Object Detection Based on Raw Bayer Images

Guoyu Lu

Abstract— Bayer pattern is a widely used Color Filter Array (CFA) for digital image sensors, efficiently capturing different light wavelengths on different pixels without the need for a costly ISP pipeline. The resulting single-channel raw Bayer images offer benefits such as spectral wavelength sensitivity and low time latency. However, object detection based on Bayer images has been underexplored due to challenges in human observation and algorithm design caused by the discontinuous color channels in adjacent pixels. To address this issue, we propose the BayerDetect network, an end-to-end deep object detection framework that aims to achieve fast, accurate, and memory-efficient object detection. Unlike RGB color images, where each pixel encodes spectral context from adjacent pixels during ISP color interpolation, raw Bayer images lack spectral context. To enhance the spectral context, the BayerDetect network introduces a spectral frequency attention block, transforming the raw Bayer image pattern to the frequency domain. In object detection, clear object boundaries are essential for accurate bounding box predictions. To handle the challenges posed by alternating spectral channels and mitigate the influence of discontinuous boundaries, the BayerDetect network incorporates a spatial attention scheme that utilizes deformable convolutional kernels in multiple scales to explore spatial context effectively. The extracted convolutional features are then passed through a sparse set of proposal boxes for detection and classification. We conducted experiments on both public and self-collected raw Bayer images, and the results demonstrate the superb performance of the BayerDetect network in object detection tasks.

I. Introduction

The goal of object detection is to localize a set of objects and classify their categories within an image. Object detection algorithms based on deep convolutional neural networks have been proposed [7] [31] [18] [9] [1]. Though all the methods mentioned above achieve promising performance in diverse aspects, they are all designed for RGB color images, which are processed using image signal processor (ISP) pipelines' inputs. However, the ISP pipeline costs excessive storage and processing time and is also susceptible to damaging or losing the primitive pixel information captured by the raw camera sensor, such as demosaicing [11]. Bayer pattern is a widely used Color Filter Array (CFA) that covers the digital image sensors for capturing the different light wavelengths. The generated raw Bayer images provide a more comprehensive spectral range as RGB color images, but in a single channel, much less than the 3 channels in RGB color images. Though raw Bayer images enjoy the benefits of less memory consumption, high speed due to the potential saving of ISP procedures, and primitive preservation of spectral information, raw Bayer images are still not used for major computer vision tasks, such as object detection. One reason is the discontinuity of spectral information in neighboring pixels across the raw Bayer images, which prohibits easy observation for human users and algorithm design that relies on smooth color changes. Especially for object detection tasks, clear object boundaries are required for tightly bounding the objects.

Guoyu Lu is with the Intelligent Vision and Sensing (IVS) Lab at the University of Georgia, USA, guoyulu62@gmail.com

Fig. 1. BayerDetect network has the capability of achieving fast and accurate object detection from 8-bit raw Bayer images.

In this work, we propose BayerDetect network, a novel object detection framework that can effectively detect objects based on raw Bayer images as Fig. 1. To capture better frequency spectrum and Bayer coordinates from the raw Bayer input via the high multi-path network, we replace the widely used ResNet-based backbone with a spectral frequency and spatial coordinate guided ResNext. The feature representations are then converted into a small number of proposals for efficiently learning and optimizing the sparse candidates instead of constraining the dense candidates from the Region Proposal Network (RPN). To overcome the lack of contextual connections among adjacent pixels as RGB color images, BayerDetect network proposes spectral and spatial attention mechanisms to establish the spectral and spatial relationship between adjacent pixels and mitigate the color disconnection issues that existed in raw Bayer images. The designated BayerDetect network can preserve the primitive spectral wavelength information from the camera sensor and capture the grid point locations from the Bayer pattern under the proposed framework for efficient data storage and computation. The overall framework is shown in Fig. 2.

The main contributions of our work are listed as follows:
1. We propose a novel framework to explore object detection based on raw Bayer images. The framework can achieve state-of-the-art detection accuracy with much lower storage

