies.

2. Related Work

2.1. Object Detection Algorithms

Object detection has been a critical area of research in computer vision, significantly advancing over the past few decades. Traditional object detection methods relied heavily on handcrafted features and shallow learning techniques. The advent of deep learning has revolutionized this field, leading to the development of more robust and accurate algorithms. Modern object detection methods are categorized into two types: two-stage detectors and one-stage detectors. Two-stage detectors, such as R-CNN [16], Fast R-CNN [15], Faster R-CNN [39], and Mask R-CNN [18], initially generate region proposals and then refine them through classification and bounding box regression, achieving high precision and efficiency. Variants like Cascade R-CNN [5] further enhance detection performance through multi-stage detection and regression. One-stage detectors, including SSD [30], YOLO, RetinaNet [26], and EfficientDet [43], predict object locations and categories in a single step, providing faster performance suitable for real-time applications. The YOLO series has evolved to YOLOv8 [44] and YOLOv10 [45], further optimizing speed and accuracy.

Despite these advancements, identifying objects reflected in surfaces like mirrors and glass remains a particularly challenging and underexplored problem. Reflections can severely distort object appearance, introducing ambiguous visual cues that complicate the detection process. Most current object detection algorithms are not designed to distinguish between real objects and their reflections, which can lead to frequent misclassifications. These algorithms often struggle to differentiate between an actual object and its mirror image, resulting in false positives and reduced accuracy, especially in environments rich in reflective surfaces, such as bathrooms, retail stores, or even city streets with glass-fronted buildings. Our work aims to address this gap by introducing a benchmark and developing specialized approaches for reflected object detection.

2.2. Object Detection Benchmarks

Object detection benchmarks play a crucial role in the development and evaluation of detection algorithms by providing standardized datasets and evaluation metrics that facilitate consistent and fair comparisons among different approaches. Over the years, several prominent benchmarks have emerged, each contributing uniquely to the field, such as PASCAL VOC[11], MS COCO [27], and ImageNet [20]. These benchmarks provide large-scale images and standardized evaluation metrics. For instance, PASCAL VOC comprises 20 categories with 11,530 images and 27,450 annotated bounding boxes. ImageNet covers 200 categories with approximately 500,000 annotated bounding boxes. MS COCO includes 91 categories, over 300,000 images, and 2.5 million annotated instances. These datasets have been instrumental in pushing the boundaries of object detection research, promoting the development of more accurate and robust models. In addition to these established benchmarks, several domain-specific benchmarks have emerged to address particular challenges in object detection. For instance, KITTI [14] focuses on autonomous driving scenarios, providing annotated data for detecting objects such as cars, pedestrians, and cyclists in street scenes. UAVDT (UAV Detection and Tracking) [10] provides benchmarks for aerial object detection, emphasizing challenges unique

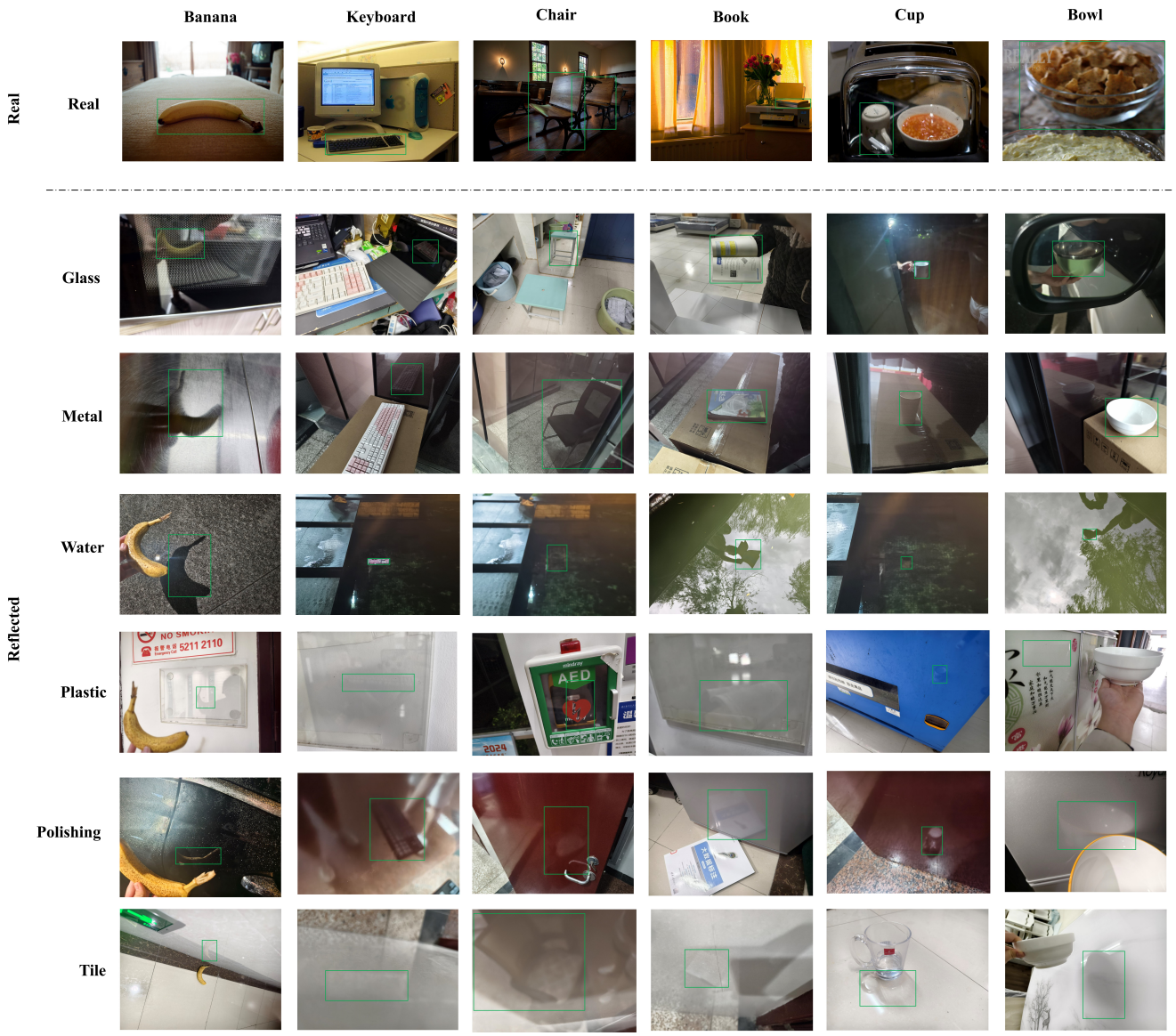
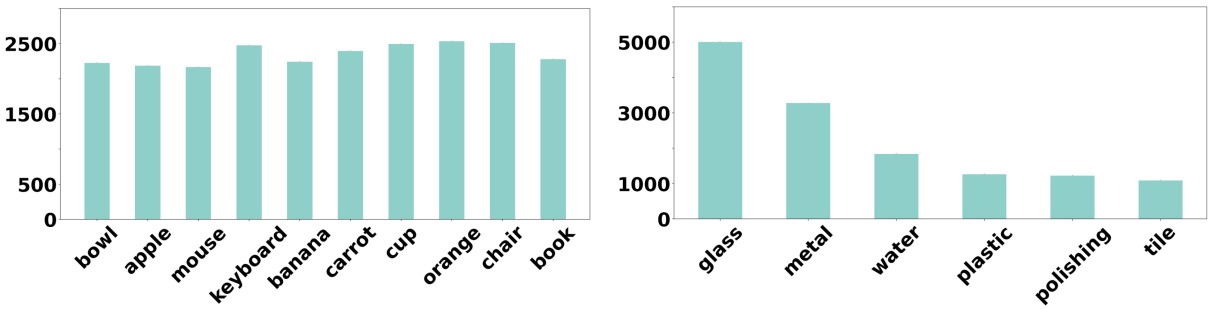


Figure 2: Samples from six categories (i.e., 'banana', 'keyboard', 'chair','book', 'cup', and 'bowl', from left to right) and their corresponding natures (i.e., 'real' and 'reflected' from top to bottom) in the ROD dataset. Note that the objects have been marked with green bounding boxes.

to unmanned aerial vehicle (UAV) imagery, such as varying altitudes and viewpoints.

