Exploiting Polarized Material Cues for Robust Car Detection

Wen Dong[1], **Haiyang Mei**[1,2], **Ziqi Wei**[3,4,*], **Ao Jin**[1], **Sen Qiu**[1], **Qiang Zhang**[1], **Xin Yang**[1,*]

[1]Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology
[2]Show Lab, National University of Singapore
[3]Institute of Automation, Chinese Academy of Sciences
[4]State Key Laboratory of Structural Analysis for Industrial Equipment, Dalian University of Technology
{dongwen, dllsja}@mail.dlut.edu.cn, haiyang.mei@outlook.com, ziqi.wei@ia.ac.cn, {qiu, zhangq, xinyang}@dlut.edu.cn

Abstract

Car detection is an important task that serves as a crucial prerequisite for many automated driving functions. The large variations in lighting/weather conditions and vehicle densities of the scenes pose significant challenges to existing car detection algorithms to meet the highly accurate perception demand for safety, due to the unstable/limited color information, which impedes the extraction of meaningful/discriminative features of cars. In this work, we present a novel learning-based car detection method that leverages trichromatic linear polarization as an additional cue to disambiguate such challenging cases. A key observation is that polarization, characteristic of the light wave, can robustly describe intrinsic physical properties of the scene objects in various imaging conditions and is strongly linked to the nature of materials for cars (*e.g.*, metal and glass) and their surrounding environment (*e.g.*, soil and trees), thereby providing ***reliable*** and ***discriminative*** features for robust car detection in challenging scenes. To exploit polarization cues, we first construct a pixel-aligned RGB-Polarization car detection dataset, which we subsequently employ to train a novel multimodal fusion network. Our car detection network dynamically integrates RGB and polarization features in a request-and-complement manner and can explore the intrinsic material properties of cars across all learning samples. We extensively validate our method and demonstrate that it outperforms state-of-the-art detection methods. Experimental results show that polarization is a powerful cue for car detection. Our code is available at https://github.com/wind1117/AAAI24-PCDNet.

Introduction

Autonomous driving and advanced driving assistance system (ADAS) rely on highly accurate road scene analysis. As a fundamental step to achieving a reliable road scene understanding, object detection has received great attention in recent years and has been significantly boosted with the development of deep neural networks (DNNs) (Lin et al. 2017; Redmon and Farhadi 2018; Sun et al. 2021). In practical road scenes, cars appear to be one of the most frequently observed yet dangerous objects, and car detection is still a challenging problem due to the large structural and appearance variations of cars in different scenes.

Although existing state-of-the-art methods explore rich contexts in multiple modalities including RGB, LiDAR, and infrared to improve the detection accuracy, these methods typically assume the imaging quality is optimal. When it comes to adverse conditions such as low light, rain, and fog, a significant accuracy drop would occur due to the poor and limited sensing scene information fed into the algorithms. For example, an RGB camera may fail to capture important visual cues under low-light conditions (Song et al. 2019; Arora et al. 2022), a LiDAR sensor may struggle to distinguish targets in complex environments due to its limited range and resolution (Qian et al. 2021; Chen et al. 2020), and, similarly, an infrared sensor may produce blurry images with low contrast when exposed to extreme weather conditions (Du et al. 2021; Sun et al. 2022). Instead, polarization, characteristic of the light wave, can robustly reveal the intrinsic physical properties of cars and their surrounding environment (*e.g.*, the surface geometric structure, roughness, and material) in various view/lighting/weather conditions. This inspires us to exploit the ***reliable*** and ***discriminative*** features provided by polarization to complement traditional RGB features for robust car detection.

Linear polarization cues, described by the angle of polarization (AoLP) and the degree of linear polarization (DoLP), might not be equally obvious/informative over different scenes and image regions, or even confound valid RGB cues. To address these challenges, we design a novel RGB-Polarization Car Detection Network (PCDNet) with RGB intensities, trichromatic AoLP and DoLP as input. PCDNet is built on three key modules: (i) Polarization Integration (PI) module that fuses AoLP and DoLP to generate a comprehensive and semantically meaningful polarization representation; (ii) Material Perception (MP) module to explore the polarization/material properties of cars across different learning samples for enhancing the polarization cues in each scene; and (iii) Cross-Domain Demand Query (CDDQ) module to dynamically integrate the informative polarization cues into RGB features based on the spatial demand map generated from RGB domain.

To train PCDNet, we introduce an RGB-Polarization car detection dataset, dubbed RGBP-Car, which consists of 1,611 RGB images and pixel-aligned trichromatic (*i.e.*, red, green and blue channels) AoLP and DoLP images, as well as corresponding annotated 31,234 bounding boxes of cars. To

*Corresponding authors

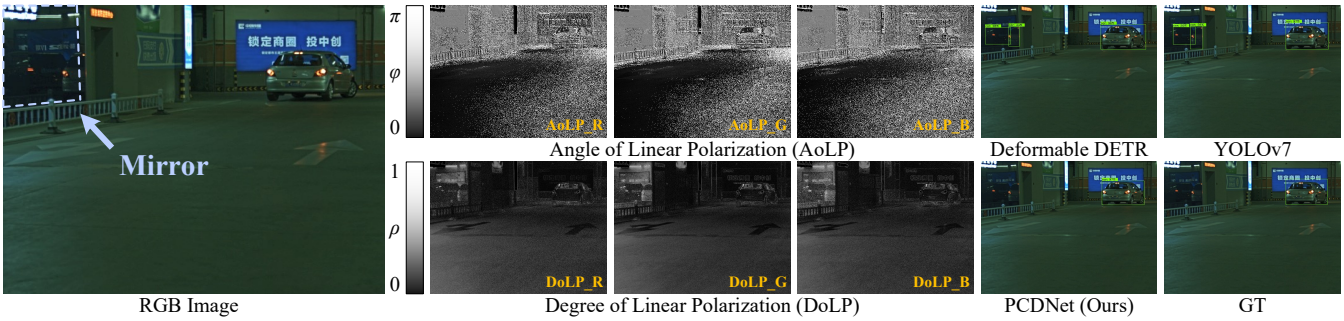


Figure 1: Car detections (indicated by green bounding boxes) obtained with the *RGB-only* methods of Deformable DETR (Zhu et al. 2020) and YOLOv7 (Wang, Bochkovskiy, and Liao 2023) compared to our *RGB-Polarization* Car Detection Network (PCDNet). Both prior methods fail to distinguish mirrored cars from real ones due to the similar visual appearances. In contrast, our method can handle such ambiguity and correctly detect the real car in scenes with the help of intrinsic material properties revealed by the polarization cues.

ensure diversity, the images in RGB-P Car are captured from various real-world traffic scenes with different view/weather/lighting conditions and vehicle densities.