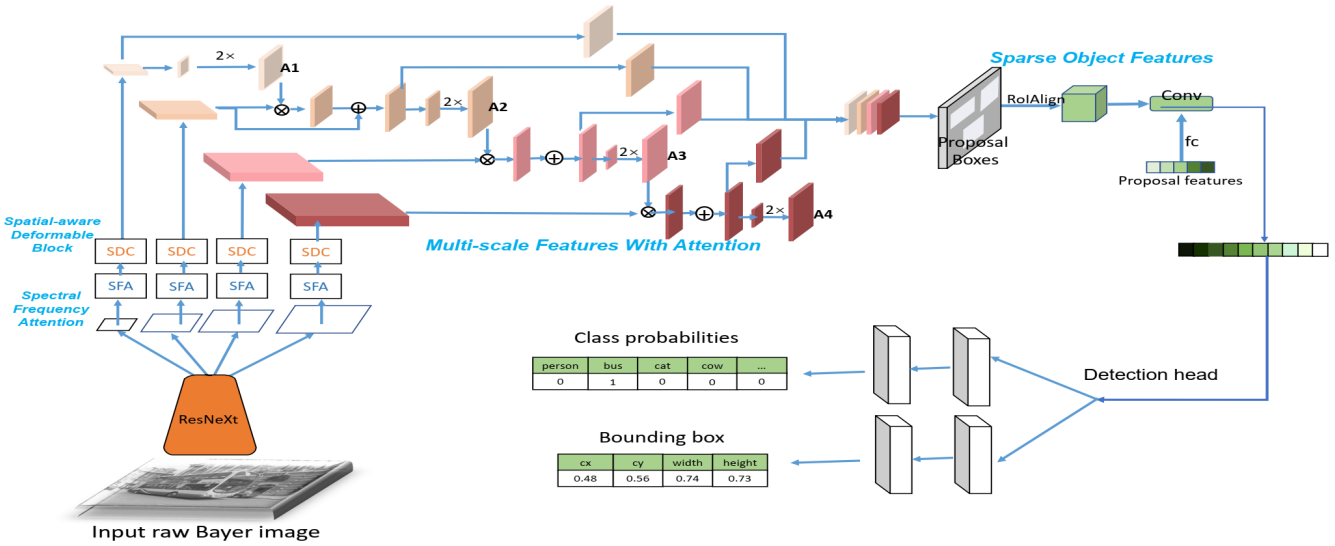

Fig. 2. Overview of our proposed BayerDetect network. A single raw Bayer image will be input into the detection network. The network will learn the spectral and spatial-aware deformable attention that builds the context relationship between Bayer pattern grids and mitigates the channel discontinuous issues of adjacent pixels of the raw Bayer images. Multiple-scale attention will be explored to detect objects of various scales in the Bayer images. The detection is accelerated through sparse proposals and features.

requirements and computation costs with Bayer input and sparse detector. To our best knowledge, BayerDetect network is one of the first detection frameworks specifically designed to detect objects on raw Bayer images, which demonstrates the suitability of raw Bayer images on detection tasks. 2. We decompose the raw Bayer images to different spectral frequencies to better recover light wavelengths recorded by the Bayer filter. The extracted spectral features are further processed by spatial coordinate attention to explore spatial dependencies among adjacent pixels. The output feature maps can sufficiently focus on high-frequency details and the Bayer grid pattern. 3. We propose spatial-aware deformable convolution to exploit spatial relationships between neighboring pixels with different color channels, which effectively mitigate the pixel disconnection issue of object boundaries.

II. Related Work

The goal of object detection is to localize and recognize each object with a bounding box in an image. It can be divided into anchor-based detectors and end-to-end detectors. Anchor-based object detector consists of single-stage detection [19] [28] [29] [34] and two stage detection [1] [18]. Among these detectors, the anchor boxes are predefined by the sliding windows and are assigned background and foreground samples. Considering the pre-defined anchor boxes are data independent, the training process usually requires a careful selection of the hyper-parameters for efficient optimization. Shape-prior-based detectors were designed to detect objects with specific shapes (e.g., parabola) [24], [26]. Recent fully convolution-based anchor-free approaches (e.g., CenterNet [5], CornetNet [14] and FCOS [39]) still require post-processing (e.g., non-maximum suppression (NMS)) steps for filtering those additional detections, which also requires careful selection on the threshold and might not be robust for complex scenes.

Different from aforementioned anchor-based approaches, end-to-end algorithms are explored [32] [30] [2] [12]. Recently, [12], DETR [2] and Deformable DETR [42] exploited the relations among the diverse objects and predictions to prevent additional post-processing steps. However, they are still relatively expensive with massive object candidates. Sparse-RCNN [35], as a purely sparse algorithm, focuses on a small group of bounding boxes and the learnable proposals instead of enumerating all dense images, hence efficient and fast. Hybrid matching [13] is further proposed to improve detection accuracy. 3D object detection can also be realized with simultaneous depth estimation [25], [16]. Queries are also applied to enhance the detection network training [6], [22]. The consistency information in videos is also applied to detect objects accurately [4]. And cross-transformer is developed for object detection [8]. Different from the approaches mentioned above, our method is highlighted to be the first to realize end-to-end object detection based on the spectral wavelength and pattern geometry of the raw Bayer images.

Most Bayer patterns using Bayer Color Filter Array (CFA) are designed for the reliable demosaicing process, which is performed to interpolate the vacant red, green, and blue values in the raw Bayer pattern images for restoring 3-channel RGB color images [17] [40] [27] [21]. Furthermore, hand-crafted algorithms are examined for color difference-based interpolation [3], edge directional interpolation [15], and reconstruction-based interpolation [33]. To encourage better image demosaicing, deep learning approaches were applied in [38] [37] [20]. On the other hand, Liu et al. [20] laid special emphasis on a self-guidance network where they used an initially estimated green channel as a guiding force to recover all of the input image's missing values by supervising the network training according to the edge loss of the reconstructed image as well as the ground truth color image. Moreover, [41] assessed the image restoration effect of restoring images from the raw domain of two different OLED displays.

Nonetheless, the direct and relatively straightforward raw Bayer images are rarely explored in computer vision algorithms, mainly because raw Bayer images are not straightforward to observe, thus adding difficulties in designing corresponding algorithms. This is especially an issue for object detection tasks, which depend on clear and continuous object boundaries to detect and bound the objects. These factors limit the broader applications of raw Bayer images. The paper demonstrates the possibility and benefits of utilizing raw Bayer images on object detection tasks based on our dedicated framework for Bayer patterns.

III. BayerDetect Object Detector

We propose BayerDetect network, a sparse object detection framework for specifically localizing and classifying objects on raw Bayer images. The proposed network introduces spectral frequency attention blocks to extract rich contextual information from the features of the multi-path group representation ResNeXt backbone. As the classical convolution operations have only fixed and limited receptive fields and are not able to adapt spatial information well for many unseen objects and scenes, we further propose utilizing spatial-aware deformable convolution for more accurate detection. This is especially helpful for raw Bayer images, as the object boundaries are not connected in color channels. The deformable convolution layers can provide attention to those disconnected pixels to extract shape information. A multi-scale attention mechanism is introduced to dynamically combine the multi-scale pyramid features to maximally preserve the low-level color and pattern features and the high-level object semantics for detection. The combined features are then aligned for each sparse proposal bounding box with the RoIAlign operation. To add more shape and geometry information to the coarse localization, the RoI features are weighted by the proposal features. Finally, the re-weighted features are forwarded to predict bounding boxes and precise object classification. The detailed analysis is elaborated in the following sections, where Sec. III-A introduces the principle and benefits of the Bayer pattern, Sec. III-B, Sec. III-C, Sec. III-D and Sec. III-E detail the architectures of each key component and the loss constraints in the network.