Despite significant advancements in object detection, no public benchmark specifically targets reflected object detection. This gap hinders the development and evaluation of algorithms for handling reflections. Reflective surfaces are common in real-world scenarios such as surveillance, autonomous driving, and smart homes. Accurate detection of reflected objects is crucial for enhancing performance and safety in these applications. This paper addresses this gap by introducing a benchmark specifically tailored for reflected object detection. This benchmark includes a variety of scenes with reflective surfaces, such as mirrors, windows, and glossy floors, providing a diverse set of scenarios where reflections are prominent. The benchmark not only serves as a tool for evaluating the performance of detection algorithms in these challenging conditions but also encourages the development of more sophisticated methods capable of distinguishing between real objects and their reflections.



(a) Number of images per category in ROD dataset.

(b) Number of images per reflected in ROD dataset.

Figure 3: Statistics for each object category and reflection nature in the dataset.

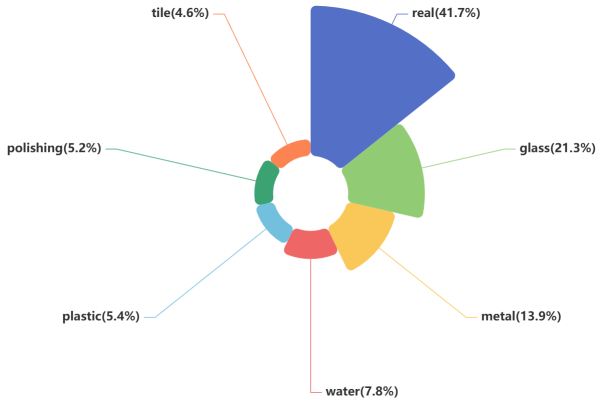


Figure 4: Number of images containing real or reflected objects in the ROD dataset.

2.3. Dealing With Mirrors and Reflections in Vision

Mirrors or other reflective surfaces are common in natural images, and can cause false positive results in the tasks of detection, segmentation, counting, robotic navigation, scene reconstruction, and etc [4, 8, 35, 25, 19, 28]. Reflection detection focuses on identifying regions in an image that contain reflections. When we take a picture through glass windows, the photographs are often degraded by undesired reflections. One of the primary approaches to dealing with reflections involves removing or suppressing the reflections in images. For instance, Abiko et al. employed generative adversarial networks (GANs) to enhance the quality of reflection removal, yielding more natural and clear images [1]. Arvanitopoulos et al. propose a single image reflection suppression method based on a Laplacian data fidelity and an l-zero gradient sparsity regularization term [3]. Particularly, mirror surface detection aims to identify and segment mirror surfaces within a scene. For instance, Yang et al. proposed to address the mirror segmentation problem with a computational approach [49]. Since then, numerous methods have been developed to address mirror detection and segmentation [35, 25, 19, 28]. As these methods have progressed, several specialized datasets have been created to assess their performance. Notably, both the MSD and Progressive Mirror Detection datasets share the goal of advancing mirror detection by providing images and annotations for training and evaluation. However, the MSD dataset [49] is smaller in scale and focuses primarily on indoor scenes, offering limited scene diversity. In contrast, the Progressive Mirror Detection dataset [25] is larger, encompassing both indoor and outdoor scenes with more diverse data and higher-quality annotations. These advancements in datasets, alongside the progression of algorithms, continue to drive innovation in dealing with mirrors and reflections.

Despite extensive research efforts dedicated to dealing with mirrors and reflections in vision, most of these works focus primarily on identifying, localizing, segmenting, and suppressing reflective regions in images. In this work, we make the first attempt to differentiate reflected objects from real ones, a critical capability for various applications, including surveillance, autonomous driving, service robots, and smart homes.

3. Benchmark for Reflected Object Detection

We construct a dedicated dataset for Reflected Object Detection (ROD) dataset, which is a dataset that contains labels of both class and object nature, with prediction bounding-box labeled for each image.

3.1. Image Collection

For image collection, We selected 10 objects with 6 common reflective surfaces in daily life, guided by the selection principles of PASCAL VOC [12] and COCO [27]. The chosen objects for ROD dataset are bowl, apple, mouse, keyboard, banana, carrot, cup, orange, chair, and book, all of which are categories included in the COCO dataset. However, gathering varied images of these objects or their reflected ones in different scenes can be challenging. To address this, we initially sourced images using web crawlers

and online repositories that focus on real-world scenarios with reflective surfaces. Additionally, we conducted field photography sessions in various environments such as homes, offices, and public spaces to capture images that include mirrors and other reflective surfaces. To ensure that the dataset was representative of real-world conditions, we made sure to capture images under various lighting conditions and from different angles. The final collection comprises 23,520 images, encompassing 10 distinct objects (bowl, apple, mouse, keyboard, banana, carrot, cup, orange, chair, and book) and 2 attributes that indicate the nature of the objects (i.e., real or reflected). These objects are represented across 6 types of reflective surfaces (i.e., glass, metal, water, plastic, and tile). The shooting location is Guilin, Nanjing and Bengbu City of China. The vivo x100 pro is used to screen the pictures while ensuring the natural and clear objects in the pictures. The resolution will be reduced to 1600x1200 and 1200x900 in the later stage, and the coding method is H.265. Fig. 2 presents some sample images from ROD dataset, demonstrating that each object category is captured in multiple scenes.

3.2. Annotation

This section provides a detailed introduction to the image annotation process, covering three aspects: category, bounding box, and the nature of the object, as follows:

- **Category:** one of: bowl, apple, mouse, keyboard, banana, carrot, cup, orange, chair, and book.
- **Bounding box:** an axis-aligned bounding box that encloses the visible part of the object in the image.
- **Nature of the object:** a real or reflected object.

We follow three steps, i.e., manual annotation, visual inspection, and box refinement, to complete the annotation of images, guided by the annotation guidelines proposed in [12] and [27]. Specifically, all the images are first annotated by an expert, i.e., a student engaged in object detection, during the initial stage. Manual annotation can lead to occasional errors or inconsistencies, prompting the verification team to carefully review the annotated files in the second step. Annotation errors identified by the validation team in the third stage will be sent back to the initial annotation stage for refinement. By employing this three-stage strategy, the dataset ensures its contained objects have high-quality annotation. We imported all the images into the label-studio tool and annotated each image carefully, exported them into COCO and VOC data set formats, and provided three formats of annotated files: json, xml and excel. Fig. 2 displays five examples of box annotations from ROD dataset.

