

Large, Complex, and Realistic Safety Clothing and Helmet Detection: Dataset and Method

Fusheng Yu, Jiang Li[✉], Xiaoping Wang[✉], Senior Member, IEEE, Shaojin Wu, Junjie Zhang, and Zhigang Zeng[✉], Fellow, IEEE

Abstract—Detecting safety clothing and helmets is paramount for ensuring the safety of construction workers. However, the development of deep learning models in this domain has been impeded by the scarcity of high-quality datasets. In this study, we construct a large, complex, and realistic safety clothing and helmet detection (SFCHD) dataset. SFCHD is derived from two authentic chemical plants, comprising 12,373 images, 7 categories, and 50,552 annotations. We partition the SFCHD dataset into training and testing sets with a ratio of 4:1 and validate its utility by applying several classic object detection algorithms. Furthermore, drawing inspiration from spatial and channel attention mechanisms, we design a spatial and channel attention-based low-light enhancement (SCALE) module. SCALE is a plug-and-play component with a high degree of flexibility. Extensive evaluations of the SCALE module on both the ExDark and SFCHD datasets have empirically demonstrated its efficacy in enhancing the performance of detectors under low-light conditions. The dataset and code are publicly available at <https://github.com/ljfrank/open/SFCHD-SCALE>.

Note to Practitioners—This paper was motivated by the challenge of safety clothing and helmet detection in construction safety inspections. At high-risk construction sites, protective gear like safety clothing and helmets is crucial for worker protection. Monitoring their use is of significant safety implication and practical value. Currently, only two helmet detection benchmarks have been published, focusing mainly on helmets and neglecting safety clothing, which could cause safety risks in real applications. Additionally, existing datasets feature overly simplified backgrounds and idealized lighting, differing greatly from actual work conditions. This may lead to poor generalization of models trained on these datasets. To tackle these issues, we introduce a large, complex, and realistic safety clothing and helmet detection (SFCHD) dataset. It presents more authentic work scenarios with intricate backgrounds, with images captured from two chemical plants’ 40 different scenes via surveillance cameras.

Manuscript received 10 August 2024; revised 10 August 2024; accepted 10 August 2024. This work was supported in part by the Interdisciplinary Research Program of HUST under Grant 2024JCYJ006, the National Natural Science Foundation of China under Grant 62236005 and 61936004, and the Fundamental Research Funds for the Central Universities of HUST under Grant 2024JYCXJJ067 and 2024JYCXJJ068. Thanks to Wuhan Qianyang Intelligent Technology Co., Ltd. for supporting this work. (*Fusheng Yu and Jiang Li contributed equally to this work. Corresponding author: Jiang Li and Xiaoping Wang.*)

The authors are with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China, also with the Institute of Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China, also with the Hubei Key Laboratory of Brain-Inspired Intelligent Systems, Huazhong University of Science and Technology, Wuhan 430074, China, and also with the Key Laboratory of Image Processing and Intelligent Control (Huazhong University of Science and Technology), Ministry of Education, Wuhan 430074, China (e-mail: yufusheng@hust.edu.cn; ljfrank@hust.edu.cn; wangxiaoping@hust.edu.cn; wushaojin@hust.edu.cn; zjunjie@hust.edu.cn; zgzeng@hust.edu.cn).

Digital Object Identifier

SFCHD offers abundant training samples and benchmarks for various detection tasks, including small object and high-low light detections. We also design the SCALE module to overcome the limitations of existing algorithms under low-light conditions. This plug-and-play module excavates valuable information from low-light images in both spatial and channel dimensions, significantly boosting detection model performance. Our dataset and method can be deployed in real-world construction safety inspections for real-time monitoring of workers’ safety gear, ensuring their personal safety.

Index Terms—Computer industry, object detection, image processing, construction safety.

I. INTRODUCTION

SAFETY issues are an enduring and significant concern across all industries, particularly in high-risk construction sites such as chemical plants. Protective equipment, including safety clothing and helmets, is crucial for safeguarding workers in high-risk construction sites. Helmets can effectively prevent head injuries caused by falling or splashing objects, while safety clothing can protect the body and arms from hazardous chemicals and liquids. The absence of safety clothing and helmets frequently results in safety accidents, exerting a devastating impact on families and society. Consequently, monitoring the usage of safety clothing and helmets in factories or construction sites is of immense safety significance and broad application value.

Currently, the conventional automatic monitoring technology for safety clothing and helmet compliance is a category of sensor-based methods [1], [2]. These techniques monitor whether workers are wearing safety attire or helmets by installing sensors on them and analyzing the emitted signals. However, sensor-based approaches necessitate substantial investment in procurement, installation, and maintenance, leading to relatively high costs. In recent years, deep learning methods have gained substantial attention in computer vision [3], [4] due to their ability to self-learn useful features from large-scale, annotated training data. In particular, convolution neural networks are widely used for image classification and object detection. For instance, LeCun et al. [5] utilized convolutional neural networks (CNNs) to recognize handwritten digits; Kolar et al. [6] applied CNNs to detect safety guardrails in construction sites; Nath et al. [7], [8] adopted CNNs to identify common construction-related objects, e.g., buildings, equipment, and workers. These endeavors have provided new insights into the monitoring of safety clothing and helmet usage.

The tremendous success of deep learning technique is largely attributed to the availability of large-scale, high-quality, and meticulously annotated datasets. To the best of our knowledge, only two benchmark datasets for helmet detection have been published to date, namely Pictor-v3 [9] and SHWD [10]. These datasets primarily focus on the detection of helmets, with insufficient consideration given to safety clothing, leading to potential safety hazards in practical applications. Moreover, the image background in existing datasets is overly simplified, and the lighting conditions are too idealized, which is significantly different from real-world work environments. Training models directly on these datasets may lead to insufficient generalization capabilities in their practical applications. Therefore, these two datasets do not fully meet the needs of real-world applications.

To address the above limitations, in this work, we contribute a large, complex, and realistic high-quality safety clothing and helmet detection (SFCHD) dataset. Fig. 1 illustrates a comparison between the SFCHD dataset and existing datasets. It is evident that our SFCHD dataset offers more realistic work scenarios and more complex background environments. The SFCHD dataset comprises 12,373 images, covering 7 categories, with a total of 50,558 labeled instances, including labels such as Person, Safety Helmet, Safety Clothing, Other Clothing, and Head. To more closely mimic real-world work environments, we add two new labels for safety helmets and clothing under varying lighting conditions and shadow coverage: Blurred Clothing and Blurred Head. All images are captured from factory surveillance cameras, encompassing 40 different scenes across two chemical plants. Each instance has been manually annotated by professional inspectors, ensuring the accuracy of the annotations. It is worth noting that our SFCHD dataset not only provides a rich set of training samples but also serves as a benchmark for the evaluation of various detection tasks, such as small object detection, and high-low light object detection.

To more closely approximate real-world factory environments, our dataset includes a multitude of images captured under extreme lighting conditions, such as overexposure or underexposure, reflecting the actualities of the real world. When confronted with these extreme conditions, typical detection models often perform suboptimally. The reason is the presence of substantial noise in the images, which has a detrimental effect on the expression of potential features. To address this challenge and enhance the model's robustness in practical application scenarios, we introduce a spatial and channel attention-based low-light enhancement (SCALE) module. The SCALE module consists of two parallel pathways, i.e., the spatial attention pathway (SAP) and the channel attention pathway (CAP). SAP focuses on extracting features from key regions in the image, enhancing the model's perception of spatial feature relationships. CAP, on the other hand, assesses the importance of different channels, learning channel-specific feature information, assisting the model in extracting features such as color or texture from the image. SAP is composed of two scale-differentiated attention, achieving refined feature fusion and latent feature representation. The design of CAP is inspired by the channel attention mechanism in SENets [11]

and incorporates a skip connection structure. Unlike many image enhancement models that require specific loss functions, the proposed SCALE is a stand-alone, plug-and-play component, which can be seamlessly integrated into detection models and trained efficiently using only detection loss.

The main contributions of this work are as follows:

- We contribute a large, complex, and realistic high-quality benchmark named SFCHD. To our knowledge, SFCHD is the inaugural factory dataset that concurrently satisfies the detection of both safety clothing and helmets.
- To address the challenge for image detection under low-light conditions, we propose a novel low-light enhancement module named SCALE. This plug-and-play module is composed of a spatial attention pathway and a channel attention pathway.
- We apply several classic object detection models on the SFCHD dataset for experimentation to verify its usability. Concurrently, we evaluate the SCALE module on both the ExDark and SFCHD datasets. The results substantiate that SCALE demonstrates significant enhancements in the detection of low-light images.