We perform extensive experiments to demonstrate the superiority of our method over competing approaches and show the importance of polarization cues for robust car detection in challenging scenes (*e.g.*, Fig. 1). In summary, our contributions are:

- the first solution to exploit both RGB and trichromatic angle/degree of linear polarization (AoLP/DoLP) cues for robust car detection;
- a new pixel-aligned RGB-P car detection dataset covering challenging dense cars and low light scenarios;
- a novel multimodal fusion network that dynamically integrates RGB and polarization features in a request-and-complement manner;
- a novel polarization cues perception strategy to implicitly explore the intrinsic material properties of cars across the whole learning samples.

Background and Related Work

Polarization has a long research history in computer vision and is widely used in many tasks such as reflection removal (Wieschollek et al. 2018; Lei et al. 2020; Li et al. 2020), surface normal and/or shape estimation (Chen et al. 2017; Kadambi et al. 2015), and semantic segmentation (Mei et al. 2022; Kalra et al. 2020). Light is an electromagnetic wave, with its electric field oscillating perpendicularly to the direction of propagation. Unpolarized light has a randomly fluctuating electric field while polarized light has a biased direction of the electric field. Common light sources like the sun and LED spotlights emit unpolarized light which would become partially/fully polarized light when passing through a linear polarizer, reflecting off certain materials, or undergoing certain types of scattering.

In this work, we focus on the linear polarization measurement captured by the off-the-shelf polarization-array CMOS sensor which can record light intensities in four polarization directions, *i.e.*, I_{0°, I_{45°, I_{90°, and I_{135°, respectively. The polarization state of the light can be described using the Stokes vector $S = [S_0, S_1, S_2, S_3]$, where S_0 stands for the total light intensity, S_1 and S_2 describe the ratio of the $0^\circ/45^\circ$ linear polarization over its perpendicular counterpart, and S_3 is the circular polarization power. The Stokes elements S_0, S_1, S_2 are formally defined as:

$$\begin{aligned} S_0 &= I_{0^\circ} + I_{90^\circ} = I_{45^\circ} + I_{135^\circ}, \\ S_1 &= I_{0^\circ} - I_{90^\circ}, \\ S_2 &= I_{45^\circ} - I_{135^\circ}. \end{aligned} \tag{1}$$

The angle of linear polarization (AoLP) ϕ and the degree of linear polarization (DoLP) ρ are then be calculated via:

$$\phi = \frac{1}{2} arctan(\frac{S_2}{S_1}), \quad \rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}. \tag{2}$$

As shown in Fig. 3, the trees, walls or sky in the background usually exhibit a low linear polarization degree while the glass, rubber and plastic parts of a car are typically with a high linear polarization degree. This observation inspires us to exploit the material cues revealed by polarization for robust car detection.

Object Detection has achieved significant progress with the revolution of deep learning. Many state-of-the-art approaches emerged, including region-based detectors (*e.g.*, Faster R-CNN (Ren et al. 2017) and EfficientDet (Tan, Pang, and Le 2020)), one-stage detectors (*e.g.*, YOLO (Redmon et al. 2016), SSD (Liu et al. 2016), and RetinaNet (Lin et al. 2017)), and anchor-free detectors (*e.g.*, FCOS (Tian et al. 2019)). These methods typically employ advanced network architectures such as ResNet (He et al. 2016), VGG (Simonyan and Zisserman 2014) and EfficientNet (Tan and Le 2019). The trend in object detection has been toward the development of large models. The Vision Transformer (Dosovitskiy et al. 2020), derived from natural language processing (Zaheer et al. 2020), has achieved notable improvements in the field of object detection. For example, some methods such as Swin Transformer (Liu et al. 2021), DETR (Carion et al. 2020), and DAB-DETR (Liu et al. 2022) have achieved remarkable results on benchmark datasets including Microsoft COCO (Lin et al. 2014) and PASCAL VOC