3.3. Dataset Statistics

The statistics of the ROD dataset are summarized in Fig. 3. Fig. 3 (a) presents a histogram showing the number of images in the dataset for each category. Observations indicate that the number of objects in each category is relatively balanced, with the 'chair' category being the most prevalent, comprising 2,512 images. Fig. 3 (b) presents a histogram that illustrates the dataset is dominated by the 'glass' category, which contains 5,005 images. The 'metal' category follows with 3,278 images, while 'water' has 1,839. The 'plastic', 'polishing', and 'tile' categories contain 1,264, 1,226, and 1,093 images, respectively. This distribution underscores the prominence of 'glass' and 'metal' as the most frequent reflective surfaces. The glass and metal materials are smooth, the reflection is more obvious, and the detailed texture of the reflected object is more, while the reflective surface such as the polishing of the plastic box is more fuzzy, the surface is more rough, and the detailed characteristics of the reflective surface are less.

Fig. 4 further illustrates the number of images containing real or reflected objects. This detailed breakdown highlights the distribution and prevalence of each object category within the dataset, offering insights into its composition and the representation of reflections. The dataset contains 9,815 real images and 13,705 images of reflected objects. A total of 15,112 images were captured using the Vivo X100 Pro camera, which includes 6,695 images at a resolution of 1600x1200 and 8,417 images at a resolution of 1200x900. Additionally, 8,407 images were sourced from open-source image sites, web crawlers, and public image portfolios. To facilitate training and evaluation, the ROD dataset is divided into two primary subsets: the training set and the test set, with a ratio of 7:3.

4. Baseline Detectors for Detecting Reflected Objects

We develop five baseline detectors based on five state-of-the-art object detection algorithms, i.e., RTMDet [32], YOLOv10 [45], YOLOv8 [44], YOLOX [13], and PPYOLOE [48], to facilitate the development of detecting reflected objects. For each model, we add an additional head or branch to predict the nature of the objects without altering the overall framework. The resulting baseline detectors are named RO-RTMDet, RO-YOLOv10, RO-YOLOv8, RO-YOLOX, and RO-PPYOLOE, respectively. Given space constraints and the fact that YOLOv10, YOLOv8, YOLOX, and PPYOLOE are all YOLO variants, we detail only RO-RTMDet, YOLOv8 and RO-YOLOv10 in the following sections. The extension to YOLOX and PPYOLOE is straightforward and will not be elaborated upon here.

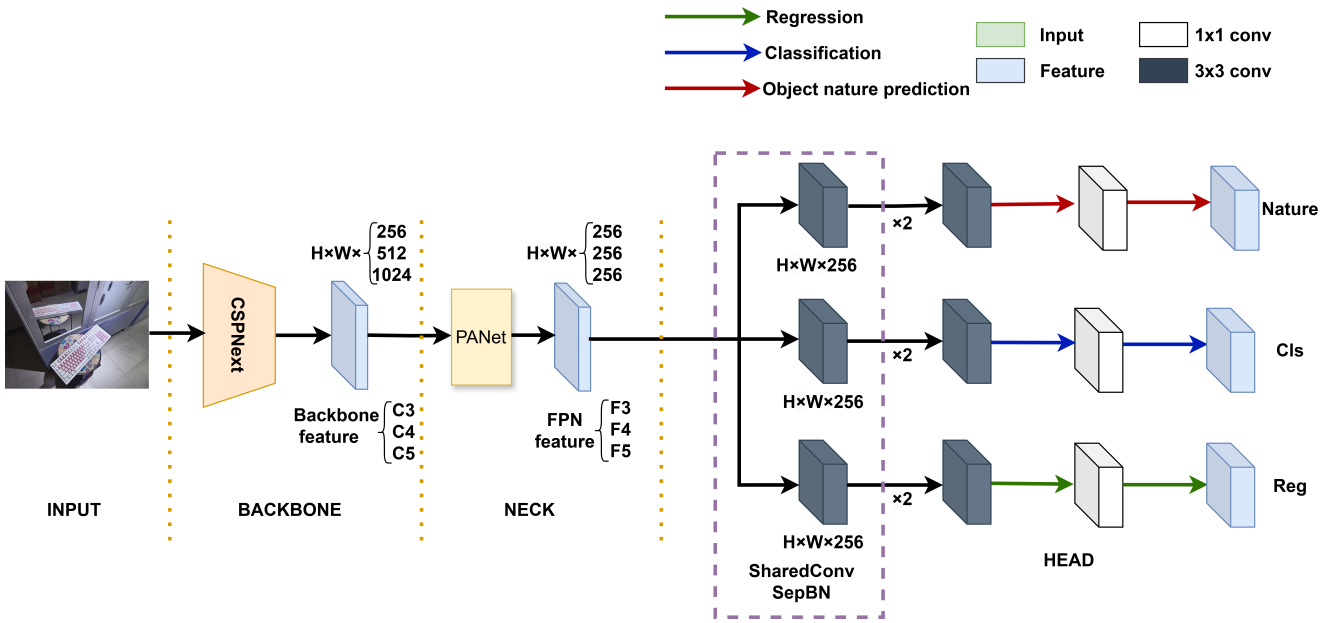


Figure 5: The network structure of the RO-RTMDet detector, inherited from RTMDet, is different from the addition of an additional branch head for the object nature.

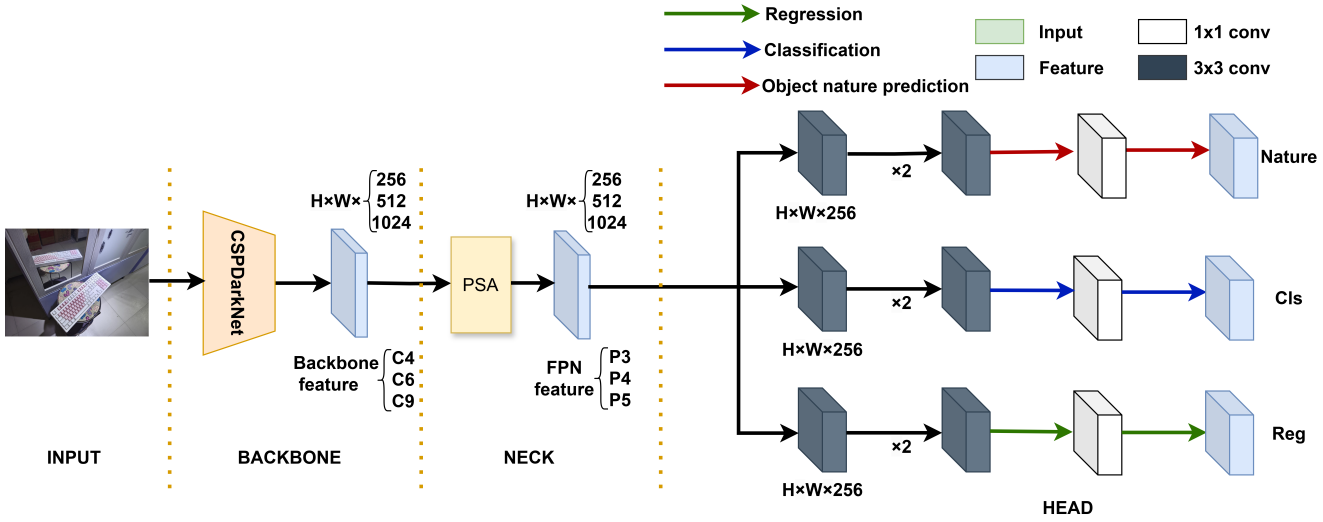


Figure 6: The network structure of the RO-YOLOv8 detector is inherited from YOLOv8, except for the addition of an additional reflected nature branch head.

