Enhancing Object Detection in Distorted Environments: A Seamless Integration of Classical Image Processing with Deep Learning Models

Abbass Zein-Eddine, Oussama Messai, Abdelouahid Bentamou, Yann Gavet

Enhancing Object Detection in Distorted Environments: A Seamless Integration of Classical Image Processing with Deep Learning Models

1st Abbass Zein-Eddine
Centre SPIN, CNRS UMR 5307 LGF
Mines Saint-Etienne, Univ Lyon
42023 Saint-Etienne, France
abbass.zeineddine@emse.fr

2nd Oussama Messai
Centre SPIN, CNRS UMR 5307 LGF
Mines Saint-Etienne, Univ Lyon
42023 Saint-Etienne, France
oussama.messai@emse.fr

3rd Abdelouahid Bentamou
Centre SPIN, CNRS UMR 5307 LGF
Mines Saint-Etienne, Univ Lyon
42023 Saint-Etienne, France
abdelouahid.bentamou@emse.fr

4th Yann Gavet
Centre SPIN, CNRS UMR 5307 LGF
Mines Saint-Etienne, Univ Lyon
42023 Saint-Etienne, France
gavet@emse.fr

Abstract*—Computer vision tasks are directly influenced by the conditions of image acquisition, especially in the context of object detection. Often, these conditions are beyond our control. In this paper, we introduce a method that seamlessly integrates with any computer vision model using deep learning to enhance its performance in distorted environments. Our method effectively mitigates the effects caused by various types of image distortions. It relies on classical image processing techniques capable of reducing distortions and enhancing image quality in a general manner, without requiring specific knowledge of the applied distortion type. Integration into any model during the preprocessing stage is straightforward. Furthermore, we've added new layers that analyze the enhanced image in a depth-wise manner, running in parallel with the model backbone. We tested the method on the object detection task using the well-known computer vision model, You Only Look Once (YOLO), and the results reveal a significant improvement in Mean Average Precision (mAP). *The implementation code can be found at: https://github.com/abbass-zain-eddine/Object-detection-under-uncontrolled-acquisition-environment.git

***Index Terms*—Object detection, image distortion, YOLO**

I. Introduction

In recent years, object detection has become an increasingly important problem in computer vision, with numerous applications in fields such as robotics, surveillance, and self-driving cars. Object detection algorithms typically rely on the detection of features or patterns in the image that correspond to objects of interest. However, object detection can be challenging when dealing with distorted or low-contrast images, where objects may be difficult to discern from their background. Such images can arise due to factors such as lighting conditions, object movement, or camera angle.

In the literature, this problem is tackled in several ways. Some propose a separate method for denoising the images before feeding them to the model [1]. This way of solving the problem, opened the door for an intensive work in the domain of image denoising. Furthermore, the emergence of transformers [2] in 2017 and its hit in the domain of natural language processing and computer vision, a plenty of research was done to integrate it in image denoising models [3], [4], [5], and [6]. Others try to enhance the features extracted from the noisy images using a separate feature enhancement network [7]. In [8] the authors propose a fully convolutional neural network which jointly denoise the input maps by learning edges and contrast details, followed by learning of residing salient details via colour spatial maps in an end-to-end fashion. Many application were targeted towered specific type of applications such as underwater images [9], drone detection [10]and vehicle detection [11] [12]. To address this challenge, image pre-processing techniques can be applied to enhance the visibility of objects in an image. These techniques aim to improve the image quality by adjusting the brightness, contrast, and color balance of the image, thereby enhancing the features that correspond to the objects of interest. Histogram equalization is one such technique that is commonly used to improve the contrast of an image by spreading out the pixel intensity values across the entire dynamic range. Other techniques, such as gamma correction, and logarithmic correction, can also be used to adjust the image's brightness and contrast. In the context of object detection, pre-processing techniques can enhance the visibility of objects in an image and improve detection performance.

In this article, we explore the use of pre-processing techniques to enhance object detection in distorted images. Specifically, we use a combination of histogram equalization and other corrections to enhance the grayscale image of the input image. Then, the enhanced grayscale image can be concatenated as a fourth channel along with the RGB image and fed to a state-of-the-art object detection model, such as $YOLO$ [13]. Furthermore, we use a well designed block to extract features

in a depth-wise manner, and finally concatinate those features to those extracted by the model backbone. The performance of this approach is tested on a benchmark object detection dataset, comparing it to conventional RGB-only approaches. Our experiments demonstrate that the enhanced grayscale image can significantly improve object detection performance in distorted images, outperforming conventional RGB-only approaches. We believe that this approach has the potential to enable more accurate object detection in challenging real-world scenarios, where images are often distorted due to several factors.