The rest of this paper is organized as follows. In Section II, we review the literatures pertinent to this work, encompassing the current state of datasets and the latest advancements of algorithms. In Section III, we provide a detailed account of the SFCHD dataset, including its construction principles, data details, and potential applications. We introduce the architecture of the SCALE module in Section IV. In Section V, we initially apply several classical models to substantiate the SFCHD's usability, followed by an assessment of the SCALE's efficacy on the ExDark and SFCHD datasets. Finally, in Section VI, we summarize this work.

II. RELATED WORK

A. Object Detection Datasets

The most remarkable datasets in the field of object detection include Pascal VOC [12], [13], MS-COCO [14], ILSVRC [15], and Open Image [16]. The Pascal Vision Object Class (VOC) Challenge is a competition to accelerate the advancement of the field of computer vision, and two versions of the datasets adopted in the competition, VOC 2007 and VOC 2012, were commonly used for standard benchmarks. The Microsoft Common Objects in Context (MS-COCO) is one of the most challenging object detection datasets available, including two commonly used versions, COCO 2005 and COCO 2018. MS-COCO contains 80 categories, whose annotated objects are divided into three categories based on scale: small objects, medium objects, and large objects. The ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) is also a competition that includes an object detection task, and its dataset serves as a benchmark for the evaluation of algorithm performance. The number of objects in the LSVRC was expanded to 200 compared to the VOC dataset, and the number of images and annotated objects are significantly larger than those in the VOC dataset. The Open Image dataset was launched by Google in 2017 and consists of 9.2 million images with 16 million bounding boxes for 600 categories on 1.9 million images used



Fig. 1. Examples of images from existing datasets and the SFCHD dataset. Here, (a)-(e), (f)-(j), and (k)-(m) are images selected from the Pictor-v3, SHWD, and SFCHD datasets, respectively.

for detection, becoming one of the largest object detection datasets.

Despite the multitude of datasets [12]–[16] available for the field of object detection, there is a relative dearth of datasets specifically tailored to industrial scenarios. The availability of high-quality and authentic datasets for safety clothing and helmet detection is crucial for ensuring engineering safety and factory compliance. Although object detection algorithms have made significant strides, their widespread application in industrial settings is still limited due to the absence of high-quality datasets. To our knowledge, there are currently only two open-source datasets available for helmet detection, namely Pictor-v3 [9] and SHWD [10]. However, the utility of these datasets is constrained, and they struggle to meet the application demands of both academia and industry. The Pictor-v3 dataset contains only 1,330 annotated images, which is insufficient to meet the contemporary requirement for large-scale data in deep learning. Moreover, the dataset is deficient in the inclusion of negative samples, leading to an imbalanced data composition and a disproportionate emphasis on color labels. In reality, safety helmets come in a variety of colors, and an overemphasis on color labels could potentially impair the model’s ability to recognize the physical attributes of the helmets. The SHWD dataset comprises 3,241 images related to safety helmets, with 10,457 instances annotated. Although there is a marked increase in the number of images compared to Pictor-v3, most images in SHWD are downloaded from the web, featuring relatively simple backgrounds that do not adequately represent the complexity of real-world scenarios. Furthermore, due to the diverse origins of the data, there is

a significant variation in image resolution within the SHWD dataset, which may adversely affect the model’s performance in practical applications. In this work, we collect a high-quality and large-scale dataset, SFCHD, that authentically reflects the complexity of industrial scenarios, to facilitate the application of object detection algorithms in the field of industrial safety. Table I shows a comparison between our SFCHD and the extant safety helmet datasets.

B. Object Detection Methods

In the field of computer vision, object detection is a fundamental task for many real-world applications, which are receiving widespread attention. There are many potential application scenarios for object detection, such as autonomous driving [17], [18], medical diagnosis [19], emotion detection [20], pedestrian detection [21], [22], etc. The early object detection algorithms are composed of hand-crafted feature extractors such as Histogram of Oriented Gradients (HOG) [23], Viola Jones detector [24], and so on. These methods have inaccurate detection results and are resistant to be applied to different datasets. In recent years, as the rapid evolution of deep learning, the reintroduction of CNN based models has brought a new landscape to the realm of object detection. CNN-based [25], [26] object detection methods can be categorized into two main types: two-stage methods and one-stage methods. Two-stage algorithms have two separate steps, i.e., searching for object proposals in images during the first stage, and then classifying and localizing them during the second stage. These methods commonly are of complex architecture and demand longer time to generate proposals. Unlike the

TABLE I
COMPARISON BETWEEN SFCHD AND EXISTING OPEN-SOURCE DATASETS FOR SAFETY HELMETS

Dataset	Year	#Category	#Sample	#Instance	Color	Task	Data Source
Pictor-v3	2020	6	1,330	9,208	RGB	Detection	Web-mined and Crowd-sourced
SHWD	2019	2	3,241	10,457	RGB	Detection	Web-mined
SFCHD	2023	7	12,373	50,558	RGB	Detection	Chemical Plant

In the SHWD dataset, only the images related to safety helmets are used. #Category, #Sample, and #Instance denote the numbers of categories, samples, and instances in the dataset, respectively.

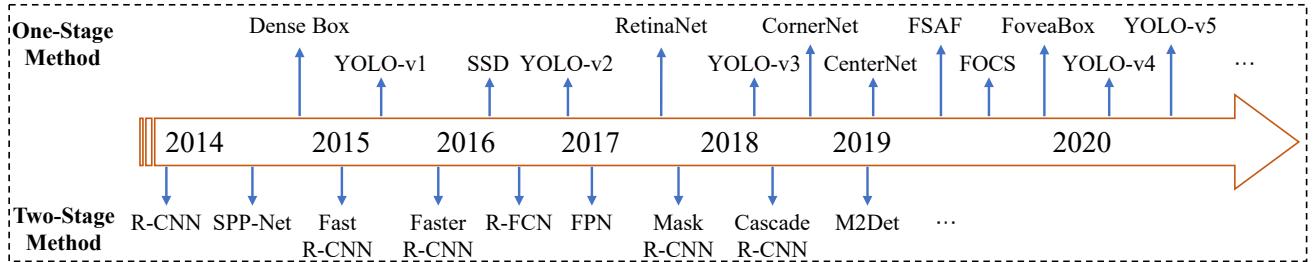


Fig. 2. The developmental trajectory of object detection methods based on deep learning. The upper part illustrates the one-stage methods and the lower part showcases the two-stage methods.

two-stage algorithms, one-stage approaches are superior in terms of real-time performance and network architecture. One-stage algorithms localize objects with the predefined boxes of different scales and aspect ratios, and classify semantic objects in a single shot adopting dense sampling [27].

RCNN [28] is a pioneering work in two-stage object detection. It employed a selective search algorithm to generate 2000 region proposals, then used a CNN to extract features from each proposal, followed by support vector machine (SVM) [29] and finally performed classification and regression. However, the slow inference speed due to the large number of region proposals limits the practicality of the model. Fast RCNN [30] further improved the speed and accuracy of RCNN [28] by sharing features among proposals and using ROI-pooling to extract fixed-size features. Faster RCNN [31] unified the region proposal network and detection network, leading to improved accuracy and inference speed. One-stage methods, such as YOLO [32], directly predicted object categories and bounding boxes in a single network pass. YOLOv2 [33] improved the accuracy and speed by introducing batch normalization and full convolutional operations. YOLOv3 [34] further improved the accuracy by using a more powerful backbone network and multiscale prediction based on Darknet-53. YOLOv4 [35] introduced additional improvements in data preprocessing, network design, and prediction. SSD [36], another one-stage method, detected objects of different scales by using feature maps from different layers. FCOS [37] predicted object boundaries and categories based on internal points and allocated objects to different pyramid levels based on their scales. YOLOv5 introduced some novel improvement ideas based on YOLOv4, which resulted in a significant performance advancement in both speed and accuracy. The development line of object detection algorithms based on deep learning is shown in Fig. 2.