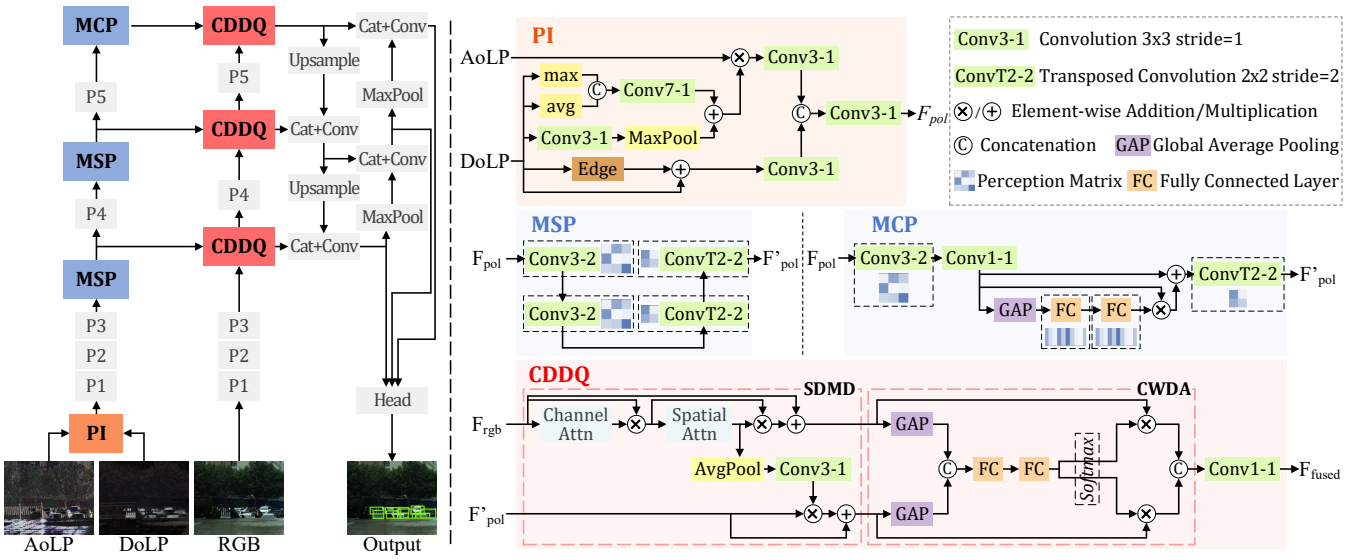


Figure 2: Overview of PCDNet and its three main modules: the Polarization Integration (PI) module, the Material Spatial/Channel Perception (MSP/MCP) module, and the Cross Domain Demand Query (CDDQ) module which contains a Spatial Demand Map Delivery (SDMD) block and a Channel Weight Dynamic Assignment (CWDA) block.

(Everingham et al. 2010). However, such models often exceed the hardware load and detection speed requirements of most restricted terminal devices. Moreover, most of them rely on clear and optimal RGB images which is hard to obtain in challenging scenes. Image enhancement and restoration on the low-quality RGB image will cost extra computing power and time. Our method differs from the above works in that we introduce reliable polarized material cues to complement traditional RGB features and design an RGB-P-based multimodal fusion network for robust detection.

Multimodal Fusion can provide rich contextual information for robust object detection (Valverde, Hurtado, and Valada 2021; Bijelic et al. 2020). Blin et al. employed a simple fusion method by stacking the multimodal data in the channel dimension to replace the original input (Blin et al. 2019). Manjunath et al. and Chen et al. adopted the concatenation (Manjunath et al. 2018) and element-wise addition (Chen et al. 2017) to fusing the low-level features of LiDAR and RGB, respectively. The attention mechanism (Vaswani et al. 2017) is also used to achieve multimodal fusion. HAFNet (Zhang et al. 2020b) developed a cross-modal attention mechanism to perform feature fusion. Mei et al. calculated dynamic fusion weights for RGB and depth, considering the quality of each modality (Mei et al. 2021). Similarly, Ji et al. used global average pooling followed by a fully connected layer to compute the channel attention weight for each modality (Ji et al. 2021). Mei et al. generated spatial attention maps based on both global and local features to guide the multimodal fusion (Mei et al. 2022). Although these methods achieve performance improvement to some extent, they perform information compensation in a passive "post" manner, resulting in the limited robustness of the model in challenging scenes. In this work, we develop a novel polarization material perception scheme to learn the intrinsic material properties of cars and a proactive multimodal fusion strategy to compensate RGB features with informative polarization cues in a "request-and-complement" manner, enhancing the robustness of car detection.

Methodology

RGB images depict objects based on color differences that match human perception. However, objects with similar colors may not have enough color contrast to show their shape. By contrast, polarized light is strongly linked to the material of objects and the orientation of the reflecting surface, enabling it to reveal material properties that make object boundaries visible even when colors are similar (*e.g.*, Fig. 3 (a) and (b)). In spite of that, polarization cues may be weak in certain lighting conditions or viewing angles (*e.g.*, Fig. 3 (c)). Including polarization measurements naively in existing car detection methods may not necessarily yield the expected performance improvement. How to effectively integrate RGB and polarization features is a key challenge to be addressed to achieve robust car detection.

We introduce a novel Polarization Car Detection Network (PCDNet) that is capable of exploring and integrating polarized material cues for robust car detection. As shown in Fig. 2, PCDNet takes as input the RGB intensity, trichromatic AoLP and DoLP, and outputs car detections. The AoLP and DoLP are first integrated into a comprehensive and semantically meaningful polarization feature representation by a Polarization Integration (PI) module for the ease of the following feature extraction and fusion. Then, the RGB and polarization features are separately fed into two branches of CSP-DarkNet (Wang et al. 2020) encoder, each consisting of five stages to extract multi-level contextual features. The Material Perception (MP) module, which aims to extract the intrinsic material properties of cars across the whole learning

samples, is applied to different levels of extracted features. The MP module has a specific design for low-level features (MSP with spatial specialization) and high-level features (MCP with channel specialization), respectively. For multimodal feature fusion, the Cross-Domain Demand Query (CDDQ) module assigns fusion weights adaptively and conducts dynamic fusion in a request-and-complement manner. Finally, based on the fused features, we adopt the anchor-based detection head from YOLOv7 (Wang, Bochkovskiy, and Liao 2023) to generate final classifications and bounding boxes.

Polarization Integration (PI)

AoLP ϕ and DoLP ρ reveal object/scene materials from two different aspects. Polarization Integration (PI) module is designed to combine them into an unified and semantically meaningful polarization representation F_{pol} for the ease of the following feature extraction and fusion. As the captured polarization angle is more likely random at regions with a low polarization degree, PI filters the AoLP measurement based on the DoLP. In addition, PI also extracts and integrates the edge information in DoLP measurement to help the distinction of objects with different materials. Formally,

$$F_{pol} = \vartheta_{3\cdot1}([\vartheta_{3\cdot1}(F_{\phi\rho}), \vartheta_{3\cdot1}(\rho + \mathbb{E}(\rho))]), \quad (3)$$

$$F_{\phi\rho} = \phi \otimes (\vartheta_{3\cdot1}([\hat{\mathbb{A}}(\rho), \hat{\mathbb{M}}(\rho)]) + \sigma(\mathbb{M}(\vartheta_{3\cdot1}(\rho)))), \quad (4)$$

where $\vartheta_{k\cdot s}$ denotes a $k \times k$ convolution with a stride of s, followed by a batch normalization and a SiLU activation function. $[\cdot]$ indicates the concatenation operation over the channel dimension. $\mathbb{E}$ refers to the Scharr edge extractor. $\otimes$ is the element-wise multiplication. $\hat{\mathbb{A}}$ and $\hat{\mathbb{M}}$ are the average and max pooling in the channel dimension, respectively. $\mathbb{M}$ is the max pooling in the spatial dimension with a kernel size of 5. And σ is the Sigmoid activation.