4.1. RO-RTMDet

The network architecture of the proposed RO-RTMDet is shown in Fig. 5. CSPNet [46] serves as the backbone, generating output features C3, C4, and C5 with 128, 256, and 512 channels, respectively. These features are fused into CSP-PAFPN [32], the neck of RO-RTMDet, which employs the same block as the backbone. The classification head and the regression head are two parallel components used for classification and regression, respectively, forming the head of the original RTMDet. Building upon the original RTMDet model, we introduce a new classification head to predict the nature of objects (i.e., real or reflected). During RO-RTMDet training, the overall loss of the model is

defined as follows:

$$L = L_{cls} + L_{reg} + \lambda L_{nat}, \tag{1}$$

where L_{cls}, L_{reg}, and L_{nat} represent the losses for classification, regression, and object nature prediction, respectively. λ is a constant that weights the loss for the reflected objects prediction head. Below are their specific definitions:

$$\begin{aligned}
L_{cls} &= \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \sum_{cls \in classes} -|y_n^{cls} - p_n^{cls}|^{\beta} \\
&\quad ((1 - y_n^{cls}) log(1 - p_n^{cls}) + y_n^{cls} log(p_n^{cls})), \\
L_{reg} &= \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \Bigg[1 \\
&\quad - (\text{IOU}(b_n^t, b_n^p) - \frac{|C - b_n^t \bigcup b_n^p|}{|C|})\Bigg], \\
L_{nat} &= \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \sum_{nat \in natures} -|y_n^{nat} - p_n^{nat}|^{\beta} \\
&\quad ((1 - y_n^{nat}) log(1 - p_n^{nat}) + y_n^{nat} log(p_n^{nat})),
\end{aligned} \tag{2}$$

where y_n^{cls} and y_n^{nat} are the labeled value of classification and the object nature, p_n^{cls} and p_n^{nat} are the corresponding predictions, N_{pos} is the number of positive anchor, β is the hyperparameter for the dynamic scale factor, which is set to 2, b_n^t and b_n^p represent the ground truth bounding boxes and the prediction, respectively; IOU and C are the IOU loss function and the smallest enclosing convex box of these two bounding boxes. We utilize the same training pipeline as RTMDet for training RO-RTMDet.

4.2. RO-YOLOv8

The network architecture of the proposed RO-YOLOv8 detector is shown in Fig. 6. RO-YOLOv8 uses a modified CSPDarknet [46] as backbone. It replaces the CSPLayer used in YOLOv5 with a C2f module [44]. These features are input into the neck to enhance feature representation, which is made up of the PAN (Path Aggregation Network). The original YOLOv8 model includes two types of heads: one for regression and another for classification tasks. In RO-YOLOv8, we introduce an additional prediction head for detecting the nature of objects. During RO-YOLOv8 training, the overall loss of the model is defined as follows:

$$L = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg} + \lambda_{dfl} L_{dfl} + \lambda_{nat} L_{nat}, \tag{3}$$

where L_{cls} and L_{nat} represent the losses for classification and the object nature prediction, respectively, while L_{reg} and L_{dfl} indicate the Complete Intersection over Union (CIoU) Loss [52] and the Distribution Focal loss (DFL) [22]. λ_{cls}, λ_{reg}, λ_{dfl}, λ_{nat} are constants to weight these loss terms. The definitions of L_{cls} and L_{nat} are provided below. The specific definition of L_{reg} and L_{dfl} are omitted, as it is too intricate to elaborate on here and may divert from the main focus of our discussion. For a comprehensive understanding of L_{reg} and L_{dfl}, we recommend referring to the detailed explanations provided in the original documentation by Zheng et al. [52] and Li et al. [22].

$$\begin{aligned}
L_{cls} &= \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \sum_{cls \in classes} y_n^{cls} log(p_n^{cls}) \\
&\quad + (1 - y_n^{cls}) log(1 - p_n^{cls}), \\
L_{nat} &= \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \sum_{nat \in natures} y_n^{nat} log(p_n^{nat}) \\
&\quad + (1 - y_n^{nat}) log(1 - p_n^{nat}),
\end{aligned} \tag{4}$$

where y_n^{cls} and y_n^{nat} are the labeled value of classification and the object nature, p_n^{cls} and p_n^{nat} are the corresponding predictions, N_{pos} is the number of positive anchor. We utilize the same training pipeline as YOLOv8 for training RO-YOLOv8.

4.3. RO-YOLOv10

The network architecture of the proposed RO-YOLOv10 detector is shown in Fig. 7. RO-YOLOv10 uses a modified CSPDarknet as backbone. It replaces the C2f module used in YOLOv8 [44] with a compact inverted block (CIB) module and introduces an efficient partial self-attention (PSA) module [45]. These features are input into the neck to enhance feature representation, which is made up of the PAN (Path Aggregation Network). The original YOLOv10 model has two types of heads: (1) a one-to-many (o2m) head for regression and classification tasks, and (2) a one-to-one (o2o) head for precise localization. In RO-YOLOv10, we add object nature prediction branch into both of these two heads. During RO-YOLOv10 training, the overall loss of the model is defined as follows:

$$\begin{aligned}
L &= L_{o2m-head} + L_{o2o-head}, \\
L_{o2m-head} &= L_{o2m-cls} + \lambda L_{o2m-nat} \\
&\quad + L_{o2m-reg} + L_{o2m-dfl}, \\
L_{o2o-head} &= L_{o2o-cls} + \lambda L_{o2o-nat} + L_{o2o-reg} \\
&\quad + L_{o2o-dfl}.
\end{aligned} \tag{5}$$

In the o2m head, $L_{o2m-cls}$ and $L_{o2m-nat}$ represent the losses for classification and object nature prediction, respectively, while L_{reg} and L_{dfl} indicate the Complete Intersection over Union (CIoU) Loss [52] and the Distribution Focal loss (DFL) [22]. Similarly, each loss function in the o2o head carries the same meaning as in the the o2m head. λ is a constant that weights the loss for the object nature prediction branch. Below, the L_{cls} and L_{nat} in the o2m head