C. Low-Light Object Detection

In low-light environments, the visibility of the image is severely compromised due to poor lighting conditions, greatly impacting the performance of detector. Initial methods for low-light detection focus on enhancing the low-light images before feeding them into detector. For instance, Zhang et al. [38] introduced KinD, which achieves image enhancement through paired training with images at different levels of illumination. Lv et al. [39] proposed a multi-branch low-light enhancement network (MBLEN), which extracts features at various levels and then merges them using a multi-branch structure to produce enhanced images. Liu et al. [40] proposed IA-YOLO, which uses a CNN-based parameter predictor to train the YOLO detector. Qin et al. [41] introduced DE-YOLO, by training the cascaded DENet and YOLOv3 models in an end-to-end manner with the standard detection loss. Cui [42] presented IAT-YOLO, which utilizes both local and global branches, incorporating transformers for image enhancement. Additionally, joint training methods have been employed, where only the loss from detection tasks is used to accomplish low-light detection.

III. PROPOSED SFCHD DATASET

Existing helmet datasets are insufficient to meet the requirements of practical scenarios. In this work, we collect real-world images from chemical plants and construct a high-quality safety clothing and helmets dataset for object detection. In the following, we introduce the construction principles, data details, and potential tasks of the proposed SFCHD dataset.

A. Construction Principle

To guarantee the high quality of the SFCHD dataset, we construct the SFCHD dataset in accordance with the following principles.

TABLE II
STATISTICS OF INSTANCE DISTRIBUTION PER CATEGORY IN THE SFCHD DATASET

Category	Person	Safety Helmet	Safety Clothing	Other Clothing	Head	Blurred Clothing	Blurred Head	Total
Training	13,528	11,378	11,781	626	961	1,053	896	40,223
Testing	3,482	2,920	3,032	154	239	271	238	10,336
Total	17,010	14,298	14,813	780	1,200	1,324	1,134	50,559

Realistic Source: Collecting realistic data is crucial for practical implementation and also makes the related studies more meaningful. Therefore, we collect on-site images of workers from surveillance cameras in two chemical plants based on specific collection criteria. For each camera, we save one image every 10 frames when there were people in the scene. This data collection process is carried out continuously for one week, resulting in approximately 40,000 captured images as the initial data source. Subsequently, we perform data filtering while ensuring an even distribution of scenes and categories. The goal is to avoid significant imbalances in the quantity of different categories. Eventually, we retain 12,373 images as the final dataset. We employ a repetitive filtering approach to ensure the authenticity and uniformity of the data.

Data Privacy: While collecting and labeling images, we strictly follow the standard de-identification procedure by removing any sensitive information (such as time, plant location, etc.), thereby ensuring that the identities of the factory and workers remain confidential.

Diverse Scenes: To ensure a diverse range of scenarios in the dataset, we collect data from two factories, capturing nearly 40 different indoor and outdoor scenes are captured. This approach ensures the dataset's scene universality.

Professional Annotation: Our annotation team consists of six professional security inspectors and the three graduate students who have undergone specialized training. Each instance in SFCHD is annotated by the six professional security inspectors. After annotation, three graduate students who are not involved in the labeling process were invited to review the annotations, ensuring that every image and every annotation in the dataset is completely accurate.

Various Categories: Our SFCHD dataset comprises five common categories observed at construction sites, including Person, Safety Helmet, Safety Clothing, Other Clothing, Head. To ensure the realism of the dataset, we add annotations for objects that are difficult to differentiate. In other words, we annotate Blurred Clothing and Blurred Head for those who are affected by shadows and strong lighting. In this manner, the SFCHD dataset encompasses a total of 7 categories, as depicted in Table II.

High-Low Light Scene Discriminate: As illustrated in Fig. 3, we categorize the entire dataset into three groups: high-light, low-light, and normal-light images, based on the intensity of light sources around person in the images. Images with an evenly distributed light source are classified as normal-light. On the other hand, images with apparent strong light sources or shadows are classified as high-light and low-light,

respectively. Although deep learning-based object detection methods have exhibited impressive results on conventional datasets, detecting objects accurately in low-quality images captured under extreme light conditions remains a significant challenge. Thus, we provide researchers with samples of extreme light conditions to encourage the development of object detection in such scenarios. The classification of all scenes is carried out by the three graduate students. Given the potential for different perspectives on classifying images, a voting system was used to resolve any disagreements, with the majority opinion determining the final classification outcome.

B. Dataset Details

In accordance with the construction principles of the dataset, we amass a large, complex, and realistic dataset for safety clothing and helmets, known as SFCHD. Subsequently, we will provide a description of the SFCHD dataset in terms of instance distribution, data partitioning, and data quality.

Instance Distribution: The SFCHD dataset comprises 12,373 images, 7 categories, and 50,552 annotations. The instance proportion of each category is shown in Fig. 4. As depicted in Fig. 4, the most prominent categories are Person, Safety Helmet, and Safety Clothing, accounting for about 33.64%, 28.28% and 29.30% of the dataset, respectively. Due to the rarity of violations in real-world chemical plants, situations where safety clothing or helmets are not worn are infrequent. As a result, the instances classified as Other Clothing and Head represent only a small portion of the dataset, 1.54% and 2.37%, respectively. The remaining categories, Blurred Clothing and Blurred Head, which represent instances captured in special environments, are relatively limited in quantity, accounting for 2.24% and 2.62%, respectively.

Data Partitioning: Our dataset is divided into training set and testing set, and the ratio is 4:1. We have 12,373 images in SFCHD dataset, including 9,898 images in the training set, 2,475 images in the testing set. The statistics of category distribution of training and testing set are also shown in Table II.

Data Quality: To facilitate model training, all images are stored in JPG format; the average resolution of the images is predominantly 1280 * 704, with the maximum resolution reaching 1920 * 1280.

C. Potential Tasks

In addition to conventional object detection, the SFCHD dataset can also serve as a benchmark for the assessment of



Fig. 3. Examples of images with different lighting conditions. Here, (a)-(c), (d)-(f), and (g)-(i) are the images based on normal-light, high-light, and low-light, respectively. In (d)-(f), it is extremely difficult to detect whether workers are wearing safety clothing and helmets because the surrounding light is too high. Similarly, in (g)-(i), too low light around workers makes it difficult to conduct object detection tasks.

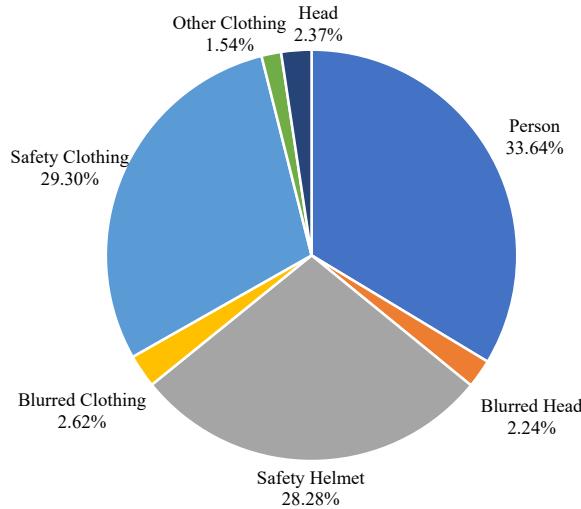


Fig. 4. The proportion of instances per category in the SFCHD dataset.

various extreme detection tasks, such as small object detection and high-low light object detection.

Small Object Detection: As per the definition provided by the international organization SPIE, a small object must not exceed 0.12% of the overall image area. Therefore, we define small objects as those whose actual bounding box occupies

less than 0.1% of the entire image area, while large objects are those whose bounding box occupies more than 0.2% of the image area. The remaining objects are classified as medium-sized objects. Table III provides detailed statistics on the number of small objects in the dataset.