A. Formulation of Raw Bayer Images

Most commercial digital cameras contain a single CCD/CMOS sensor that can capture the light intensity but not its wavelength or color. Generally, a Bayer color filter array (CFA) is more productive because it overlays the image sensor to comprehensively cover the field of view and generate different filtered color information, which eventually leads to a raw Bayer image (image mosaic). Hence, normal RGB color S is possible to be recovered with colors from the split spectral channels S^R, S^G and S^B, where $S = S^R \cup S^G \cup S^B$. S^B and S^R together occupy a quarter of all pixels, whereas S^G occupies half of all image pixels that existed in a quincunx lattice.

A demosaicing method is required to recover a full 3-channel RGB image by interpolating the missing color

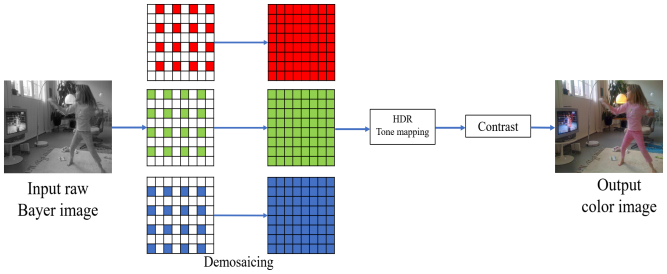

Fig. 3. An illustration of how the Bayer CFA is used in a typical camera ISP pipeline.

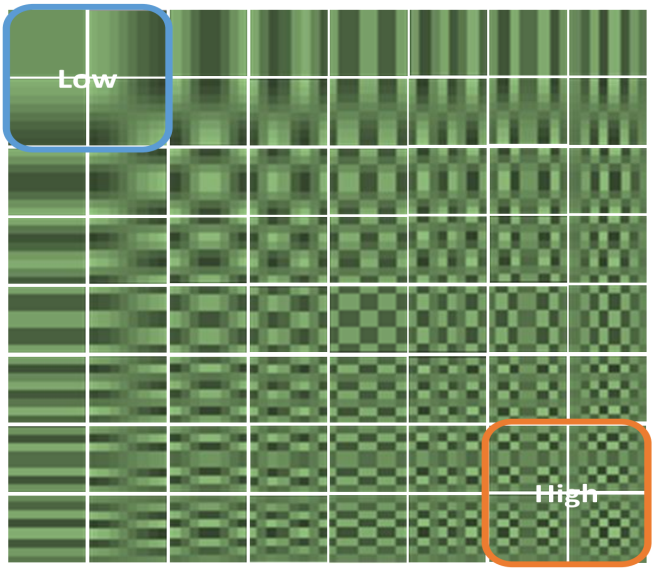

Fig. 4. An illustration of the transformed 2D DCT coefficients.

channels based on the raw Bayer image. This demosaicing method is effective in large areas with constant colors and smooth surfaces. For scenes featuring high contrast areas, such as those where colors constantly change and objects move, demosaicing results in unwanted loss of details, which could lead to color bleeding and artifacts such as zippering. In the meantime, post-processing phases like demosaicing add extra computation time and complexity in real applica-tions. Hence, raw Bayer images are more appropriate for end-to-end detection because the raw Bayer images preserve the most primitive color information at a low cost. An illustration of the raw Bayer image with zoom-in regions and standard post-processing phases is shown in Fig. 3.

B. Spectral Frequency Attention

As introduced above, the adjacent pixels of raw Bayer images present different spectral channels. So we introduce our dedicated spectral frequency attention block for exploring the different spectral information captured by the primitive camera sensor and the spatial coordinate attention for enhancing a wide range of contextual information on each local Bayer pattern. The attention blocks are integrated into the multi-path group representation ResNeXt for comprehensive feature extraction.

Considering the raw Bayer CFA inherently captures different light wavelengths via the Bayer pattern and the high-frequency details (e.g., fine boundaries) are important for ob-

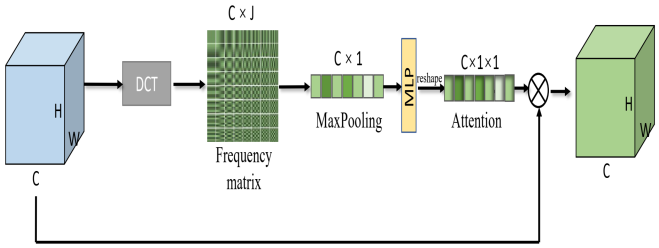

Fig. 5. The detailed structure of the spectral frequency attention block. The input feature map is decomposed to the spectral frequency domain. The output feature map is weighted by the attention from MLP and shares the same dimension as the input feature map.

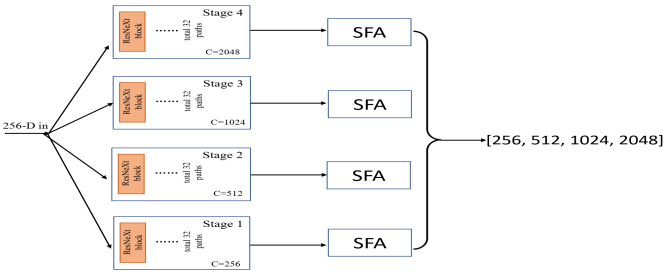

Fig. 6. An illustration of the ResNeXt block with our spectral frequency attention.