Material Perception (MP)

Despite the cars in different scenarios may have diverse visual appearances such as different colors and textures, they are typically share similar materials including glass, rubber, metal, etc. Fortunately, polarization can robustly reveal the intrinsic physical properties of these materials. Inspired by this, we design the Material Perception (MP) strategy to explore the discriminative and invariant material features of cars across different scenarios. Considering the different characteristics of different levels of features, *i.e.*, low-level features have larger spatial sizes and keep rich and detailed low-level information while high-level features contain more semantic cues distributed in more feature channels, MP is instantialized as Material Spatial/Channel Perception (MSP/MCP) modules for the low-/high-level polarization features, respectively. Formally, MSP/MCP can be described as:

$$MSP(F) = \varrho_{2\cdot2}(\varrho_{2\cdot2}(\vartheta_{3\cdot2}(\vartheta_{3\cdot2}(F)))), \quad (5)$$

$$MCP(F) = \varrho_{2\cdot2}(\mathcal{M}((\vartheta_{1\cdot1}(\vartheta_{3\cdot2}(F))))), \quad (6)$$

$$\mathcal{M}(x) = x + x \otimes \sigma(m_2(m_1(\bar{\mathbb{A}}(x)))), \quad (7)$$

where $\varrho_{k\cdot s}$ denotes a $k \times k$ transposed convolution with a stride of s, followed by a batch normalization and a SiLU activation function. $\mathcal{M}(\cdot)$ is the perception scheme with two independent perception matrices in fully connected layers named m_1 and m_2. And $\bar{\mathbb{A}}$ is global average pooling.

Datasets	Pol.	Pixel align	Images Train / Test	Cars Train / Test
PolarLITIS	Mono	×	2569 1640 / 929	17428 6061 / 11367
RGBP-Car (Ours)	Tri	✓	2601 1611 / 990	31234 19582 / 11652

Table 1: Comparison of existing car detection datasets with polarization measurements.

Cross Domain Demand Query (CDDQ)

Polarization and RGB features are different types of representations of the scenes and simply combing them may dilute the useful clues of the cars originally presented in the individual modality or amplify the background interference. We address this issue by introducing the Cross-Domain Demand Query (CDDQ) module for effective multimodal feature fusion, taking into account both the context and quality of each modality feature. CDDQ takes as input the RGB features F_{rgb} and polarization representations F'_{pol}. It first utilizes a Spatial Demand Map Delivery (SDMD) block to generate enhanced RGB features F^*_{rgb} and distill informative and required polarization features F^*_{pol} and then obtains fused features F_{fused} through a Channel Weight Dynamic Assignment (CWDA) block:

$$\begin{aligned} F^*_{rgb} &= F_{rgb} + F''_{rgb} = F_{rgb} + \mu \otimes F'_{rgb} \\ &= F_{rgb} + \mu \otimes \eta \otimes F_{rgb}, \end{aligned} \quad (8)$$

$$\eta = \sigma(\vartheta_{1\cdot1}(\bar{\mathbb{M}}(F_{rgb})) + \vartheta_{1\cdot1}(\bar{\mathbb{A}}(F_{rgb}))), \quad (9)$$

$$\mu = \sigma(\vartheta_{7\cdot1}([\hat{\mathbb{A}}(F'_{rgb}), \hat{\mathbb{M}}(F'_{rgb})])), \quad (10)$$

$$F^*_{pol} = F'_{pol} + \vartheta_{3\cdot1}(\mathbb{A}(\mu)) \otimes F'_{pol}, \quad (11)$$

$$F_{fused} = \vartheta_{1\cdot1}([\alpha \times F^*_{rgb}, \beta \times F^*_{pol}]), \quad (12)$$

$$\alpha, \beta = \delta(\langle\sigma(fc(si(fc([\bar{\mathbb{A}}(F^*_{rgb}), \bar{\mathbb{A}}(F^*_{pol})])))))\rangle), \quad (13)$$

where η is the channel attention vector. μ is the spatial attention map which also serves as the guidance of the informative polarization cues request process. $\mathbb{A}$ is an average pooling with a kernel size of 3. $\mathbb{M}$/$\bar{\mathbb{A}}$ denotes global max/average pooling. α and β are dynamic fusion weights assigned to the RGB and polarization features, respectively, with a constraint of being non-negative and summing up to 1 for each channel position. fc is the fully connected layer and $\langle\cdot\rangle$ is the split operation over the channel dimension. si and δ are the SiLU and Softmax activation functions, respectively.