TABLE III
CATEGORY DISTRIBUTION OF OBJECTS WITH DIFFERENT SIZES IN THE SFCHD DATASET

Category	Total	Large	Medium	Small
Safety Helmet	14,298	3,551	3,030	7,717
Head	1,200	95	164	941
Blurred Clothing	1,324	281	313	730
Blurred Head	1,134	15	41	1,078

High-Low Light Detection: A critical reason why object detection tasks are challenging under extreme light conditions is the lack of relevant datasets. According to the light conditions of the area where the object is located, we divide the images in the dataset into high-light image, low-light image and normal-light image. Fig. 3 illustrates the quality of each type of image. The quantity distribution is shown in Fig. 5, and we can conclude that the samples under normal-light are the most, while the samples under high- and low-light are

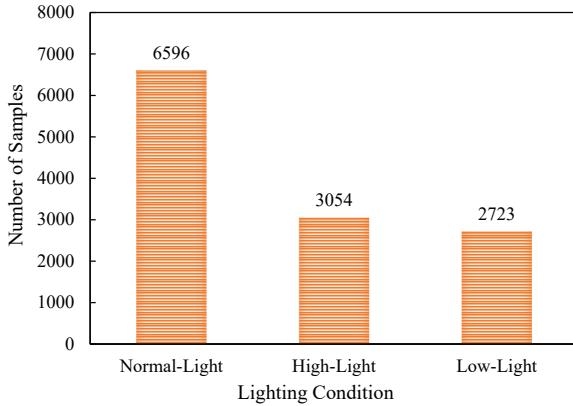


Fig. 5. The quantity distribution under different lighting conditions in the SFCHD dataset.

basically the same.

IV. PROPOSED SCALE MODULE

To address the challenges faced by object detection models in low-light environments, we design a spatial and channel attention-based low-light enhancement (SCALE) module. SCALE, as a plug-and-play component, integrates seamlessly into traditional object detection workflows, as depicted in Fig. 6. Specifically engineered for extreme low-light conditions, the SCALE module adaptively enhances images, consisting primarily of two components, i.e., spatial attention pathway (SAP) and channel attention pathway (CAP). SAP enhances the network's ability to focus on important areas of the image by assigning differentiated weights to pixels in various locations, thereby more effectively understanding and capturing the spatial structure and its correlations within the image. The CAP, on the other hand, guides the network to concentrate on the most informative and significant channels by assigning weights, while suppressing features from redundant channels, further improving the quality and accuracy of feature expression. Features processed by the SCALE module are then fed into the object detector. The entire pipeline relies solely on detection loss for end-to-end training, simplifying the training process and enhancing efficiency. This design not only improves the model's performance under low-light conditions but also maintains the simplicity of model training.

A. Spatial Attention Pathway

For the SAP branch, the low-light image I is initially input into three convolutional kernels of varying sizes to extract features from multiple scales. Subsequently, the obtained feature maps are concatenated along the channel dimension to enrich the feature space. This process can be mathematically represented as:

$$\begin{aligned} F_k &= \text{Conv}_k(I), \quad k \in \{1, 3, 5\}, \\ F_{Conv} &= \text{Cat}[F_1, F_3, F_5], \end{aligned} \quad (1)$$

where I denotes an input with a resolution of $H \times W \times C$; Conv_k denotes the convolutional operation with a kernel of size k , e.g., Conv_1 is the convolution with a kernel of size

1; F_k denotes the corresponding convolution branch output; Cat denotes the concatenation operation across the channel dimension; F_{Conv} denotes the feature map after passing through parallel convolution branches.

The obtained feature map F_{Conv} is then directed into two independent spatial attention units (i.e., SA in Fig. 6). Each Spatial Attention unit primarily consists of pooling, convolution, and element-wise multiplication operations. Initially, we apply global max pooling and global average pooling to the feature map. The global max pooling helps the network to filter the most salient feature, focusing on areas of the low-light image that contain key information while disregarding the effects of darkness and noise. The global average pooling smooths the feature value, making the extracted feature more stable and representative. Subsequently, these two pooled feature maps are concatenated, fed into a convolution for fusion, and normalized through a sigmoid function. Ultimately, the normalized fused feature is element-wise multiplied with each channel of the original feature map, yielding the weighted spatial attention feature maps F_{SA_k} ($k \in \{3, 5\}$). The process above can be mathematically expressed as:

$$F_{SA_k} = F_{Conv} \otimes \sigma(\text{Conv}_k(\text{Cat}[F_{Max}, F_{Avg}])), \quad (2)$$

where F_{Max} and F_{Avg} represent the values after applying max pooling and average pooling to F_{Conv} , respectively; σ represents the sigmoid activation function, and \otimes is the element-wise multiplication. The spatial attention feature maps F_{SA_3} and F_{SA_5} are concatenated, then passed through a convolution of size 3 to produce the output of the SAP branch:

$$\begin{aligned} F'_{Conv} &= \text{Conv}_3(\text{Cat}[F_{SA_3}, F_{SA_5}]), \\ I_{SAP} &= I \odot F'_{Conv}, \end{aligned} \quad (3)$$

where \odot denotes Hadamard multiplication, I_{SAP} represents the output of SAP branch.

B. Channel Attention Pathway

For CAP branch, before entering the channel attention unit (CA in Fig. 6), the low-light image I is first input into a residual block. This residual block primarily consists of convolution, batch normalization, and residual operations, which can be mathematically expressed as:

$$\begin{aligned} \hat{F}_{Conv} &= \text{Conv}_1(I), \\ F_{CBL} &= \text{LR}(\text{BN}(\text{Conv}_1(I))), \\ F_{Res} &= \hat{F}_{Conv} + \text{LR}(\text{BN}(\text{Conv}_3(F_{CBL}))), \end{aligned} \quad (4)$$

where Conv_1 and Conv_3 denote convolution operations with kernel sizes of 1 and 3, respectively; LR represents the LeakyReLU activation function, and BN denotes batch normalization; a residual block comprises two CBL blocks, and each CBL block includes a convolution layer, LeakyReLU function, and batch normalization; F_{CBL_1} signifies the feature map after the first CBL block, and F_{Res} indicates the final output of the residual block.

Inspired by SENet [11], the obtained feature map F_{Res} is then directed into the channel attention unit. This unit primarily consists of pooling, linear transformation, and element-wise multiplication operations. Initially, an average pooling

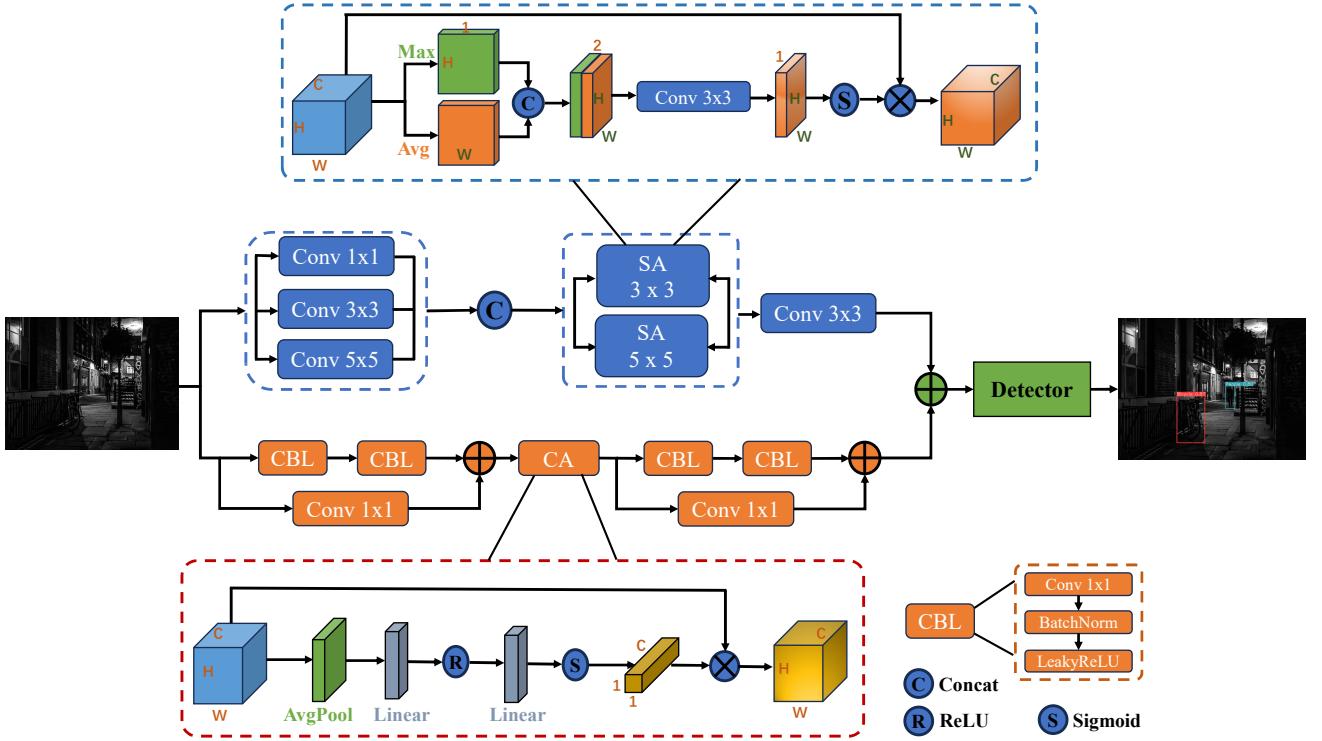


Fig. 6. Network structure of spatial and channel attention-based low-light enhancement (SCALE) module.