ject localization, we propose decomposing the image features to the frequency spectrum via Discrete Cosine Transform (DCT). DCT is able to represent an image in the format of a combination of sinusoids at different magnitudes and frequencies, as Fig. 4. Provided with the input feature map $X \in R^{M \times N}$, the 2D DCT spectrum $F \in R^{M \times N}$ can be expressed as:

$$F_{ij} = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} a_{i,j} \cos \frac{\pi(2m+1)i}{2M} \cos \frac{\pi(2n+1)j}{2N} \quad (1)$$

where $a_{0,0}$ is $\sqrt{\frac{1}{MN}}$, corresponding to the lowest frequency. For all other frequencies, $a_{i,j}$ can be computed as $\sqrt{\frac{2}{MN}}$. Given the input feature map $X \in R^{C \times H \times W}$, the corresponding DCT coefficients are in $R^{F \times C \times H \times W}$ for the selected F frequency components. By conducting DCT on the input feature map and summarizing the output across depth channels, the frequency matrix is obtained with a dimension of $C \times J$. Here, $J = H \times W$. The embedding matrix is then used to select the highest frequency response via max pooling. The final output feature map is weighted by the input feature map and the frequency attention vector, as Fig. 5.

The introduced spectral frequency attention block is integrated into the ResNeXt as the backbone of the proposed BayerDetect network. Extending the idea of ResNet [10] and Inception [36], ResNeXt utilizes a multi-path strategy to stack the ResNet blocks and the grouping convolution operation in Inception to aggregate the useful information to increase accuracy without increasing complexity of the model. An illustration of a basic block from ResNeXt integrating with our introduced modules is shown in Fig. 6.

C. Multi-scale Features With Attention

Object detection usually relies on high-level features that represent more object semantics and class information. However, to fully exploit the Bayer pattern, we propose highlighting the effect of the low-level features on the high-level features to enrich the meaningful shape and color information.

For two feature maps with diverse scales, we predict scale-aware attention map to enhance the benefits of the Bayer pattern on the detection performance. As illustrated in Fig. 7, for each feature map from a smaller scale, we first forward it to a 3×3 convolutional layer activated by the sigmoid function to generate an attention map which has the same dimension as the input. Then, the attention map is up-sampled to the same size as the larger-scale feature map. Element-wise multiplication is utilized to highlight the larger-scale feature map with the learned attention map dynamically. Finally, the attended larger feature map is added to the original input feature map via the residual connection to prevent potential feature degradation. Given the input smaller scale feature map O, larger scale feature map L, and the up-sampled learned attention map $A\uparrow$, the output re-weighted feature map R can be expressed as: $R = O \cdot A\uparrow + L$ where $R, A\uparrow$ and L share the same dimension of $H \times W \times C$.

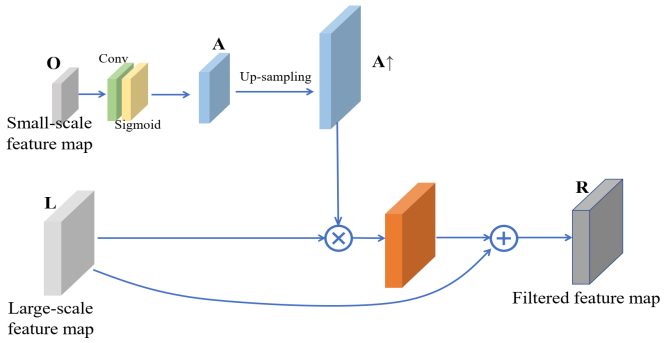

Fig. 7. An illustration of multi-scale feature attention mechanism.

D. Spatial-aware Deformable Convolution for Bayer Pattern

Due to spectral discontinuity, object boundaries are not closely connected in position. Normal convolutional operations utilize the fixed locations and receptive field with an input feature map and output the new features by computing the weighted sum. However, fixed convolutions are not best suited for dramatic object shape changes, especially for Bayer images with pixels in neighboring locations representing different color channels. Therefore, in addition to spectral attention, inspired by the deformable convolution, we propose a novel spatial-aware deformable convolution that is more suitable for Bayer grids. Different from the normal deformable convolution, the offset for each convolutional element within a sample grid is separately learned and does not share the common parameter, which improves the capability of adaptively capturing object shapes and spatial information. More specifically, we propose the spatial-aware deformable convolution connecting the backbone and the Feature Pyramid Network (FPN) in order to more efficiently

	Train	Validation	Test
Images	3,500	500	500
Boxes	33,572	3,455	3,502

TABLE I

DETAILS OF OUR COLLECTED RAW BAYER IMAGE DATASET.

model the geometric information that existed in different locations.

Given that the input feature map from the backbone is X, the output feature map is Y and each position on the output feature map is $p = (x', y')$, the 3×3 deformable convolution at p can be expressed as:

$$Y(p) = \sum_{i=0}^{8} W(p_{ori}) \cdot I(p + D \cdot p_{ori} + Off(n)) \quad (2)$$

where W is a matrix of weights. p_{ori} represents the positions of the regular grid, and D is the dilation rate. For the normal convolution, the spatial position will be $p + D \cdot p_{ori}$. In comparison, for the spatial-aware deformable convolution, the output spatial position will be $p + D \cdot p_{ori} + Off(n)$, which is able to better generalize and more sensitive to irregular shapes and structures. An illustration of the proposed spatial-aware block is shown in Fig. 8.

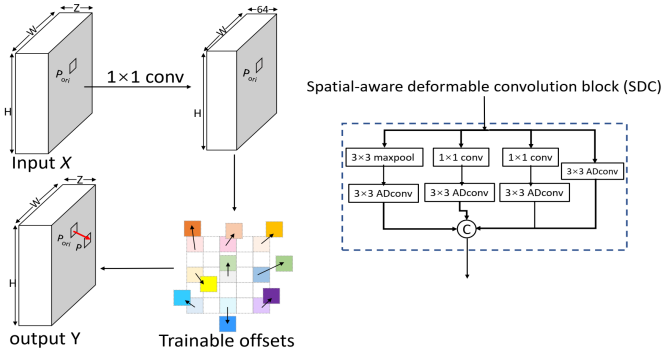

Fig. 8. Overview of the introduced spatial-aware deformable convolution and its block in the network.