RGB-P Car Detection Dataset

We construct the first pixel-aligned RGB-polarization car detection dataset called RGBP-Car with trichromatic polarization measurements. We record cars in diverse traffic scenes using FLIR-Blackfly-S, a polarized color camera that simultaneously obtain pixel-aligned polarization measurements in four linear polarization directions (0°, 45°, 90°,

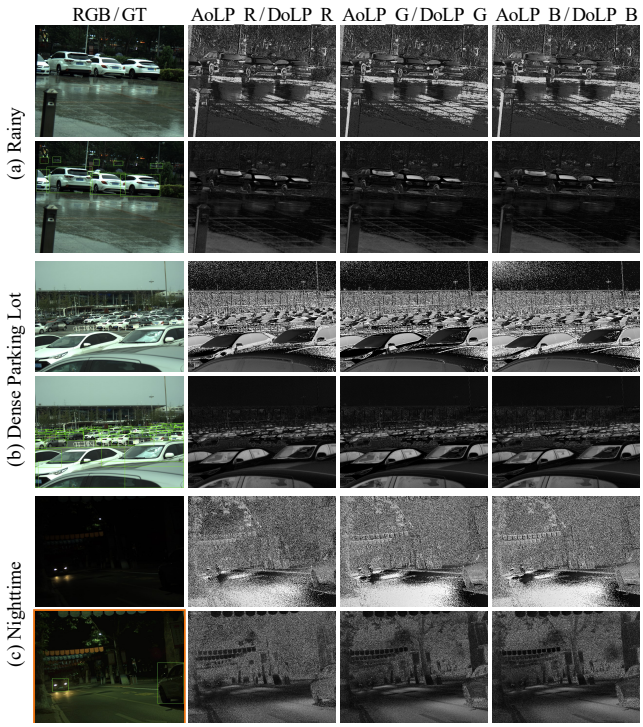
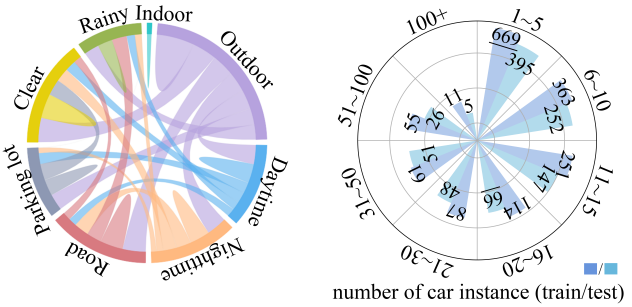


Figure 3: RGBP-Car Examples. The first column displays the RGB intensity (top) and the corresponding annotation (bottom). The next three columns show the AoLP (top) and DoLP (bottom) measurements for the red, green, and blue channels, respectively. From top to bottom are scenes of stopped cars in a rainy parking lot, dense cars in an outdoor parking lot, and driving cars on a clear night road, respectively. (The low-light RGB image is enhanced by ZeroDCE (Guo et al. 2020) (with orange frame) for visualization.)



(a) relationship among scenes (b) car instance ln distribution

Figure 4: The images in our RGB-P Car dataset vary in terms of (a) scenarios and (b) the number of car instances.

Methods	Pub'Year	Backbone	AP	AP50	AP75
Faster R-CNN$^{\ddagger\triangle}$	NeurIPS'15	Res50	44.8	75.4	45.4
SSD$^{\dagger\circ}$	ECCV'16	VGG16	25.5	52.6	22.6
Cascade R-CNN$^{\ddagger\triangle}$	CVPR'18	Res50	45.8	73.2	47.8
CornerNet$^{\dagger\circ}$	ECCV'18	Res50	19.8	47.4	29.6
P-SSD I$^{\star\dagger\circ}$	ITSC'19	VGG16	25.9	53.1	22.7
P-SSD S$^{\star\dagger\circ}$	ITSC'19	VGG16	23.0	48.9	20.1
FCOS$^{\dagger\circ}$	ICCV'19	Res50	23.1	50.9	18.4
DH R-CNN$^{\ddagger\triangle}$	CVPR'20	Res50	32.7	65.3	28.2
Dynamic R-CNN$^{\ddagger\triangle}$	ECCV'20	Res50	46.2	74.2	48.0
EfficientDet$^{\ddagger\triangle}$	CVPR'20	D3	45.3	73.0	46.3
VarifocalNet$^{\dagger\circ}$	CVPR'21	Res50	44.2	73.5	44.4
D-DETR$^{\dagger\circ}$	ICLR'21	Res50	43.8	74.9	44.3
DDOD$^{\dagger\circ}$	MM'21	Res50	43.5	73.0	43.3
TOOD$^{\dagger\triangle}$	ICCV'21	Res50	44.3	74.3	44.6
YOLOX$^{\dagger\circ}$	arXiv'21	YOLOX-l	54.3	82.5	56.7
YOLOv7$^{\dagger\triangle}$	CVPR'23	Dark53	57.6	84.3	60.3
RTMDet$^{\dagger\circ}$	arXiv'22	RTMDet-l	53.9	81.4	56.7
DINO$^{\dagger\circ\S}$	ICLR'22	Res50	52.7	81.8	54.8
YOLOv8$^{\dagger\circ}$	-'23	YOLOv8-l	56.8	83.6	59.0
PCDNet*	**Ours**	Dark53	**58.5**	**85.2**	**61.5**

Table 2: Quantitative comparison against state-of-the-art polarization-based detectors ($\star$), single-stage detectors ($\dagger$), two-stage detectors ($\ddagger$), anchor-based detectors ($\triangle$), anchor-free detectors ($\circ$), and self-supervised method ($\S$).

and 135°) for each color channel (*i.e.*, R, G, and B). RGBP-Car contains 2601 RGB, AoLP, and DoLP image triplets. Each image has manually labeled bounding boxes indicating the position and size of each car. To ensure the diversity and challenge of our dataset, we take the RGB-P images under different weather conditions (clear and rainy), different lighting conditions (daytime and nighttime), different driving environments (indoor, outdoor, road and parking lot), and different car densities. Fig. 3 gives representative examples and Fig. 4 analyzes (a) the relationship among different scenes and (b) the density distribution of car instances. Tab. 1 further shows the superiority of our RGBP-Car over existing car detection datasets with polarization measurements.

Assessment

Experimental Setup

We implement our PCDNet in PyTorch (Paszke et al. 2019) and train it for 300 epochs with the batch size of 32 on two NVIDIA GeForce RTX 3090 GPUs. We use stochastic gradient descent (SGD) (Amari 1993) with a momentum of 0.937 and a weight decay of 5×10^{-4} during training. The initial learning rate is set to 0.01 and decayed to 0.001 using a cosine annealing schedule. We initialize PCDNet randomly and load the weights of CSPDarknet53 (Wang et al. 2020) pre-trained on ImageNet (Deng et al. 2009) for the encoder part. To increase the diversity and complexity of the training samples, we apply data augmentations including random cropping, random flipping, and mosaic (Redmon and Farhadi 2018). We use the evaluation metrics of Microsoft COCO (Lin et al. 2014) for validation.