operation is employed to down-sample the feature map. This operation not only reduces the complexity of the feature map but also diminishes noise and detail interference, capturing local statistical information. Subsequently, the pooled feature map undergoes two linear transformations and is normalized through a sigmoid function. Ultimately, the resulting feature map is element-wise multiplied with the original feature map across channels to yield the channel attention feature map F_{CA} . This process assigns different weights to each feature channel, enabling the network to learn the most informative feature channels. The computation can be mathematically represented as:

$$F_{CA} = F_{Res} \otimes \sigma(\text{Lin}_2(\text{ReLU}(\text{Lin}_1(F'_{Avg})))), \quad (5)$$

where F'_{Avg} denotes the value after applying the average pooling operation to F_{Res} ; Lin_1 and Lin_2 represent two successive linear transformations. F_{CA} is processed through a residual block to obtain the output of the CAP branch:

$$\begin{aligned} \tilde{F}_{Conv} &= \text{Conv}_1(F_{CA}), \\ \tilde{F}_{CBL} &= \text{LR}(\text{BN}(\text{Conv}_1(F_{CA}))), \\ \tilde{F}_{Res} &= \tilde{F}_{Conv} + \text{LR}(\text{BN}(\text{Conv}_3(\tilde{F}_{CBL}))), \\ I_{CAP} &= I \odot \tilde{F}_{Res}, \end{aligned} \quad (6)$$

where \odot denotes Hadamard multiplication, I_{CAP} denotes the output of CAP branch.

V. EXPERIMENT

Centered around the SFCHD dataset and the SCALE module, our experiments are composed of three parts. Initially, we utilize a variety of classic detection methods to perform

experiments on the existing helmet datasets and our SFCHD dataset, validating the practicality of our dataset. Subsequently, we incorporate the SCALE module into detection algorithms and conduct experiments on the public low-light dataset to confirm the effectiveness of the SCALE module. Finally, we integrate the SCALE module with the SFCHD dataset, demonstrating its capability to significantly improve detection accuracy under challenging conditions.

A. Usability of the SFCHD Dataset

In this subsection, we employ several detection models to conduct experiments on the existing helmet datasets and proposed SFCHD dataset, reporting their outcomes to substantiate the usability of our dataset.

1) *Experimental Setup: Datasets*: We compare the propose SFCHD with the Pictor-v3 [9] and SHWD [10] datasets. Pictor-v3 contains 774 crowd-sourced and 698 web-mined images, where the crowd-sourced and web-mined images contain 2,496 and 2,230 instances of workers, respectively. SHWD provides the dataset used for both safety helmet wearing and human head detection. It includes 7,581 images with 9,044 human safety helmets wearing objects (positive) and 111,514 normal head objects (not wearing or negative).

Methods: We utilize classic detection methods to verify the usability of our dataset, including SSD [36], Faster RCNN [31], RetinaNet [43], FCOS [37], TOOD [44], VFNet [45], YOLOv5¹. SSD was a one-stage, multi-scale detection algorithm that utilized a convolutional pyramid for feature extraction and performed classification and regression across

¹<https://github.com/ultralytics/yolov5>

TABLE IV
COMPARISONS OF DIFFERENT METHODS ON THE PICTOR-V3, SHWD, AND SFCHD DATASETS

Method	Backbone	Pictor-v3		SHWD		SFCHD (ours)	
		mAP(0.50)	mAP(0.50:0.95)	mAP(0.50)	mAP(0.50:0.95)	mAP(0.50)	mAP(0.50:0.95)
SSD	VGG16	85.5	48.8	80.8	57.4	72.8	41.5
Faster RCNN	ResNet-50	90.6	53.4	84.8	63.1	76.4	50.3
FCOS	ResNet-50	89.5	52.4	85.8	63.9	76.4	49.6
VFNet	ResNet-50	91.4	55.2	85.7	63.9	76.4	51.0
RetinaNet	ResNet-50	90.5	54.4	85.5	63.6	75.9	48.9
TOOD	ResNet-50	91.5	55.8	86.7	64.4	78.9	52.3
YOLOv5	CSPDarknet53	88.2	53.6	84.0	63.9	74.1	49.6

feature maps. Faster RCNN was a pioneering two-stage object detection framework that identified potential object regions, or anchor boxes, and then precisely classified and localized the objects within these regions. RetinaNet introduced the Focal Loss to address the class imbalance in one-stage detectors, enhancing performance without altering the underlying network structure. FCOS redefined object detection by offering an anchor-free and proposal-free approach, operating pixel-by-pixel for a streamlined, full convolutional network-based solution. TOOD proposed a task-aligned one-stage detection method with a novel head structure and learning method, focusing on improving detection for specific tasks. VFNet designed the Varifocal Loss function to predict IoU-aware classification scores and introduced a star-shaped bounding box feature representation for refining object localization. YOLOv5 was a one-stage object detection algorithm, improved upon its predecessors by offering superior speed and accuracy in one-stage object detection.

Details: All experiments are conducted using MMDetection², with dataset partitions following the default configuration. The experiments were optimized using the SGD optimizer. The momentum and weight decay are set to 0.9 and 0.0001, respectively. During the training and testing process, the image size was adjusted to 1333 * 800. The batch size is set to 2 for a single GPU, and distributed training is conducted using 4 GPUs. All models are trained on NVIDIA GeForce RTX 3090 GPUs, each with 24GB of graphics memory. We evaluate the mean average precision (mAP) of the object detection to measure the performance of all models fairly. The mAP(0.50:0.95) metric measures the mAP across different Intersection over Union (IoU) thresholds, ranging from 0.50 to 0.95 with an interval of 0.05. Similarly, mAP(0.50) is the mAP at a single IoU threshold of 0.50.

2) *Experimental Results:* Table IV presents a comparative analysis of various methods across different datasets. We primarily rely on two metrics, i.e., mAP(0.50) and mAP(0.50:0.95), to gauge the adaptability of different networks to each dataset. It is observable from the table that our SFCHD records the lowest scores in mAP, followed by the Pictor-V3 and SHWD datasets. This is primarily attributed to the higher complexity of our dataset, which encompasses intricate backgrounds, diverse categories, and undesirable lighting conditions, thereby presenting greater challenges. Undoubtedly, models trained on datasets that more closely mimic real-world scenarios and are more challenging tend to exhibit more effective performance in practical applications.

Table V delineates the performance of distinct categories in our dataset. It is observable from the table that YOLOv5 achieves a heightened level of precision in the identification of the Person, Safety Helmet, and Safety Clothing categories. This elevated accuracy may be attributed to the abundant data representation for these classes, thereby equipping the network with the capacity to absorb a more extensive array of characteristics. In the case of the Other Clothing and Head categories, despite their limited proportion within the dataset, the pronounced divergence in their features from the Safety Clothing and Safety Helmet facilitates the network's capability to distinguish and assimilate these variations. In contrast, the Blurred Clothing and Blurred Head categories, constrained by a scarcity of samples in the dataset and further impeded by complex environmental factors such as shadows and intense lighting, present a formidable challenge for the network to discern efficacious features, thereby leading to diminished discernment efficacy.

In Table VI, we compare the mAP(0.50:0.95) and inference speed of various models on the SFCHD dataset. It can be observed that TOOD achieves the highest mAP, but its FPS is only 17.0; SSD attains the highest FPS, but its mAP is lower than all other models, with a score of only 41.5%. Upon comprehensive consideration, YOLOv5 realizes the optimal balance between accuracy and speed.

B. Validity of the SCALE Module

In this subsection, we integrate the SCALE module with detection algorithms and experiment on the public low-light dataset to validate the efficacy of SCALE.