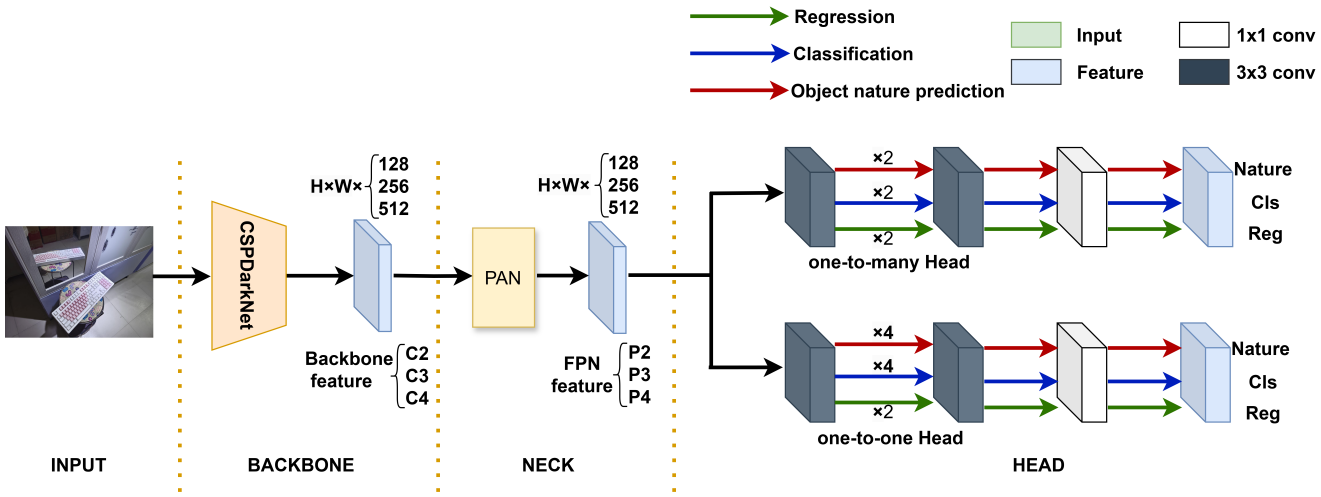


Figure 7: The network structure of the RO-YOLOv10 detector is inherited from YOLOv10, except for the addition of an additional reflected nature branch head.

are used as examples to provide their specific definitions. The specific definition of L_{reg} and L_{dfl} are omitted, as it is too intricate to elaborate on here and may divert from the main focus of our discussion. For a comprehensive understanding of L_{reg} and L_{dfl}, we recommend referring to the detailed explanations provided in the original documentation by Zheng et al. [52] and Li et al. [22].

$$
\begin{aligned}
L_{cls} = & \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \sum_{cls \in classes} y_n^{cls} log(p_n^{cls}) \\
& + (1 - y_n^{cls}) log(1 - p_n^{cls}), \\
L_{nat} = & \frac{1}{N_{pos}} \sum_{n=1}^{N_{pos}} \sum_{nat \in natures} y_n^{nat} log(p_n^{nat}) \\
& + (1 - y_n^{nat}) log(1 - p_n^{nat}),
\end{aligned}
\tag{6}
$$

where y_n^{cls} and y_n^{nat} are the labeled value of classification and the object nature, p_n^{cls} and p_n^{nat} are the corresponding predictions, N_{pos} is the number of positive anchor. We utilize the same training pipeline as YOLOv10 for training RO-YOLOv10.

5. Evaluation

5.1. Evaluation Metrics

In the experiment, the proposed baseline detectors are evaluated for the performance by using two common metrics, i.e., average precision (AP) and mean average precision (mAP). IOU (Intersection over Union) measures the overlap between the predicted bounding box (bbox) and the ground truth bbox. In object detection tasks, a complete prediction comprises two main components: first, the model must identify specific objects within a given image, and second, it needs to accurately determine their respective locations. Specifically, precision is the proportion of objects predicted by the model that match the real objects, whereas recall measures the proportion of real objects detected by the model. These two measures are combined in mAP, which highlights the significance of properly balancing each during the evaluation process.

Guided by the COCO evaluation [27], three IoU thresholds are used: fixed thresholds at 0.5 and 0.75 and a range threshold from 0.5 to 0.95 with a step size of 0.05. The corresponding average precisions (APs) are evaluated under these IoU thresholds, denoted as AP@0.5, AP@0.75, and AP@[.50:.05:.95], respectively. In the experiment, COCO mAP is employed to evaluate the performance of detectors in detecting reflected objects. Following [37, 17, 47, 23, 53], we use APc, APn, and APcn to represent the precision metrics for predicting the object's category, the object's nature, and their combination, respectively. Additionally, an extra prefix 'm' is added to represent mean AP, i.e., mAP.

APcn represents an evaluation metric that integrates both category and nature characteristics. The output of a standard object detection task consists of three components: bounding box (bbox), identifier (id), and confidence scores. The fusion strategy for these components is as follows:

- **Bounding-box:** Reusing bounding boxes for both category and nature attributes.
- **ID:** Since the model must infer both class and reflected nature, two decoupled heads, for cls and nature, are employed. To ensure consistency, a unified encoding method is required. In the dataset annotation, we assigned the category to the least significant bit and the

Table 1: The evaluation results of the five proposed baseline detectors, i.e., RO-YOLOv8, RO-YOLOv10, RO-RTMDet, RO-YOLOX, and RO-PPYOLOE, on the ROD dataset. It is important to note that APc, APn, and APcn represent the precision metrics for predicting the object's category, the object's nature, and the combination of both.

Method	{APc, APn, APcn}@0.5	{APc, APn, APcn}@0.75	{mAPc, mAPn, mAPcn}
RO-YOLOv8	(0.795,0.729,**0.571**)	(0.741,0.684,0.541)	(**0.683**,0.637,**0.522**)
RO-YOLOv10	(**0.812**,**0.790**,0.570)	(**0.744**,**0.731**,**0.542**)	(0.679,**0.677**,0.515)
RO-RTMDet	(0.720,0.474,0.537)	(0.656,0.438,0.511)	(0.601,0.406,0.480)
RO-YOLOX	(0.754,0.735,0.490)	(0.654,0.638,0.445)	(0.574,0.558,0.378)
RO-PPYOLOE	(0.713,0.565,0.538)	(0.649,0.514,0.512)	(0.598,0.476,0.481)

reflected nature to the next significant bit. Given that there are 10 categories in the ROD dataset, our encoding should be in decimal format. Consequently, the merged identifier can be expressed as ID_c+10ID_n. For instance, if the identified category is 'cup', with ID_c marked as 6, and it is in the reflected nature with ID_n marked as 1, the fused ID_{cn} would be calculated as 6+10x1=16. This indicates that the fused ID_{cn} will only yield the correct value if both the category and reflected nature identifiers are accurate.

- **Scores:** Calculate the geometric mean of the APc and APn scores.

$$\begin{aligned} Bbox_{cn} &= Bbox_c = Bbox_n, \\ ID_{cn} &= ID_c + 10ID_n, \\ Scores_{cn} &= \sqrt{Scores_c \times Scores_n}. \end{aligned} \tag{7}$$

Since the value of ID_{cn} must satisfy the condition that both the category and reflected nature are correct simultaneously, the value of APcn should be lower than that of APc and APn. This relationship is also demonstrated in the subsequent experiments.

5.2. Evaluation Results

Overall performance. We conducted a comprehensive evaluation of the five baseline detectors proposed in this paper—RO-YOLOv8, RO-YOLOv10, RO-RTMDet, RO-YOLOX, and RO-PPYOLOE—on the ROD dataset. All models were trained on Nvidia Tesla P40 GPUs with a batch size of 32 over 300 epochs, ensuring that each model achieved convergence. For instance, RO-PPYOLOE converges at the 30th epoch, while RO-YOLOv10 converges at the 90th epoch. This trend indicates that the average precision (AP) values for each model gradually increase, followed by a slow decline after reaching the convergence epoch, ultimately attaining optimal performance at the point of convergence. Table 1 presents the evaluation results using the three accuracy metrics defined in Section 5.1, namely APc, APn, and APcn. It can be seen that RO-YOLOv10 is the best-performing detector, significantly outperforming the other detectors in all AP metrics.