E. Sparse Object Proposals and Features

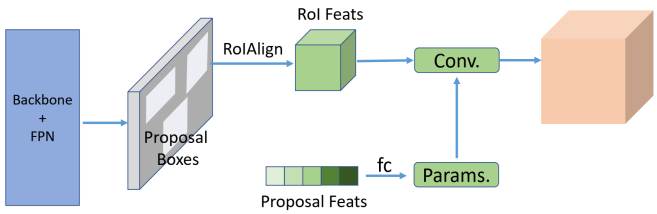

Fig. 9. An illustration of the generation of sparse object features from our network. RoI features and the corresponding proposal features are interacted with to enhance the object localization estimation.

With the introduced attention-based ResNeXt and FPN for generating dynamic multiscale feature maps, a sparse object detection method based on [35] is adopted to prevent dense and redundant object detection candidates for further improving efficiency and accuracy.

More specifically, a small set of proposal boxes in the size of $N \times 4$ is designed as trainable parameters for preventing the dense object positional candidates across the entire image Bayer grids. RoI features are then extracted from the proposal boxes by RoIAlign. Moreover, a group of high-dimension proposal features is utilized for better encoding the rich object characteristics such as shapes. The RoI features, and the proposal features are combined with a one-to-one interaction for generating the feature maps for later object localization and classification. An illustration of the sparse RoI feature generation is depicted in Fig. 9.

Fig. 10. Samples of our collected raw Bayer image dataset.

Object classification and bounding box regression loss: To accurately classify objects, focal loss L_{cls} [19] is deployed to obtain matching between the predicted classification and the ground truth category label. Another major loss is for regressing the bounding boxes. Unlike many existing detectors to estimate bounding box dimensions under initial assumptions, we directly estimate the bounding boxes' positions and dimensions. The most commonly used L1 loss is applied for regression. However, as our network has multiple scales of bounding box predictions, the loss function should be invariant to bounding box scales, especially when the bounding box prediction is far from the ground truth. To mitigate this influence, we further utilize a combination of L1 and the scale-invariant generalized IoU loss L_{GIoU}. Therefore, the comprehensive bounding box regression loss becomes:

$$L_{reg} = \lambda_{L1} \| b_i - \tilde{b}_i \|_1 + \lambda_{giou} L_{GIoU}(b_i, \tilde{b}_i) \quad (3)$$

where b_i indicates the i th ground truth bounding box location and $\tilde{b}_i$ indicates the i th estimated bounding box. λ_{L1} and λ_{giou} are hyper-parameters for L_1 and L_{GIoU} set to be 5.0 and 2.0 respectively. The overall objective function is a sum of the abovementioned losses across the number of detected objects as $L_{all} = \lambda_{cls} L_{cls} + \lambda_{reg} L_{reg}$, where λ_{cls} and λ_{reg} are set as 2.0 and 1.0.

IV. EXPERIMENTS

We first evaluate our approach on the Bayer images generated from the public MS COCO dataset and our collected real-world raw Bayer image dataset. The Pascal VOC dataset is also used to further verify the generality of our method. We then perform comprehensive ablation studies regarding network structures, input image type, and losses.

Dataset: Our experiments are conducted on the Bayer images generated from the challenging MS COCO benchmark, including 80 categories. Models are trained on the official COCO train2017 split and evaluated on the validation set. The trained model is further verified on the VOC 2007 test set without training for generalization validation. We reverse-engineered the RGB images to 8-bit Bayer images and used the simulated data for training and testing. We mimicked the mosaic image with Bayer filter to extract only a single

Method	Backbone	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [19]	ResNet-50	16.7	37.4	58.8	41.4	22.4	41.4	49.1
RetinaNet [19]	ResNeXt-50	16.1	37.7	58.9	41.6	22.5	42.0	49.3
Faster-RCNN [31]	ResNet-50	20.2	37.9	58.8	41.1	22.4	41.1	49.1
Faster-RCNN [31]	ResNeXt-50	19.3	38.5	59.1	41.3	22.9	41.6	49.8
Cascade RCNN [1]	ResNet-50	14.2	40.2	59.6	43.5	23.8	44.1	52.8
Cascade RCNN [1]	ResNeXt-50	13.7	40.8	59.7	44.0	23.8	44.6	52.9
DETR [2]	ResNet-50	18.2	15.4	29.4	14.5	4.3	15.1	26.7
DETR [2]	ResNeXt-50	17.7	15.6	29.9	14.7	4.5	15.4	27.0
Deformable DETR [42]	ResNet-50	20.6	37.2	55.5	40.5	21.1	40.7	50.5
Deformable DETR [42]	ResNeXt-50	20.1	37.8	55.9	41.0	21.3	41.1	50.6
Sparse RCNN [35]	ResNet-50	22.8	42.8	59.3	45.9	24.5	44.9	54.0
Sparse RCNN [35]	ResNeXt-50	22.2	42.9	59.5	45.9	24.7	45.0	54.2
Ours	ResNet-50	27.9	45.9	61.1	47.0	24.9	45.2	54.9
Ours	ResNeXt-50	27.1	46.2	61.4	47.2	25.1	45.8	55.2

TABLE II

COMPARISON WITH DIFFERENT OBJECT DETECTORS ON THE SIMULATED RAW BAYER COCO 2017 VALIDATION DATASET.