1) *Experimental Setup: Dataset:* To substantiate the efficacy of the SCALE module, we conduct extensive experiments on the ExDark [46] dataset. The dataset is a benchmark specifically designed to assess object detection algorithms under low-light conditions. It comprises a total of 7,363 low-light images, categorized into 10 varying illumination conditions, ranging from extremely dim to moderately low light, across 12 classes, with a total of 23,710 annotated instances. The

²<https://github.com/open-mmlab/mmdetection>

TABLE V
PERFORMANCE OF DIFFERENT CATEGORIES IN THE SFCHD DATASET

Method	Person	Safety Helmet	Safety Clothing	Other Clothing	Head	Blurred Clothing	Blurred Head	mAP(0.50:0.95)	mAP(0.50)
	AP(0.50:0.95)								
SSD	60.2	56.5	55.7	45.2	38.4	20.5	14.2	41.5	72.8
Faster RCNN	71.2	64.7	64.6	54.5	49.3	27.2	20.5	50.3	76.4
FCOS	68.4	63.4	64.6	54.0	48.0	28.9	19.6	49.6	76.4
VFNet	73.1	66.2	64.3	54.0	52.5	25.1	22.0	51.0	76.4
RetinaNet	71.1	63.5	64.5	51.8	48.5	27.1	15.9	48.9	75.9
TOOD	72.9	66.0	65.9	56.2	52.8	29.6	22.3	52.3	78.9
YOLOv5	72.7	66.4	63.7	54.9	50.7	21.2	18.9	49.6	74.1

TABLE VI
RESULTS OF MAP AND SPEED FOR DIFFERENT METHODS IN THE SFCHD DATASET

Method	Backbone	mAP(0.50:0.95)	FPS
SSD	VGG16	41.5	57.5
Faster RCNN	ResNet-50	50.3	25.9
FCOS	ResNet-50	49.6	27.5
VFNet	ResNet-50	51.0	11.7
RetinaNet	ResNet-50	48.9	27.0
TOOD	ResNet-50	52.3	17.0
YOLOv5	CSPDarknet53	49.6	46.3

dataset is partitioned into a training set with 5,896 images and a testing set with 1,467 images. The images within the dataset encompass a multitude of scenarios from indoor to outdoor settings and from urban to natural landscapes, ensuring the diversity and applicability of the dataset.

Methods: The models we select primarily encompass three categories. The first category consists of low-light image enhancement methods, including KinD [38], MBLLEN [39], and Zero-DCE [47]. KinD, inspired by Retinex theory, decomposed the image into two components, one responsible for light regulation (illumination) and the other for degradation elimination (reflectance). MBLLEN extracted rich features at different levels, enhanced them through multiple sub-networks, and ultimately produced the output image through multi-branch fusion. Zero-DCE formulated the task of light enhancement as a specific curve estimation for images with a deep network, training a lightweight deep network to estimate the dynamic range adjustment of pixels and higher-order curves for a given image. The second category is low-light image object detection methods, including MAET [48], DENet [41], IAT-YOLO [42], and PE-YOLO [49]. MAET, in a self-supervised manner, considered the physical noise model and image signal processing to learn the intrinsic visual structure by encoding and decoding real-world illumination degradation transformations. DENet enhanced the model's performance under adverse weather conditions by using a Laplacian pyramid to decompose each input image into a low-frequency component and several high-frequency components. IAT-YOLO proposed a lightweight

and fast illumination adaptive transformer for normalizing sRGB images under low-light or underexposed/overexposed conditions. PE-YOLO introduced a detail processing module composed of a context branch and an edge branch to enhance image details, as well as a low-frequency enhancement filter to obtain low-frequency semantic information and avoid high-frequency noise. The third category consists of classical one-stage object detection methods, including FCOS [37], VFNet [45], and TOOD [44]. These methods are primarily utilized for cascading with the SCALE module to form new end-to-end models.

Details: The inaugural category of models is trained on the LOL [50] dataset, subsequently employing the trained models to process the ExDark dataset, thereby generating enhanced and brightened imagery. These images are then subsequently input into the YOLOv3 [34] detector to procure the detection outcomes of the low-light image enhancement techniques. In contrast, the second category of models directly integrates low-light images into the network for training, culminating in the ultimate detection results. To ensure a fair comparative analysis, the detector employed across both categories of methods is the YOLOv3. Throughout the training and testing phases, the dimensions of the images are uniformly adjusted to 608 * 608. For training, the batch size for an individual GPU is configured to 4, and distributed training is executed across 4 GPUs. The optimization algorithm utilized during training is SGD, with the learning rate set to 0.001 and the weight decay factor designated as 0.0005. The third category of methods is to validate the enhancement effect of the SCALE module under low-light conditions. We integrate the SCALE module with existing object detection algorithms to obtain the detector's output results. In this category, experiments are conducted based on the MMDetection framework. During training and testing, the size of the images is uniformly adjusted to 1333 * 800. The batch size for a single GPU is set to 4, and distributed training is performed using the same 4 GPUs. The optimizer for training is the SGD, with the learning rate set to 0.01 and the weight decay factor set to 0.0001.

2) *Experimental Results:* In Table VII, we report the performance comparison between the proposed SCALE-YOLO and existing models, encompassing both low-light image enhancement methods and low-light image object detection methods.

TABLE VII
PERFORMANCE COMPARISONS BETWEEN OUR SCALE-YOLO AND EXISTING MODELS ON THE EXDARK DATASET

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP(0.50)
	AP(0.50)												
YOLOv3	79.8	75.3	78.1	92.3	83.0	68.0	69.0	79.0	78.0	77.3	81.5	55.5	76.4
KinD	80.1	77.7	77.2	93.8	83.9	66.9	68.7	77.4	79.3	75.3	80.9	53.8	76.3
MBLLEN	82.0	77.3	76.5	91.3	84.0	67.6	69.1	77.6	80.4	75.6	81.9	58.6	76.8
Zero-DCE	84.1	77.6	78.3	93.1	83.7	70.3	69.8	77.6	77.4	76.3	81.0	53.6	76.9
MAET	83.1	78.5	75.6	92.9	83.1	73.4	71.3	79.0	79.8	77.2	81.1	57.0	77.7
DENet	80.4	79.7	77.9	91.2	82.7	72.8	69.9	80.1	77.2	76.7	82.0	57.2	77.3
IAT-YOLO	79.8	76.9	78.6	92.5	83.8	73.6	72.4	78.6	79.0	79.0	81.1	57.7	77.8
PE-YOLO	84.7	79.2	79.3	92.5	83.9	71.5	71.7	79.7	79.7	77.3	81.8	55.3	78.0
SCALE-YOLO (Ours)	81.3	79.3	78.2	93.9	84.2	75.5	74.9	82.3	81.0	77.5	82.5	57.3	79.0

These models include both low-light image enhancement methods and low-light image object detection methods.

TABLE VIII
PERFORMANCE IMPROVEMENTS OF THE SCALE MODULE ON THE EXDARK DATASET

Method	Bicycle	Boat	Bottle	Bus	Car	Cat	Chair	Cup	Dog	Motorbike	People	Table	mAP(0.50)
	AP(0.50)												
FCOS	75.5	64.4	68.0	86.8	78.5	69.3	55.4	71.7	70.0	64.8	72.3	46.7	68.6
FCOS+SCALE	75.1	66.6	73.5	89.9	78.9	67.0	57.2	72.8	74.2	67.3	72.0	45.9	70.0
VFNet	77.4	70.5	76.6	90.6	81.8	67.1	59.4	71.8	72.6	70.6	77.7	53.3	72.5
VFNet+SCALE	79.4	70.2	76.5	89.7	81.7	71.9	60.9	71.5	75.0	71.2	77.4	55.5	73.4
TOOD	77.0	69.2	72.2	90.0	80.0	72.6	63.0	71.8	71.0	71.9	76.2	52.2	72.3
TOOD+SCALE	77.9	70.0	78.3	90.0	80.7	69.1	62.0	72.4	73.7	69.2	78.1	54.2	73.0

Compared to YOLOv3, SCALE-YOLO demonstrates a 2.6% improvement in mAP. KinD, MBLLEN, and Zero-DCE are image enhancement models aimed at improving image details by enhancing brightness, but they also tend to amplify image noise, affecting feature extraction and representation. Compared to these methods, SCALE-YOLO achieves better detection performance with improvements of 2.7%, 2.2%, and 2.1%, respectively. From the table, it can be discerned that compared to image enhancement methods, object detection methods exhibit superior performance. This indicates that the enhancement networks' comprehension of images is more effective than simply increasing image contrast and brightness. In comparison to MAET, SCALE-YOLO achieves a 1.3% increase in mAP, against DENet, the mAP improvement is 1.7%, compared to IAT-YOLO, the mAP enhancement is 1.2%, and when juxtaposed with PE-YOLO, the mAP increase amounts to 1%. SCALE-YOLO employs spatial attention to learn the correlation of different regions in low-light images and channel attention to learn the correlation of different channels, thereby focusing more on important feature channels.