Additionally, the evaluation results of the five baseline detectors on target categories show that the AP under fixed IoU thresholds (especially 0.5 and 0.75 IoU) and the average AP for categories are both lower than those for target attributes. For all these detectors, the difference between category AP and attribute AP exceeds 2%. Notably, the RO-YOLOX detector demonstrates the largest gap, reaching 20%. This indicates that in the ROD dataset, identifying and locating the objects themselves is more challenging than recognizing their attributes.

It is also worth noting that when traditional object detection is combined with target attribute prediction to form a compound task, known as reflective object detection, the performance of the detectors is lower than when handling each task individually. This suggests that the reflective object detection task proposed in this paper is more challenging than traditional object detection. This performance drop highlights the importance of developing more specialized algorithms and training strategies to better address the complexity of this compound task.

Performance on per Category. To get a deeper analysis and understanding of the performance in detecting the nature of objects using our proposed baseline detectors, we further conduct performance evaluations on each category. Table 2 presents the mAP_{cn} of the five detectors evaluated on ROD dataset.

As observed, these five detectors perform best on the chair category, while they perform worst on the carrot category. For the chair category, the mAP_{cn} values all exceed 90%, except for the RO-YOLOX detector. For the banana category, the mAP_{cn} values are all below 80%, as for the carrot category with RO-RTMDet, and RO-PPYOLOE even scoring below 50%. This disparity can be explained by the fact that images typically contain only one chair object, often presented at a standard size. In contrast, images frequently contain a large number of carrot objects, leading to crowding and occlusion. Please refer to Figure 8 for a visual comparison of examples from these five categories. The first row shows qualitative results for the keyboard category from the five detectors. The challenges of detecting keyboards on the screen are exacerbated by their size and the reflective properties of the screen itself. Compared

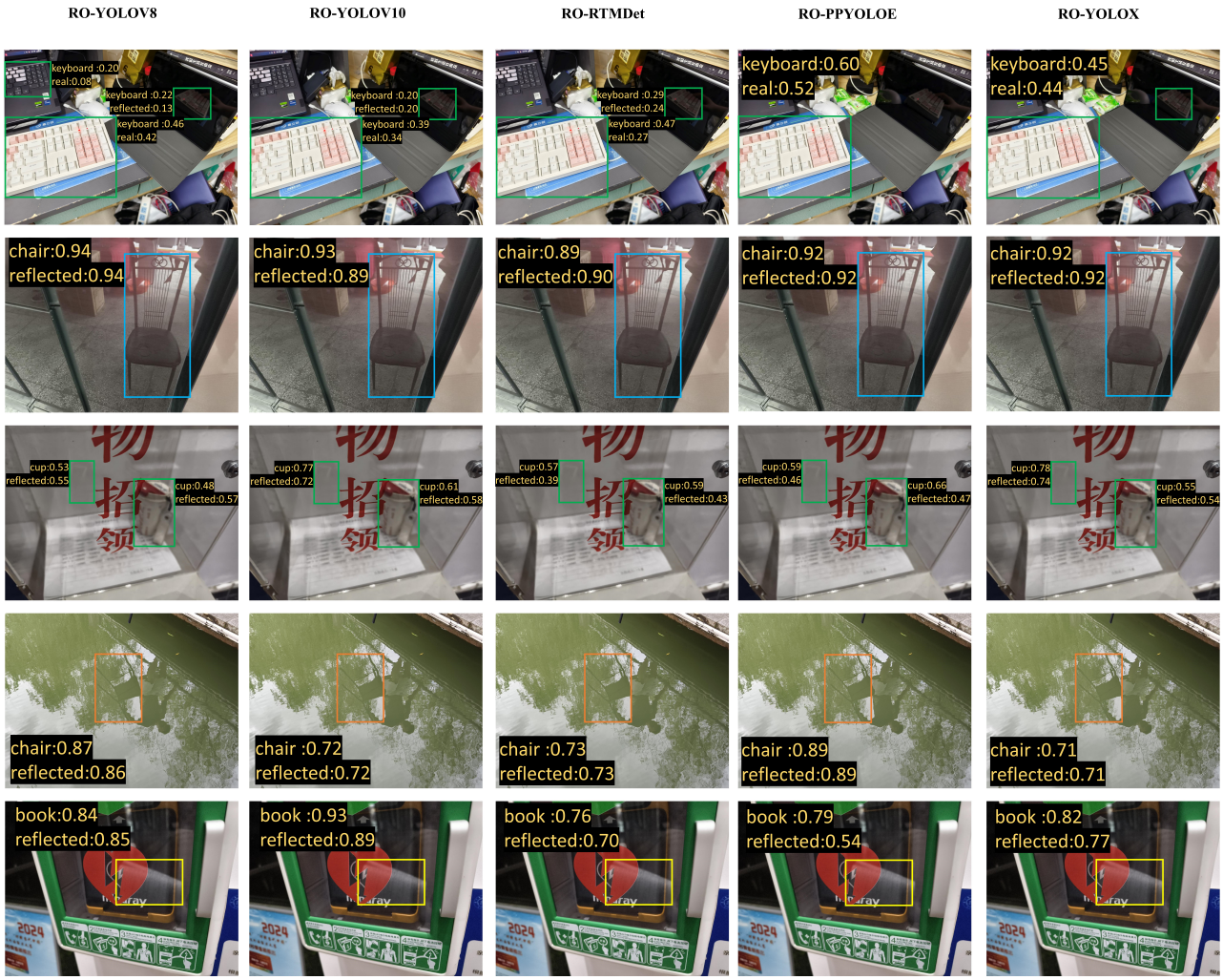


Figure 8: A qualitative comparison of the five detectors on 5 samples from the keyboard, chair, cup and the book category, respectively. Note that all these detectors successfully detect the object but fail to detect the keyboard correctly. The predicted bounding box, object category, object nature, and the corresponding scores have been marked in the images.

Table 2: Comparison of the mAP_{cn} of the five baseline detectors on the ROD dataset. It is important to note that mAP_{cn} is the mAP for prediction of the composite of the object's category and its nature.

	bowl	apple	mouse	keyboard	banana	carrot	cup	orange	chair	book
mAP_{cn}(RO-YOLOv8)	**0.766**	0.672	0.874	0.883	**0.794**	0.673	0.847	0.721	**0.924**	0.792
mAP_{cn}(RO-YOLOv10)	0.764	**0.731**	0.849	0.858	**0.794**	**0.811**	**0.881**	**0.782**	0.915	**0.801**
mAP_{cn}(RO-RTMDet)	0.755	0.637	**0.891**	**0.889**	0.770	0.457	0.821	0.499	0.922	0.557
mAP_{cn}(RO-YOLOX)	0.713	0.556	0.797	0.838	0.729	0.742	0.787	0.763	0.875	0.744
mAP_{cn}(RO-PPYOLOE)	0.755	0.608	0.885	0.879	0.768	0.463	0.832	0.489	0.923	0.827

to a mirror, the screen has a lower reflectance coefficient, which complicates the recognition of keyboards in reflections. This reduced reflectance hampers detectors such as RO-YOLOX and RO-PPYOLOE in accurately identifying the reflective features of keyboards. Additionally, the similar color and appearance between carrots and oranges further confuse detectors like RO-PPYOLOE, leading to misclassification in object categorization. In contrast, these detectors excel in detecting chairs due to the high reflectance of mirrors and the absence of cluttered backgrounds. Furthermore, the experimental results in Table 2 also indicate that the performance of the same detector varies across dif-

Figure 9: A qualitative evaluation was conducted on 12 samples from ROD dataset. The first two rows display examples accurately predicting the nature of objects using RO-YOLOv10 and RO-RTMDet detectors, while last row shows error detection results generated by these two detectors. Note that the predicted bounding box, object category, object nature, and the corresponding scores have been marked in the images.