Qualitative and Quantitative Evaluation

We extensively compare our PCDNet with 19 state-of-the-art methods by retraining and testing all methods on the RGB-P Car dataset using their original settings. The compared methods include two-stage detectors such as EfficientDet (Tan, Pang, and Le 2020) and the R-CNN family (Ren

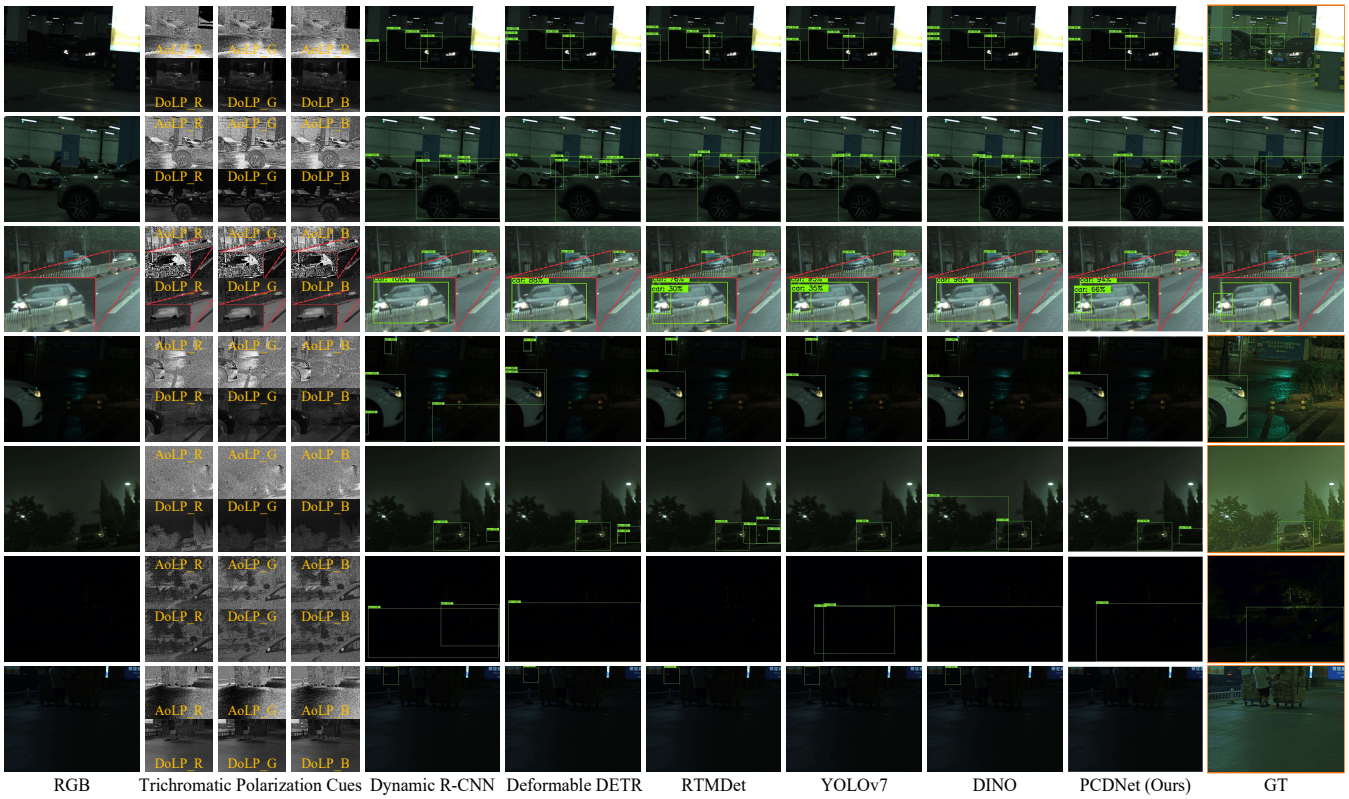


Figure 5: Qualitative comparison of PCDNet against state-of-the-art detectors retrained on RGB-P Car dataset.

et al. 2017; Cai and Vasconcelos 2019; Zhang et al. 2020a), and one-stage detectors such as SSD (Liu et al. 2016), and YOLO family (Ge et al. 2021; Wang, Bochkovskiy, and Liao 2023; Jocher 2023). These methods also comprise anchor-based methods such as the R-CNN family and YOLOv7 (Wang, Bochkovskiy, and Liao 2023), and anchor-free methods such as CornerNet (Law and Deng 2018), VarifocalNet (Zhang et al. 2021), and YOLOv8 (Jocher 2023). Some detectors use traditional convolutional networks such as FCOS (Tian et al. 2019) and RTMDet (Lyu et al. 2022) while others use transformer structures, such as DeformableDETR (Zhu et al. 2020) and DINO (Zhang et al. 2022) that employs self-supervised learning. We also include the P-SSD (Blin et al. 2019) that utilizes polarization information. The quantitative evaluation results are reported in Tab. 2. We can see that our method outperforms all competing state-of-the-art methods.

Fig. 5 further qualitatively demonstrates the benefits of our method: a) in poorly lit indoor parking lots, distinguishing black cars behind pillars is extremely challenging (the first two rows). The compared methods tend to conflate the shadow and the black car (*i.e.*, merging cars on either side of the pillar into a single entity or treating partial views of the car as one object) while our PCDNet can handle such ambiguities; b) in the third example, all methods except our PCDNet fail to detect a partially visible car obstructed by another car or misplace it with the previous car; c) in the fourth example, RGB-based methods wrongly identify distant pedestrians as cars, but our PCDNet method can effectively eliminate such interference with the help of polarization cues; d) the fifth and sixth examples depict black cars in an outdoor parking lot at night which are very hard to be distinguished in the RGB image. Despite the enhancement through ZeroDCE (Guo et al. 2020), the sixth example remains unclear. By contrast, polarization imaging is robust to low light conditions, enabling our robust car detector PCDNet; and e) the last row shows a virtual car reflected in a mirror located at the upper-left corner of the image. The mirrored virtual car and the rest of the mirror regions exhibit similar and smooth AoLP, providing useful cues for PCDNet to recognize this region as background.