Table VIII reveals the performance enhancements achieved by integrating the SCALE module into existing single-stage object detection algorithms. The results indicate that on the ExDark dataset, the performance of all three detectors is significantly bolstered. Specifically, the FCOS+SCALE version,

which integrates the SCALE module, improves the mAP by 1.4% compared to the original FCOS; the VFNet+SCALE version, with the addition of the SCALE module, sees a 0.9% increase in mAP; and TOOD+SCALE enhances performance by 0.7% over the original TOOD. These outcomes demonstrate that the SCALE module effectively augments these detectors' ability to comprehend and recognize images under low-light conditions, thereby elevating detection capabilities.

TABLE IX
ABLATION ANALYSIS FOR DIFFERENT PATHWAYS IN OUR SCALE MODULE

Method	SAP	CAP	mAP(0.50)
YOLOv3	-	-	76.4
SCALE-YOLO (Ours)	✓	✗	77.3
SCALE-YOLO (Ours)	✗	✓	77.8
SCALE-YOLO (Ours)	✓	✓	79.0

To validate the efficacy of the spatial and channel attention pathways within the SCALE module, we conduct ablation studies on these two components, as reflected in Table IX. The experimental results indicate that when only the SAP is activated, the model's mAP increased by 0.9% compared to YOLOv3, confirming that SAP can effectively focus on key information in low-light images. When only the CAP is

TABLE X
PERFORMANCE IMPROVEMENTS OF THE SCALE MODULE ON THE SFCHD DATASET

Method	Person	Safety Helmet	Safety Clothing	Other Clothing	Head	Blurred Clothing	Blurred Head	mAP(0.50:0.95)	mAP(0.50)
	AP(0.50:0.95)								
FCOS	68.4	63.4	64.6	54.0	48.0	28.9	19.6	49.6	76.4
FCOS+SCALE	68.5	63.8	64.7	53.0	47.6	28.4	20.5	49.5	76.3
VFNet	73.1	66.2	64.3	54.0	52.5	25.1	22.0	51.0	76.4
VFNet+SCALE	73.2	66.5	64.4	53.9	52.1	25.8	23.7	51.4	76.6
TOOD	72.9	66.0	65.9	56.2	52.8	29.6	22.3	52.3	78.9
TOOD+SCALE	72.9	66.2	66.2	56.2	51.6	29.6	23.5	52.4	79.3
YOLOv8	73.6	64.9	67.5	58.5	45.5	32.3	23.9	52.2	77.9
YOLOv8+SCALE	74.4	66.1	68.8	58.4	47.5	33.2	25.2	53.3	78.6

enabled, the mAP increased by 1.8%, demonstrating that CAP aids the model in identifying and enhancing channels carrying significant information. By comparing the performance improvements of the two, we can infer that inter-channel relationships may be weighted more heavily than spatial location information. When both SAP and CAP are employed jointly, the model's mAP is further enhanced by 2.6%. The aforementioned results substantiate that the SCALE module can generate more discriminative feature representations, thereby facilitating the detector's profound understanding of images in low-light environments.

C. Application of SCALE on SFCHD Dataset

In this subsection, we apply the SCALE module to the SFCHD dataset to demonstrate that SCALE can effectively enhance the detection quality on the SFCHD dataset.

1) *Experimental Setup:* To verify the performance of the SCALE module on the SFCHD dataset, we select a range of object detection models for testing, including FCOS [37], VFNet [45], TOOD [44], and YOLOv8³. We integrate the proposed SCALE module into these detectors to construct an end-to-end low-light object detection model, which is optimized solely by a single loss function. In terms of experimental setup, we continue with the configuration from previous experiments, utilizing distributed training across 4 GPUs and employing the SGD optimizer. The learning rate is uniformly set to 0.01. For the FCOS, VFNet, and TOOD models, the input image size is adjusted to 1333 * 800 pixels, the Batch Size per GPU is 2, and the weight decay factor is 0.0001. For YOLOv8, the image size is adjusted to 640 * 640 pixels, the Batch Size per GPU is increased to 16, and the weight decay factor is set to 0.0005.

2) *Experimental Results:* Table X provides a detailed exposition of the enhancement effects of the SCALE module on object detection performance within the SFCHD dataset. The experimental outcomes demonstrate that the integration of our SCALE module contributes to the improvement of detection accuracy for the VFNet, TOOD, and YOLOv8

detection algorithms. Specifically, the incorporation of the SCALE module into VFNet resulted in a 0.4% increase in mAP(0.50:0.95) compared to the original model; the TOOD variant, upon incorporating the SCALE module, experience a 0.4% enhancement in mAP(0.50); and the amalgamation of the SCALE module with YOLOv8 realized a 1.1% increment in mAP(0.50:0.95). These data robustly confirm the capability of the SCALE module in augmenting the detector's comprehension of low-light image information. However, on the FCOS model, the introduction of the SCALE module did not lead to a significant performance improvement. An analysis of the reasons suggests that this may be due to the fact that, in addition to extreme lighting conditions, images in the SFCHD dataset are also accompanied by a plethora of other types of noise. The confluence of these factors may pose a considerable challenge to the FCOS model, impacting the further enhancement of model performance. Therefore, for datasets that encompass extreme complexity, such as SFCHD, the SCALE module, when combined with high-performing detectors, can better leverage its advantages in low-light enhancement, showcasing its potential and value in the field of object detection.

VI. CONCLUSION

In this work, we focus on the complex environments of real-world construction sites, proposing an industrial dataset and a low-light image enhancement module to advance artificial intelligence research in high-risk construction sites and its application in practical scenarios. Specifically, we have constructed a large-scale, highly complex, and realistic dataset called SFCHD for safety clothing and helmet detection. This dataset encompasses 12,373 images across 40 different scenarios, covering 7 categories with a total of 50,552 annotated instances. These images are sourced from authentic construction sites, ensuring the practical applicability of the data. To validate the usability of the SFCHD dataset, we experimented with various classic object detection algorithms and compared them with other relevant datasets. Moreover, addressing the performance limitations of existing algorithms under low-light conditions, we designed a plug-and-play low-light enhancement module named SCALE. This module extracts valid information from

³<https://github.com/ultralytics/ultralytics>

low-light images in both spatial and channel dimensions, significantly enhancing the performance of object detection models. Experimental results on the ExDarK and SFCHD datasets demonstrate that the introduction of the SCALE module has led to substantial improvements in the performance of object detectors under low-light conditions.

Looking forward, we plan to apply the SFCHD dataset to a broader range of computer vision tasks, such as instance segmentation and image classification, to further expand its application scope and impact. Concurrently, we will continue to explore and propose more innovative and advantageous methods to effectively address challenges like complex background noise in industrial datasets, providing stronger support for computer vision research in industrial settings.