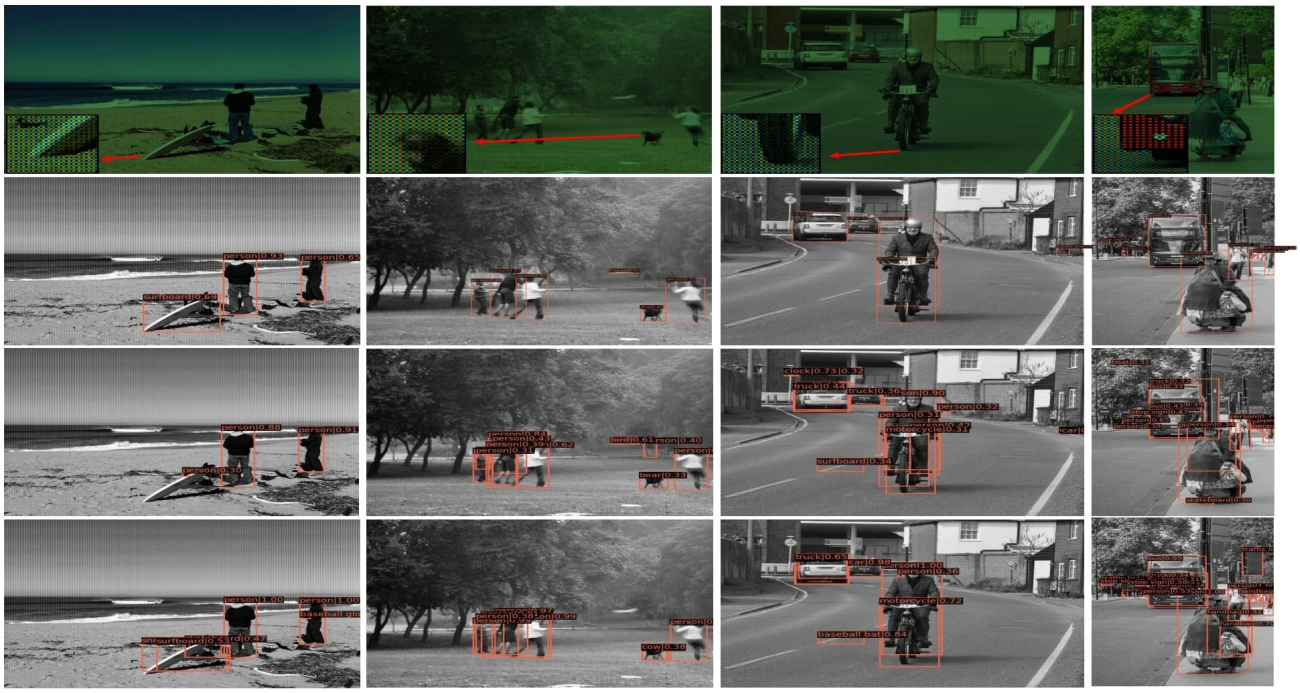

Fig. 11. Comparison of qualitative results on the simulated COCO dataset. From top to bottom: input raw Bayer image with color pattern overlaid on it for easier observation, our result, Cascade RCNN [1] detection, Sparse RCNN [35] detection.

channel for each 3-channel pixel in the sorted order of a specific pattern. In this work, we are mimicking RGGB as the most common Bayer pattern. For testing the framework on the real-world raw Bayer images, we collect and label the dataset based on the raw Bayer images captured by Nikon D3500 digital camera, which consists of 10 categories. The split of training, validation, and testing set is depicted in Table I, and the sample images of our dataset are provided in Fig. 10.

Implementation Details: ResNeXt is used as the backbone network. The base learning rate is $2.5e-5$, which gradually decreases from the 32nd epoch by a factor of 0.1. We used AdamW [23] optimizer with weight decay 0.0001. Models are trained for 42 epochs.

A. Results in Comparison with State-of-the-art

First, we quantitatively compare the proposed method with recent state-of-the-art approaches. As shown in Table II, our proposed method outperforms state-of-the-art detectors, especially by a large margin when compared with Cascade RCNN [1] and Sparse RCNN [35]. on AP and AP_L. The same evaluation results are also reported on our real collected raw Bayer image dataset in Table III. We note our method exhibits high accuracy (AP = 52.5) with the ResNeXt depth as only 50. The comparison results in Table III demonstrated that the proposed method performs well on raw Bayer images from real-world scenes. Meanwhile, due to the single-channel image processing, the network complexity is reduced based on our dedicated network structure, which results in a significant increase in detection speed in terms of FPS.

We qualitatively evaluate the performance of the object detection as shown in Fig. 11 and Fig. 12. It can be easily noticed that our detector significantly improved [1] [35] in preventing miss-predictions of persons and backpacks. And our detection matches well with the labeled ground truth annotations.

To widely validate the performance of the proposed network and test the generalization capability on different

Method	Backbone	FPS	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
RetinaNet [19]	ResNet-50	11.2	42.1	65.8	46.3	25.0	46.3	54.9
RetinaNet [19]	ResNeXt-50	10.7	42.8	65.9	46.7	25.4	46.6	55.1
Faster-RCNN [31]	ResNet-50	13.4	42.4	66.1	46.5	25.3	46.0	54.3
Faster-RCNN [31]	ResNeXt-50	13.0	42.7	66.7	46.8	25.7	46.5	54.9
Cascade RCNN [1]	ResNet-50	9.9	45.7	66.9	48.9	26.9	49.3	59.0
Cascade RCNN [1]	ResNeXt-50	9.7	46.0	67.1	49.0	27.0	49.6	59.4
DETR [2]	ResNet-50	13.1	17.6	33.1	16.5	4.8	16.2	30.1
DETR [2]	ResNeXt-50	12.6	18.1	33.4	16.8	5.7	16.4	30.5
Deformable DETR [42]	ResNet-50	14.0	42.4	62.2	45.3	23.7	45.3	56.8
Deformable DETR [42]	ResNeXt-50	13.5	43.0	62.9	45.6	24.0	45.8	57.0
Sparse RCNN [35]	ResNet-50	15.8	48.9	66.5	51.2	27.3	50.4	58.9
Sparse RCNN [35]	ResNeXt-50	15.1	49.2	66.9	51.8	27.8	50.7	59.2
Ours	ResNet-50	19.3	51.8	68.7	52.0	27.9	50.7	61.6
Ours	ResNeXt-50	18.6	52.5	69.4	52.6	28.4	51.4	61.9

TABLE III

COMPARISON WITH DIFFERENT STATE-OF-THE-ART OBJECT DETECTORS ON OUR COLLECTED REAL-WORLD RAW BAYER IMAGE DATASET.