Table 3: The ablation study of the RO-YOLOv10 model is conducted on ROD dataset using various weighting coefficients.

λ	{APc, APn, APcn}@0.5	{APc, APn, APcn}@0.75	{mAPc, mAPn, mAPcn}
0.2	(**0.821**, 0.778, 0.565)	(**0.760**, 0.729, 0.536)	(**0.694**, 0.678, 0.515)
0.4	(0.815, 0.786, 0.569)	(0.754, 0.732, 0.537)	(0.688, 0.680, 0.518)
0.6	(0.819, **0.796**, 0.579)	(0.754, **0.742**, 0.548)	(0.689, **0.692**, 0.525)
0.8	(0.819, 0.791, **0.584**)	(0.754, 0.739, **0.554**)	(0.690, 0.687, **0.528**)
1.0	(0.812, 0.790, 0.570)	(0.744, 0.731, 0.541)	(0.679, 0.677, 0.515)
1.2	(0.793, 0.776, 0.552)	(0.729, 0.720, 0.519)	(0.666, 0.666, 0.493)
1.4	(0.808, 0.782, 0.559)	(0.744, 0.723, 0.529)	(0.678, 0.670, 0.501)
1.6	(0.800, 0.774, 0.556)	(0.737, 0.715, 0.525)	(0.670, 0.659, 0.498)
1.8	(0.797, 0.774, 0.545)	(0.732, 0.715, 0.516)	(0.665, 0.660, 0.487)
2.0	(0.791, 0.770, 0.549)	(0.725, 0.714, 0.517)	(0.662, 0.658, 0.489)

ferent categories. This disparity may be attributed to the inherent differences in object characteristics, such as size, shape, texture, and background environment in the images, as well as the imbalance in category distribution. These results underscore the importance of considering object-specific challenges when detecting reflective objects.

Qualitative Evaluation. Given the potential for overwhelming viewers with too many methods in a single image, Fig. 9 presents qualitative detection results from just the RO-YOLOv10 and RO-RTMDet detectors. The first two rows display eight correctly predicted samples, while the third row shows examples where the detectors inaccurately predicted the object's nature. In these cases, reflected objects might blend into low-light backgrounds or lack distinct texture features (i.e., the first and second samples), or their mirrored background may resemble the real background (i.e., the third and fourth samples), leading to missed or inaccurate detections. This evaluation highlights that in complex scenes, the detectors are prone to struggle with accurately identifying the nature of objects.

In view of the shortcomings of the model itself in processing complex reflective surface information, more powerful feature extraction modules, such as the module based on attention mechanism, can be considered to make the model more focused on the key features of the reflected object and ignore the interference information brought by the reflected surface. For object occlusion, we can learn from some advanced algorithms in the field of object detection, such as multi-view information fusion or occlusion reasoning mechanism based on deep learning, to improve the detection performance of the model in occlusion scenes. At the same time, how to improve the training strategy of the model is discussed, such as adding more adversarial training samples, simulating various complex reflection and occlusion scenes, so that the model can learn more robust feature representation.

5.3. Ablation Study

Since RO-YOLOv10 obtained the best mAP value in the comparison experiment and the model converged at the 90th epoch, the ablation experiment would keep the batch size and other hyperparameters consistent and take the 90th epoch in each experiment. We train the RO-YOLOv10 model on ROD dataset using different weighting coefficients, i.e., λ in Eq. (5), which varies from 0.2 to 2.0 in steps of 0.2, in order to study the impact of the coefficient for weighting the loss of predicting the nature of objects. This experiment aims to determine how different weightings influence the model's ability to balance the two tasks: detecting the objects and predicting their nature. By adjusting λ, we can observe how the model prioritizes the nature prediction task relative to the conventional object detection task. Table 3 presents the experimental results for the mAP and the AP at fixed IoUs (0.5 and 0.75). RO-YOLOv10 achieves some of the best AP values when λ is set to 0.6 or 0.8. Although APcreaches its maximum value at λ = 0.2, its corresponding APcn is significantly low, resulting in poor performance. A notable discrepancy between APc and APn is evident.

Overall, as λ varies from 0.2 to 1.6, APc initially increases before decreasing, while APn consistently rises, reaching optimal values at 0.6 and 0.8, respectively. A compromise is achieved at λ =0.8, which serves as the default setting, where APcn attains its highest value. Specifically, the maximum mAPc of 0.694 occurs at λ = 0.2, the maximum mAPn of 0.692 occurs at λ = 0.6, and the maximum mAPcn of 0.528 occurs at λ = 0.8. The results suggest that there may be a counteracting impact between object localization and prediction of objects' nature when these two tasks are done concurrently as a composite task. More effective methods for mitigating this counteracting effect are needed.

6. Conclusions

In this paper, we investigated the underexplored challenge of reflective object detection and introduced the Reflective Object Detection (ROD) dataset, an extensive benchmark specifically designed for this task. ROD dataset includes 10 categories, 6 reflective surfaces and 23,520 images of real or reflected objects in various backgrounds, accompanied by standard annotations of bounding boxes and the nature of the objects (real or reflected), distinguishing it from traditional object detection benchmarks. In addition to introducing ROD dataset, we adapted five state-of-the-art object detection models to this challenging task and presented baseline results. The experimental findings reveal the limitations of current methods when applied to reflected object detection, underscoring the necessity for specialized approaches. By releasing ROD dataset, we aim to foster and advance future research in detecting reflected objects. This dataset provides a valuable resource for developing and evaluating new methods, ultimately contributing to improved performance in applications such as surveillance, autonomous driving, service robots, smart homes, and medical imaging. While the current ROD dataset encompasses diverse reflective surfaces and object categories, it lacks scene complexity and environmental variety. Baseline models exhibit suboptimal performance in handling intricate reflections. Future research should prioritize dataset expansion through multi-environment image collection under varying illumination, coupled with advanced deep learning frameworks integrating reflection-aware mechanisms with complementary techniques (e.g., semantic segmentation, object tracking) to address complex vision tasks.

Acknowledgement

The authors would like to acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 61573183). This research was also made possible through collaborative efforts and resources from the institutions involved. Special thanks are extended to the reviewers for their insightful feedback and suggestions.