REFERENCES

- [1] A. Kelm, L. Laußat, A. Meins-Becker, D. Platz, M. J. Khazaee, A. M. Costin, M. Helmus, and J. Teizer, "Mobile passive radio frequency identification (RFID) portal for automated and rapid control of personal protective equipment (PPE) on construction sites," *Automation in Construction*, vol. 36, pp. 38–52, 2013.
- [2] B. Naticchia, M. Vaccarini, and A. Carbonari, "A monitoring system for real-time interference control on large construction sites," *Automation in Construction*, vol. 29, pp. 148–160, 2013.
- [3] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.
- [4] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.
- [5] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [6] Z. Kolar, H. Chen, and X. Luo, "Transfer learning and deep convolutional neural networks for safety guardrail detection in 2d images," *Automation in Construction*, vol. 89, pp. 58–70, 2018.
- [7] N. Nath, T. Chaspary, and A. Behzadan, "A transfer learning method for deep neural network annotation of construction site imagery," in *Proceedings of the 18th International Conference on Construction Applications of Virtual Reality*, 2018, pp. 1–10.
- [8] N. D. Nath, T. Chaspary, and A. H. Behzadan, "Single- and multi-label classification of construction objects using deep transfer learning methods," *J. Inf. Technol. Constr.*, vol. 24, pp. 511–526, 2019.
- [9] N. D. Nath, A. H. Behzadan, and S. G. Paal, "Deep learning for site safety: Real-time detection of personal protective equipment," *Automation in Construction*, vol. 112, p. 103085, 2020.
- [10] M. Gochoo, "Safety helmet wearing dataset," in *Mendeley Data*, 2021.
- [11] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.
- [12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [13] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, 2015.
- [14] 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*, 2014, pp. 740–755.
- [15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [16] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Mallochi, A. Kolesnikov, T. Duerig, and V. Ferrari, "The open images dataset v4," *International Journal of Computer Vision*, vol. 128, no. 7, pp. 1956–1981, 2020.
- [17] D. Feng, A. Harakeh, S. L. Waslander, and K. Dietmayer, "A review and comparative study on probabilistic object detection in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 9961–9980, 2022.
- [18] S. Wang, Y. Sun, Z. Wang, and M. Liu, "ST-TrackNet: A multiple-object tracking network using spatio-temporal information," *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 284–295, 2024.
- [19] R. Elakkiya, V. Subramaniyaswamy, V. Vijayakumar, and A. Mahanti, "Cervical cancer diagnostics healthcare system using hybrid object detection adversarial networks," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1464–1471, 2022.
- [20] J. Li, X. Wang, G. Lv, and Z. Zeng, "GA2MIF: Graph and attention based two-stage multi-source information fusion for conversational emotion detection," *IEEE Transactions on Affective Computing*, vol. 15, no. 1, pp. 130–143, 2024.
- [21] J. U. Kim, S. Park, and Y. M. Ro, "Uncertainty-guided cross-modal learning for robust multispectral pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1510–1523, 2022.
- [22] S. Du, P. Cai, T. Hu, and T. Ikenaga, "Automatic foreground detection at 784 fps for ultra-high-speed human-machine interactions," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 4, pp. 3587–3600, 2022.
- [23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893 vol. 1.
- [24] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.
- [25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
- [26] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.
- [27] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, vol. 126, pp. 1–19, 2022.
- [28] 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*, June 2014.
- [29] J. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft, Tech. Rep. MSR-TR-98-14, April 1998.
- [30] R. Girshick, "Fast R-CNN," in *Proceedings of the 2015 IEEE International Conference on Computer Vision*, December 2015.
- [31] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
- [32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.
- [33] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017.
- [34] ———, "YOLOv3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.
- [35] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," *ArXiv*, vol. abs/2004.10934, 2020.
- [36] 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*, 2016, pp. 21–37.
- [37] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2019.
- [38] Y. Zhang, J. Zhang, and X. Guo, "Kindling the darkness: A practical low-light image enhancer," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1632–1640.
- [39] F. Lv, F. Lu, J. Wu, and C. S. Lim, "MBLLEN: Low-light image/video enhancement using cnns," in *British Machine Vision Conference*, 2018.

- [40] W. Liu, G. Ren, R. Yu, S. Guo, J. Zhu, and L. Zhang, "Image-adaptive yolo for object detection in adverse weather conditions," in *AAAI Conference on Artificial Intelligence*, 2021.
- [41] Q. Qin, K. Chang, M. Huang, and G. Li, "DENet: Detection-driven enhancement network for object detection under adverse weather conditions," in *Proceedings of the Asian Conference on Computer Vision (ACCV)*, December 2022, pp. 2813–2829.
- [42] Z. Cui, K. Li, L. Gu, S. Su, P. Gao, Z. Jiang, Y. Qiao, and T. Harada, "You only need 90k parameters to adapt light: a light weight transformer for image enhancement and exposure correction," in *Proceedings of the 33rd British Machine Vision Conference*, 2022.
- [43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, Oct 2017, pp. 2980–2988.
- [44] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2021, pp. 3490–3499.
- [45] H. Zhang, Y. Wang, F. Dayoub, and N. Sunderhauf, "VarifocalNet: An iou-aware dense object detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2021, pp. 8514–8523.
- [46] Y. P. Loh and C. S. Chan, "Getting to know low-light images with the exclusively dark dataset," *Computer Vision and Image Understanding*, vol. 178, pp. 30–42, 2019.
- [47] C. Guo, C. Li, J. Guo, C. C. Loy, J. Hou, S. Kwong, and R. Cong, "Zero-reference deep curve estimation for low-light image enhancement," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1777–1786.
- [48] Z. Cui, G.-J. Qi, L. Gu, S. You, Z. Zhang, and T. Harada, "Multitask AET with orthogonal tangent regularity for dark object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021, pp. 2553–2562.
- [49] X. Yin, Z. Yu, Z. Fei, W. Lv, and X. Gao, "PE-YOLO: Pyramid enhancement network for dark object detection," in *Artificial Neural Networks and Machine Learning – ICANN 2023*, 2023, pp. 163–174.
- [50] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *Proceedings of the 29th British Machine Vision Conference*, 2018, pp. 1–12.



Fusheng Yu received the BE degree in automation from Zhengzhou University, Henan, China, in 2021. He is currently working toward the MS degree in electronic information with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China, and the Hubei Key Laboratory of Brain-inspired Intelligent Systems. His current research interests include object detection, image retrieval, and self-supervised learning.



Jiang Li is a doctoral candidate in intelligence science and technology (AI), studying at the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), the Institute of Artificial Intelligence, HUST, the Hubei Key Laboratory of Brain-inspired Intelligent Systems, HUST, and the Key Laboratory of Image Processing and Intelligent Control (HUST), Ministry of Education, Wuhan, China. He has received the bachelor's degree in software engineering in 2018, and master's degree in computer technology in 2021. His current research interests include affective computing, multimodal learning, and graph-based learning.



Xiaoping Wang (Senior Member, IEEE) received the B.S. and M.S. degrees in automation from Chongqing University, Chongqing, China, in 1997 and 2000, respectively, and the Ph.D. degree in systems engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003. Since 2011, she has been a Professor with the School of Artificial Intelligence and Automation, HUST, the Hubei Key Laboratory of Brain-inspired Intelligent Systems, HUST, and the Key Laboratory of Image Processing and Intelligent Control (HUST), Ministry of Education, Wuhan, China. Her current research interests include affective computing, memristors and its applications to memory storage, modeling, and simulation.



Shaojin Wu received the BE degree in automation from Zhengzhou University, Henan, China, in 2021. He is currently working toward the MS degree in electronic information with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China, and the Hubei Key Laboratory of Brain-inspired Intelligent Systems, China. His current research interests include object detection, image generation, and multimodal



Junjie Zhang received the BE degree in measurement and control technology and instrumentation from Northeastern University at Qinhuangdao, Hebei, China, in 2021. He is currently working toward the MS degree in control science and engineering with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology (HUST), the Key Laboratory of Image Processing and Intelligent Control of the Education Ministry of China, and the Hubei Key Laboratory of Brain-inspired Intelligent Systems, China. His current research interests include object detection, keypoint detection.



Zhigang Zeng (Fellow, IEEE) received the Ph.D. degree in systems analysis and integration from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2003. He is currently a Professor at the School of Artificial Intelligence and Automation, HUST, the Hubei Key Laboratory of Brain-inspired Intelligent Systems, HUST, and the Key Laboratory of Image Processing and Intelligent Control (HUST), Ministry of Education, Wuhan, China. He has published over 300 international articles. His current research interests include neural networks, computational intelligence, stability analysis of dynamic systems, and pattern recognition, and associative memories.

Dr. Zeng has been a member of Editorial Boards of Neural Networks since 2012, Cognitive Computation since 2010, and Applied Soft Computing since 2013. He was an Associate Editor of IEEE Transactions on Neural Networks from 2010 to 2011. He has been an Associate Editor for the IEEE Transactions on Cybernetics since 2014 and the IEEE Transactions on Fuzzy Systems since 2016.