This work makes three contributions, which are stated below:

- We present a general methodology that does not require any knowledge about the distortion type, to enhance the performance of object detection models under an uncontrolled acquisition environment.
- We extend the use of depth-wise convolution followed by deformable convolution for object detection problems to reduce the effect of distorted channels on the other channels during convolution.
- The proposed method can be easily added in parallel to the backbone of the used object detection model to leverage its performance.

The remainder of this paper is organized as follows. In Section II, we describe the proposed approach in detail. In Section III, we present the experimental results and evaluate the performance of our approach. Finally, in Section IV, we conclude the paper and discuss future research directions.

II. Materials and methods

In this section, we describe the considered dataset and the proposed method for the enhancement of object detection in distorted mediums.

A. Dataset

The dataset used in this study, called CD-COCO [14] is derived from the original well known COCO dataset [15].It contains the same 123k images but with additional dedicated distortions applied to them. The choice of the distortion type is correlated to the scene type (indoor/outdoor) and the scene context (the objects present and the scene depth). Likewise, the distortion severity level would be assigned according to the object type and position (pixel and depth) for local distortions or atmospheric distortions (rain and haze). Fig. 1 shows some images of the CD-COCO dataset. The dataset is composed of 123K images split into three sets, respectively the training set with 95K images, the validation set with 5K images, and the test set with 23K images. The CD-COCO dataset includes a variety of local and global distortions that are applied to objects or specific areas within an observed scene. Local distortions, such as blur motion, defocus blur, and backlight illumination, are weighted and adjusted according to the object's position in the scene, taking into account both its 2D spatial position and depth. The database also contains global distortions resulting from camera parameters and acquisition conditions, such as noise sensitivity, defocus or instabilities, atmospheric turbulence, and image artifacts. The dataset considers various atmospheric and weather factors, including rain and haze phenomena, which can impact image acquisition quality. Other limitations related to camera sensors, such as noise sensitivity, contrast sensitivity, and spatial resolution, are also included. Global blur may be caused by camera motion and/or optical defocus, while local motion blur is the result of moving objects.

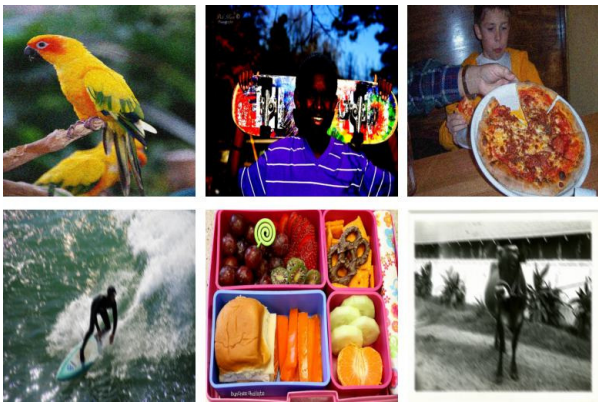

Fig. 1. Example of images from the CD-COCO dataset. The images suffer from noise due to acquisition environment.

B. Proposed method

1) Methodology: The idea behind our method is to be able to propose a block that can be easily added to several deep learning object detection models in the literature and that will enhance their performance. In addition this block should be able to function regardless of the context of the images and the type of the distortions. Deep learning models can be seen as multiple interconnected layers that are responsible for transforming input data into meaningful output predictions. The backbone, the neck, and the head represent the main components of these models, each with a unique role in the overall architecture. The backbone, also known as the feature extractor, is responsible for extracting high-level features from the input data, such as edges, corners, and textures. It typically consists of a series of convolutional layers that learn to recognize patterns in the input data and encode them as feature maps. The neck, also known as the feature fusion layer, is responsible for combining the features extracted by the backbone into a single, more robust representation. This is typically achieved using operations such as concatenation, addition, or multiplication, and helps to improve the model's ability to generalize to new data. The head, also known as the output layer, is responsible for producing the final predictions based on the features generated by the backbone and fused by the neck. The head can take many different forms, depending on the specific task being performed, such as classification, segmentation, or object detection. It typically consists of one or more fully connected layers that map the features to the output space. It is easy to prove that any