References

[1] R. Abiko and M. Ikehara. Single image reflection removal based on gan with gradient constraint. *IEEE Access*, 7:148790–148799, 2019. 5

[2] A. B. Amjoud and M. AMROUCH. Object detection using deep learning, cnns and vision transformers: A review. *IEEE Access*, 11:35479–35516, 2023. 1

[3] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk. Single image reflection suppression. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1752–1760, 2017. 5

[4] R. Bajcsy, S. W. Lee, and A. Leonardis. Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation. *International Journal of Computer Vision*, 17:241–272, 1996. 5

[5] Z. Cai and N. Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3

[6] A. Chakraborty, M. Islam, F. Shahriyar, S. Islam, H. U. Zaman, and M. Hasan. Smart home system: a comprehensive review. *Journal of Electrical and Computer Engineering*, 2023(1):7616683, 2023. 2

[7] D. Damodaran, S. Mozaffari, S. Alirezaee, and M. J. Ahamed. Experimental analysis of the behavior of mirror-like objects in lidar-based robot navigation. *Applied Sciences*, 2023. 2

[8] A. DelPozo and S. Savarese. Detecting specular surfaces on natural images. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 5

[9] J. Díaz, E. Ros, S. Mota, G. Botella, A. Cañas, and S. Sabatini. Optical flow for cars overtaking monitor: the rear mirror blind spot problem. *Ecovision (European research project)*, 2003. 2

[10] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018. 3

[11] M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge, 2012. Accessed: 2024-06-27. 3

[12] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 5, 6

[13] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6

[14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 3

[15] R. Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[17] Y. Guo, Y. Chen, J. Deng, S. Li, and H. Zhou. Identity-preserved human posture detection in infrared thermal images: A benchmark. *Sensors*, 23(1):92, 2022. 9

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[19] G.-P. Ji, K. Fu, Z. Wu, D.-P. Fan, J. Shen, and L. Shao. Full-duplex strategy for video object segmentation. *Computational Visual Media*, 9:155–175, 2021. 2, 5

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 3

[21] D. Li, H. Hagura, T. Miyabashira, Y. Kawai, and S. Ono. Traffic mirror detection and annotation methods from street images of open data for preventing accidents at intersections by alert. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3256–3262. IEEE, 2023. 2

[22] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *ArXiv*, abs/2006.04388, 2020. 8, 9

[23] Y. Li, Y. Wu, X. Chen, H. Chen, D. Kong, H. Tang, and S. Li. Beyond human detection: A benchmark for detecting common human posture. *Sensors*, 23(19):8061, 2023. 9

[24] J. Lin, X. Y. Tan, and R. W. H. Lau. Learning to detect mirrors from videos via dual correspondences. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9109–9118, 2023. 1, 2

[25] J. Lin, G. Wang, and R. W. H. Lau. Progressive mirror detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3694–3702, 2020. 2, 5

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[27] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3, 5, 6, 9

[28] F. Liu, Y. Liu, J. Lin, K. Xu, and R. W. H. Lau. Multi-view dynamic reflection prior for video glass surface detection. In *AAAI Conference on Artificial Intelligence*, 2024. 2, 5

[29] L. Liu, W. Ouyang, X. Wang, P. W. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128:261 – 318, 2018. 1

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. 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*, pages 21–37. Springer, 2016. 3

[31] V. N. Lu, J. Wirtz, W. H. Kunz, S. Paluch, T. Gruber, A. Martins, and P. G. Patterson. Service robots, customers and service employees: what can we learn from the academic literature and where are the gaps? *Journal of Service Theory and Practice*, 2020. 2

[32] C. Lyu, W. Zhang, H. Huang, Y. Zhou, Y. Wang, Y. Liu, S. Zhang, and K. Chen. Rtmdet: An empirical study of designing real-time object detectors. *arXiv preprint arXiv:2212.07784*, 2022. 6, 7

[33] D. Marikyan, S. Papagiannidis, and E. Alamanos. A systematic review of the smart home literature: A user perspective. *Technological Forecasting and Social Change*, 138:139–154, 2019. 2

[34] I. Noriaki, O. Shintaro, S. Yoshihiro, O. Kazuya, and R. Grewe. Collision risk prediction utilizing road safety mirrors at blind intersections. In *27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration*, number 23-0164, 2023. 2

[35] D. Owen and P.-L. Chang. Detecting reflections by combining semantic and instance segmentation. *ArXiv*, abs/1904.13273, 2019. 1, 2, 5

[36] D. Park and Y. H. Park. Identifying reflected images from object detector in indoor environment utilizing depth information. *IEEE Robotics and Automation Letters*, 6:635–642, 2021. 2

[37] L. Qin, H. Zhou, Z. Wang, J. Deng, Y. Liao, and S. Li. Detection beyond what and where: a benchmark for detecting occlusion state. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 464–476. Springer, 2022. 9

[38] B. R. Ray, M. Aalsma, N. D. Zaller, E. B. Comartin, and E. Sightes. The perpetual blind spot in public health surveillance. *Journal of correctional health care : the official journal of the National Commission on Correctional Health Care*, 2022. 2

[39] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[40] Y. Shen and W. Q. Yan. Blind spot monitoring using deep learning. *2018 International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pages 1–5, 2018. 2

[41] C. Singhal and S. Barick. Ecms: Energy-efficient collaborative multi-uav surveillance system for inaccessible regions. *IEEE Access*, 10:95876–95891, 2022. 2

[42] B. K. Sovacool and D. D. F. Del Rio. Smart home technologies in europe: A critical review of concepts, benefits, risks and policies. *Renewable and sustainable energy reviews*, 120:109663, 2020. 2

[43] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020. 3

[44] Ultralytics. Yolov8: Real-time object detection and image segmentation, 2023. Accessed: 2024-06-27. 3, 6, 8

[45] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, and G. Ding. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*, 2024. 3, 6, 8

[46] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 390–391, 2020. 7, 8

[47] Y. Wu, H. Ye, Y. Yang, Z. Wang, and S. Li. Liquid content detection in transparent containers: A benchmark. *Sensors*, 23(15):6656, 2023. 9

[48] S. Xu, X. Wang, W. Lv, Q. Chang, C. Cui, K. Deng, G. Wang, Q. Dang, S. Wei, Y. Du, et al. Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*, 2022. 6

[49] X. Yang, H. Mei, K. Xu, X. Wei, B. Yin, and R. W. H. Lau. Where is my mirror? *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8808–8817, 2019. 1, 5

[50] C. Zhang, F. Steinhauser, G. Hinz, and A. Knoll. Traffic mirror-aware pomdp behavior planning for autonomous urban driving. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 323–330. IEEE, 2022. 2

[51] Z.-Q. Zhao, P. Zheng, S. tao Xu, and X. Wu. Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 30:3212–3232, 2018. 1

[52] Z. Zheng, P. Wang, D. Ren, W. Liu, R. Ye, Q. Hu, and W. Zuo. Enhancing geometric factors in model learning and inference for object detection and instance segmentation. *IEEE Transactions on Cybernetics*, 52:8574–8586, 2020. 8, 9

[53] H. Zhou, Y. Wu, J. Li, L. Pan, H. Ye, and S. Li. Beyond animal detection: a benchmark for detecting animal age group. In *Fifth International Conference on Artificial Intelligence and Computer Science (AICS 2023)*, volume 12803, pages 506–515. SPIE, 2023. 9

[54] Z. Zou, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111:257–276, 2019. 1