Ablation Study

Impact of Spectral Intensity and Polarization Cues. We conduct a series of ablation experiments to demonstrate the effects of spectral intensity and polarization cues on car detection (Tab. 3). The results show that: a) combining different forms of polarization cues with RGB as the input of PCDNet can improve the car detection accuracy (C, D, F, G, K and L are higher than B); b) DoLP cues have a greater impact than AoLP cues (D, J and L are better than C, I and K, respectively); c) stacking AoLP and DoLP on RGB in the channel dimension does not boost performance (E is slightly lower than B), possibly because the characteristic gap between different modalities hinders effective features extrac-

	PCDNet Input	AP	AP50	AP75
A	RGB, AoLP and DoLP (original)	58.5	85.2	61.5
B	RGB only	57.6	84.3	60.2
C	RGB and AoLP	58.0	84.6	60.7
D	RGB and DoLP	58.3	85.4	61.1
E	stacked RGB, AoLP and DoLP	57.5	84.3	59.9
F	RGB and stacked I	58.0	84.1	61.0
G	RGB and stacked S	57.8	84.8	60.4
H	Gray only	57.4	84.3	60.0
I	Gray and mono AoLP	57.5	84.5	60.5
J	Gray and mono DoLP	57.6	84.9	60.1
K	RGB and mono AoLP	57.9	84.6	60.5
L	RGB and mono DoLP	58.2	84.9	60.6
M	Enhanced RGB	57.4	84.0	60.0

Table 3: Quantitative comparisons of ablation with different inputs. "stacked I" denotes the stacked intensity measurements with a linear polarization angle of 0°, 45° and 135° and "stacked S" refers to the stacked Stokes elements S0, S1 and S2 (Blin et al. 2019).

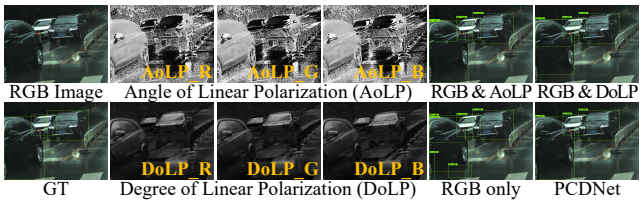


Figure 6: Qualitative comparison of ablation with different inputs. The model with RGB intensity only is susceptible to interference from ghost car caused by water on the road.

tion; d) spectral intensity and polarization are more beneficial than monochromatic intensity and polarization for car detection (comparing paired *B* and *H*, *C* and *K*, *D* and *L*, *I* and *K*, *J* and *L*); e) enhancing RGB image via ZeroDCE (Guo et al. 2020) is less effective than introducing polarization (*M* performs worse than *C*-*G*, *K* and *L*). Fig. 6 provides visual support for these observations.

Influence of PCDNet Components. First, we investigate the performance of different strategies for fusing AoLP and DoLP inputs. From Tab. 4(*A*-*D*), we observe that our PI module is more effective than the simple fusion methods including concatenation, addition and element-wise multiplication. Second, by removing MP module 4(*E*) from the original PCDNet (A), the detection performance declines. This demonstrates that exploring the polarized material features of cars across all learning samples is useful. We also explore the influence of applying MSP and MCP on different levels of features. The results in Tab. 4(*A*,*F*-*G*) show that applying MSP on shallower features and MCP on deeper features can yield better performance. Finally, we validate the effectiveness of CDDQ module. Removing the CDDQ module (*I*) from PCDNet (*A*), which causes the feature extraction processes of the RGB and polarization to be independent from each other, leads to the performance drop. We also demonstrate the benefits of the CWDA and SDMD in the CDDQ module by removing either of them (*J* and *K*).

	Ablation	AP	AP50	AP75
A	PCDNet (original)	58.5	85.2	61.5
B	Input RGB and [AoLP DoLP]	58.2	85.4	60.9
C	Input RGB and AoLP+DoLP	58.1	84.8	60.5
D	Input RGB and AoLP*DoLP	58.1	84.8	60.5
E	A *w/o* MP	56.9	84.2	59.2
F	A *w/* M(S-S-S)P	58.2	85.2	60.8
G	A *w/* M(S-C-C)P	58.2	85.0	60.9
H	A *w/* M(C-C-C)P	58.1	85.0	61.1
I	A *w/o* CDDQ	58.0	84.7	60.8
J	A *w/o* SDMD	58.2	85.2	60.8
K	A *w/o* CWDA	58.3	85.1	61.1

Table 4: Quantitative comparisons of ablation with different modules demonstrate that all component of PCDNet contributes to the overall performance. We used sequences of three letters separated by '-' and enclosed in parentheses to represent different combinations of MSP and MCP.

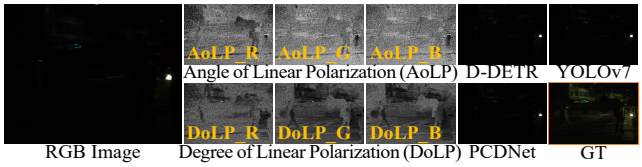


Figure 7: PCDNet has limited ability to handle extreme HDR or heavy motion blur cases.

Limitations

When both the RGB intensity and the polarization measurement yield weak car signals, our method's effectiveness declines. Specifically, in low-light scenarios, when a car approaches on an unlit road, the strong light from its headlights can create a "hole" in both the RGB and polarization and obscure the entire car. We illustrate such an example in Fig. 7 where the extreme HDR and heavy motion blur in the captured image limit its depiction of both RGB and polarization. In these challenging scenarios, prior RGB-based methods and even human vision are powerless.