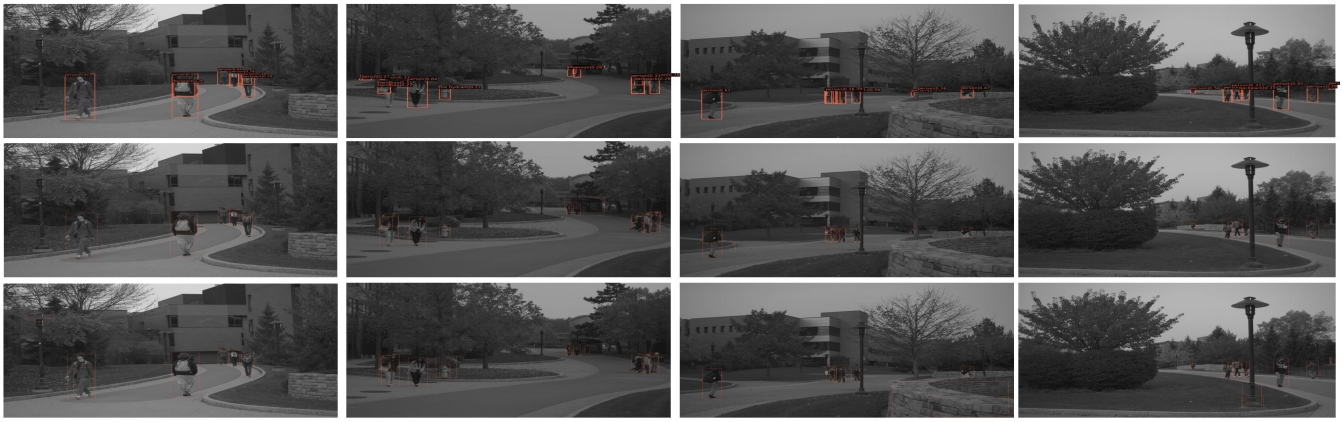

Fig. 12. Comparison of qualitative results on the real collected raw Bayer image dataset. From top to bottom: our result, Cascade RCNN [1] detection, Sparse RCNN [35] detection.

	Method	Backbone	mAP
VOC 2017	RetinaNet	ResNet-50	36.4
	Cascade RCNN	ResNet-50	46.5
	Sparse RCNN	ResNet-50	49.7
	Our	ResNet-50	52.1
	Our	ResNeXt-50	53.6

TABLE IV

DETECTION RESULTS ON BAYER IMAGES GENERATED FROM VOC 2017 TEST SPLIT.

Method	Type	AP	δ(%)	Params (M)
RetinaNet	RGB	39.0	–	90.4
	Gray	37.9	1.1 ↑	81.3
	Bayer	37.7	1.3 ↑	81.3
Ours	RGB	46.3	–	35.3
	Gray	44.2	1.1 ↓	26.2
	Bayer	46.2	0.1 ↓	26.2

TABLE V

ABLATION STUDY ON THE EFFECT OF DIFFERENT INPUT IMAGE TYPES ON THE DETECTION PERFORMANCE AND THE NUMBER OF NETWORK PARAMETERS.

scenes, we evaluate our network on VOC 2017 dataset in Table IV. Our method leads to about 2.4% mAP improvement in comparison with the top performing method [35] on VOC 2017 test dataset with the same ResNet-50 backbone structure and 3.9% improvement using ResNeXt-50 backbone.

B. Ablation Study

Ablation study on network inputs: Table V compares the accuracy and the computation cost from different types of image inputs. It can be observed through having 8-bit channels, our proposed method shows significantly higher performance on the Bayer images than the results from the grayscale images, which indicates the proposed network is able to learn richer location and shape contexts from the Bayer pattern. When comparing the detection accuracy between our results from RGB images and Bayer inputs, there is a significant decrease in network complexity, which indicates that our method can reduce storage for both images and the network.

Spectral Freq Att.	Multi-scale Att.	Spatial-aware Deform	L_{GIoU}	AP	δ(%)
✓	–	–	–	44.0	–
✓	–	–	–	44.8	0.8 ↑
✓	✓	–	–	45.1	1.1 ↑
✓	✓	✓	–	45.9	1.9 ↑
✓	✓	✓	✓	46.2	2.2 ↑

TABLE VI

ABLATION STUDY ON THE EFFECT OF DIFFERENT COMPONENTS AND LOSSES.

Ablation study on key components: We further conduct a more detailed analysis of the effect of each key component in our network design, as shown in Table VI. One can observe that spectral frequency attention and multi-scale feature attention together contribute a noticeable 1.1% improvement compared with the naive implementation without any proposed component. A spatial-aware deformable convolution block further achieves a 0.8% gain following the attention blocks. The incorporation of the generalized IoU loss for reducing the unbalanced data distribution further raises the mIoU to 46.2%, which outperforms the naive implementation by almost 2.2%.

Proposal	AP	AP_{50}	AP_{75}	FPS
100	43.9	60.2	46.1	27.5
200	45.0	60.9	46.3	27.4
400	46.2	61.4	47.2	27.1
800	46.7	61.8	48.1	26.2

TABLE VII

ABLATION STUDY ON THE EFFECT OF THE DIFFERENT NUMBERS OF PROPOSALS.

Ablation study on proposal numbers: We further explore the effect of proposal numbers on our method in Table VII. The number of proposals from 100 to 400 increases the detection accuracy dramatically. This accuracy enhancement slows down from proposal number 400 to 800. Meanwhile, the detection speed in terms of FPS decreases dramatically. Therefore, we selected 400 proposals for our proposed method based on the trade-off of detection accuracy and speed.

V. CONCLUSION

We propose BayerDetect network, an end-to-end object detection framework to detect and recognize objects on raw Bayer images. With the 8-bit raw Bayer images without post-processing ISP from the camera sensor, our approach is able to save both computation time and memory. This architecture explores the spectrum and geometry characteristics of the Bayer pattern to mitigate the neighboring pixels' channel intersection issues of Bayer images, which achieves consistent and significant performance gains on challenging public dastasets and our collected raw Bayer image dataset.

ACKNOWLEDGEMENT

This publication is based upon work supported by NSF under Awards No. 2105257 and 2126643.

REFERENCES

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In CVPR, pages 6154–6162, 2018.
[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, pages 213–229, 2020.
[3] K-H Chung and Y-H Chan. Color demosaicing using variance of color differences. TIP, 15(10):2944–2955, 2006.
[4] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In CVPR, pages 13678–13688, 2022.
[5] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In ICCV, pages 6569–6578, 2019.
[6] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In CVPR, pages 5364–5373, 2022.
[7] Ross Girshick. Fast r-cnn. In ICCV, pages 1440–1448, 2015.
[8] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. Few-shot object detection with fully cross-transformer. In CVPR, pages 5321–5330, 2022.
[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In ICCV, pages 2961–2969, 2017.
[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.
[11] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. ToG, 33(6):1–13, 2014.
[12] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In CVPR, 2018.
[13] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In CVPR, 2023.
[14] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In ECCV, pages 734–750, 2018.
[15] Wonjae Lee, Seongjoo Lee, and Jaeseok Kim. Cost-efffective color filter array demosaicing using spatial correlation. TCE, 52(2):547–554, 2006.
[16] Chengyao Li, Jason Ku, and Steven L Waslander. Confidence guided stereo 3d object detection with split depth estimation. In IROS, pages 5776–5783, 2020.
[17] Xin Li, Bahadir Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In VCIP, volume 6822, pages 489–503, 2008.
[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In CVPR, pages 2117–2125, 2017.
[19] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In ICCV, 2017.
[20] Lin Liu, Xu Jia, Jianzhuang Liu, and Qi Tian. Joint demosaicing and denoising with self guidance. In CVPR, pages 2240–2249, 2020.
[21] Shumin Liu, Jiajia Chen, Yuan Xun, Xiaojin Zhao, and Chip-Hong Chang. A new polarization image demosaicking algorithm by exploiting inter-channel correlations with guided filtering. TIP, 29:7076–7089, 2020.
[22] Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, and Xiaodan Liang. Capdet: Unifying dense captioning and open-world detection pretraining. In CVPR, 2023.
[23] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.
[24] Guoyu Lu, Li Ren, Jeffrey Caplan, and Chandra Kambhamettu. Stromule branch tip detection based on accurate cell image segmentation. In ICIP, pages 3300–3304, 2017.
[25] Yawen Lu, Qianyu Guo, and Guoyu Lu. A geometric convolutional neural network for 3d object detection. In GlobalSIP, 2019.
[26] Yawen Lu and Guoyu Lu. 3d modeling beneath ground: Plant root detection and reconstruction based on ground-penetrating radar. In WACV, pages 68–77, 2022.
[27] Daniele Menon and Giancarlo Calvagno. Color image demosaicking: An overview. SPIC, 26(8-9):518–533, 2011.
[28] Han Qiu, Yuchen Ma, Zeming Li, Songtao Liu, and Jian Sun. Borderdet: Border feature for dense object detection. In ECCV, pages 549–564, 2020.
[29] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
[30] Mengye Ren and Richard S Zemel. End-to-end instance segmentation with recurrent attention. In CVPR, pages 6656–6664, 2017.
[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS, 28, 2015.
[32] Amaia Salvador, Miriam Bellver, Victor Campos, Manel Baradad, Ferran Marques, Jordi Torres, and Xavier Giro-i Nieto. Recurrent neural networks for semantic instance segmentation. arXiv preprint arXiv:1712.00617, 2017.
[33] Ling Shao and Amin Ur Rehman. Image demosaicing using content and colour-correlation analysis. SP, 103:84–91, 2014.
[34] Lin Song, Yanwei Li, Zhengkai Jiang, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. Fine-grained dynamic head for object detection. NeurIPS, 33:11131–11141, 2020.
[35] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In CVPR, pages 14454–14463, 2021.
[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In AAAI, volume 31, 2017.
[37] Daniel Stanley Tan, Wei-Yang Chen, and Kai-Lung Hua. Deepdemosaicking: Adaptive image demosaicking via multiple deep fully convolutional networks. TIP, 27(5):2408–2419, 2018.
[38] Runjie Tan, Kai Zhang, Wangmeng Zuo, and Lei Zhang. Color image demosaicking via deep residual learning. In ICME, 2017.
[39] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In ICCV, 2019.
[40] Chao Zhang, Yan Li, Jue Wang, and Pengwei Hao. Universal demosaicking of color filter arrays. TIP, 25(11):5173–5186, 2016.
[41] Yuqian Zhou, David Ren, Neil Emerton, Sehoon Lim, and Timothy Large. Image restoration for under-display camera. In CVPR, pages 9179–9188, 2021.
[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. ICLR, 2021.