Conclusion

In this paper, we present PCDNet, the first solution that leverages both RGB intensities and trichromatic angle/degree of linear polarization (AoLP/DoLP) cues for robust car detection in challenging scenarios. PCDNet comprises three key modules: the Polarization Integration (PI) module, the Material Perception (MP) module, and the Cross-Domain Demand Query (CDDQ) module. The PI module fuses AoLP and DoLP to generate a comprehensive polarization representation. The MP module explores the polarization/material properties of cars across different learning samples and the CDDQ module proactively integrates RGB features and polarization representations in a request-and-complement manner. Extensive experiments show that PCDNet outperforms existing methods, especially in challenging scenarios. We also introduced a new pixel-aligned RGB-P car detection dataset covering diverse scenarios, which can promote the use of polarization in relevant visual tasks.

Acknowledgements

This work was supported in part by National Key Research and Development Program of China (2022ZD0210500/2021ZD0112400), the National Natural Science Foundation of China under Grants 62102058/62272081/61972067/62332019, the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01), and the open funding of State Key Laboratory of Structural Analysis for Industrial Equipment.

References

Amari, S.-i. 1993. Backpropagation and stochastic gradient descent method. *Neurocomputing*.

Arora, N.; Kumar, Y.; Karkra, R.; and Kumar, M. 2022. Automatic vehicle detection system in different environment conditions using fast R-CNN. *Multimedia Tools and Applications*.

Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; and Heide, F. 2020. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Blin, R.; Ainouz, S.; Canu, S.; and Meriaudeau, F. 2019. Road scenes analysis in adverse weather conditions by polarization-encoded images and adapted deep learning. In *ITSC*.

Cai, Z.; and Vasconcelos, N. 2019. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Chen, G.; Wang, F.; Qu, S.; Chen, K.; Yu, J.; Liu, X.; Xiong, L.; and Knoll, A. 2020. Pseudo-image and sparse points: Vehicle detection with 2D LiDAR revisited by deep learning-based methods. *IEEE Transactions on Intelligent Transportation Systems (TITS)*.

Chen, X.; Ma, H.; Wan, J.; Li, B.; and Xia, T. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*.

Du, S.; Zhang, P.; Zhang, B.; and Xu, H. 2021. Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4. *IEEE Access*.

Everingham, M.; Van Gool, L.; Williams, C. K.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International journal of computer vision*.

Ge, Z.; Liu, S.; Wang, F.; Li, Z.; and Sun, J. 2021. Yolox: Exceeding yolo series in 2021. *arXiv*.

Guo, C.; Li, C.; Guo, J.; Loy, C. C.; Hou, J.; Kwong, S.; and Cong, R. 2020. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ji, W.; Li, J.; Yu, S.; Zhang, M.; Piao, Y.; Yao, S.; Bi, Q.; Ma, K.; Zheng, Y.; Lu, H.; et al. 2021. Calibrated RGB-D salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Jocher, G. 2023. YOLOv8 - Ultraytics — Revolutionizing the World of Vision AI. https://ultralytics.com/yolov8. Accessed April 23, 2023.

Kadambi, A.; Taamazyan, V.; Shi, B.; and Raskar, R. 2015. Polarized 3d: High-quality depth sensing with polarization cues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Kalra, A.; Taamazyan, V.; Rao, S. K.; Venkataraman, K.; Raskar, R.; and Kadambi, A. 2020. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Law, H.; and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lei, C.; Huang, X.; Zhang, M.; Yan, Q.; Sun, W.; and Chen, Q. 2020. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, R.; Qiu, S.; Zang, G.; and Heidrich, W. 2020. Reflection separation via multi-bounce polarization state tracing. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *International Conference on Learning Representations (ICLR)*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; and Chen, K. 2022. RTMDet: An Empirical Study of Designing Real-Time Object Detectors.

Manjunath, A.; Liu, Y.; Henriques, B.; and Engstle, A. 2018. Radar based object detection and tracking for autonomous driving. In *Proceedings of the IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Mei, H.; Dong, B.; Dong, W.; Yang, J.; Baek, S.-H.; Heide, F.; Peers, P.; Wei, X.; and Yang, X. 2022. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Qian, K.; Zhu, S.; Zhang, X.; and Li, L. E. 2021. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Redmon, J.; Divvala, S.; Girshick, R.; and Farhadi, A. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An Incremental Improvement.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv*.

Song, H.; Liang, H.; Li, H.; Dai, Z.; and Yun, X. 2019. Vision-based vehicle detection and counting system using deep learning in highway scenes. *European Transport Research Review*.

Sun, P.; Zhang, R.; Jiang, Y.; Kong, T.; Xu, C.; Zhan, W.; Tomizuka, M.; Li, L.; Yuan, Z.; Wang, C.; et al. 2021. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sun, Y.; Cao, B.; Zhu, P.; and Hu, Q. 2022. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Tan, M.; Pang, R.; and Le, Q. V. 2020. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Valverde, F. R.; Hurtado, J. V.; and Valada, A. 2021. There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, C.-Y.; Bochkovskiy, A.; and Liao, H.-Y. M. 2023. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, C.-Y.; Liao, H.-Y. M.; Wu, Y.-H.; Chen, P.-Y.; Hsieh, J.-W.; and Yeh, I.-H. 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*.

Wieschollek, P.; Gallo, O.; Gu, J.; and Kautz, J. 2018. Separating reflection and transmission images in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big bird: Transformers for longer sequences. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Zhang, H.; Chang, H.; Ma, B.; Wang, N.; and Chen, X. 2020a. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection.

Zhang, H.; Wang, Y.; Dayoub, F.; and Sunderhauf, N. 2021. Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, P.; Du, P.; Lin, C.; Wang, X.; Li, E.; Xue, Z.; and Bai, X. 2020b. A hybrid attention-aware fusion network (HAFNet) for building extraction from high-resolution imagery and LiDAR data. *Remote Sensing*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv*.