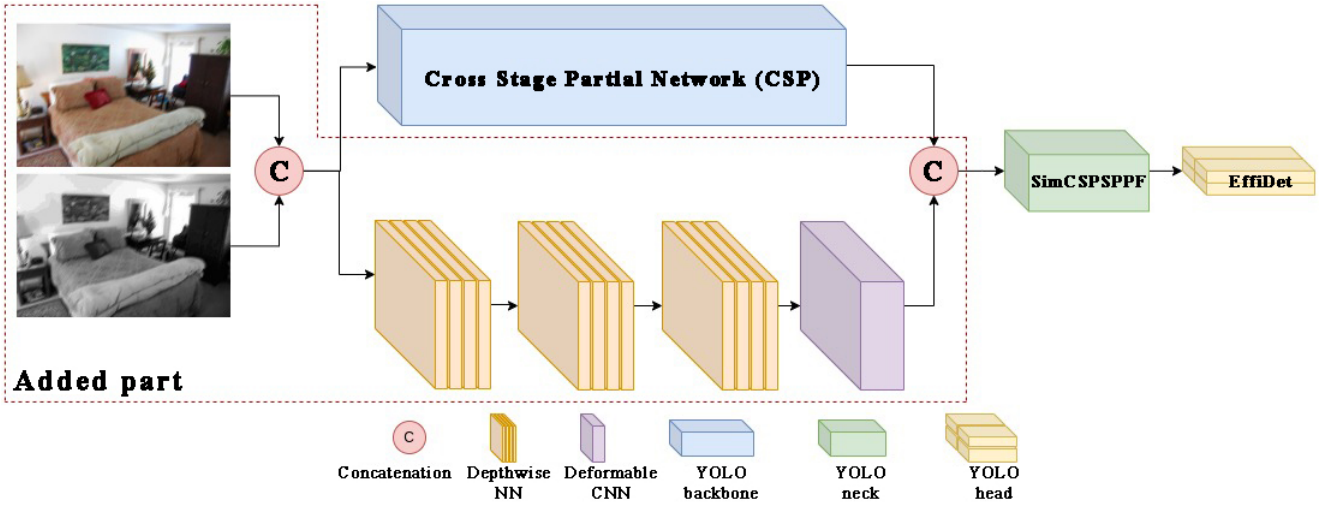

Fig. 2. The structure of the YOLOv6 large model version after fusing it with our method. the Cross Stage Netwrok used as backbone, Simplified CSP with spatial pixel pair feature [16] [17] as a neck and EffiDet head [18]

modification at the level of the backbone will influence the whole model. Consequently, our method aims to perform two main tasks. First, it uses classical image processing to generate an enhanced gray scale image from the original RGB image, and this is performed at the level of pre-processing. Second, it adds a block of layers that extract features separately from each of the input channels, and this is performed at the level of the model backbone, see Fig. 2.

2) Image enhancement by classical methods: To achieve our goal, we employed several image enhancement techniques to generate an enhanced grayscale image from RGB images, which we then concatenated as a fourth layer to our $YOLO$ model. Note that in the equations describing the three used methods I will represent the input image. First, we used logarithmic transformation to the input image (1):

$$Log(I) = c \log(1 + I) \tag{1}$$

where c is a constant factor, and $\log$ is the natural logarithm function. This formula compresses the pixel values in the image such that the dark pixels are expanded more than the bright pixels, which can reveal details in the darker regions of the image.

We also performed gamma correction on the input image using the equation (2):

$$Gamma(I) = I^{\gamma} \tag{2}$$

where γ is the gamma value. A gamma value greater than 1 can make the image brighter, while a gamma value less than 1 can make the image darker. This formula raises the pixel values of the image to a power of gamma to achieve the desired effect.

Finally, we performed histogram equalization on the input image using the equation (3):

$$HistEqu(I) = \frac{L-1}{N} \sum_{i=0}^{N-1} h(i) \sum_{j=0}^{i} p_r(j) \tag{3}$$

where L is the number of possible pixel values (in our case, 256 for 8-bit grayscale images), N is the total number of pixels in the image, $h(i)$ is the histogram of pixel values up to intensity i, and $p_r(j)$ is the normalized cumulative distribution function of pixel intensities. This formula redistributes the pixel values in the image such that the histogram of the output image has a uniform distribution, which can improve its contrast.

All those equations are applied sequentially each one on the output of the previous one as shown in equation (4), where O is the final output enhanced grayscale image.

$$O = HistEqu(Gamma(Log(I))) \tag{4}$$

By concatenating the enhanced grayscale image as a fourth channel to the RGB image and feeding it to our $YOLO$ model, we were able to improve its performance in object detection tasks by enhancing the visibility of important features in the input image.

3) Depthwise feature extraction block: The second part of the proposed method is to treat each channel of the new four channel input image separately. The motivation behind that can be seen from two points of view. First, in some distortion types each layer from the RGB is differently affected. That is to say, separating the channels during the convolution stage will lead to the reduction of the effect of the noise on some of the channels. Thus by applying several depth-wise convolution layers the model will highlight the features of each layer alone. Second, we will also have the opportunity to convolve on the fourth layer which is the enhanced gray scale image separately and extract features also with a lower noise environment. Moreover, the same four channels image is fed to the original model backbone which will convolve on the four channels together normally. Finally, the depth-wise convolution layers are followed by the deformable convolution layer which will join all the results. The use of the deformable convolution will increase spatial flexibility, improve the robustness to occlusion

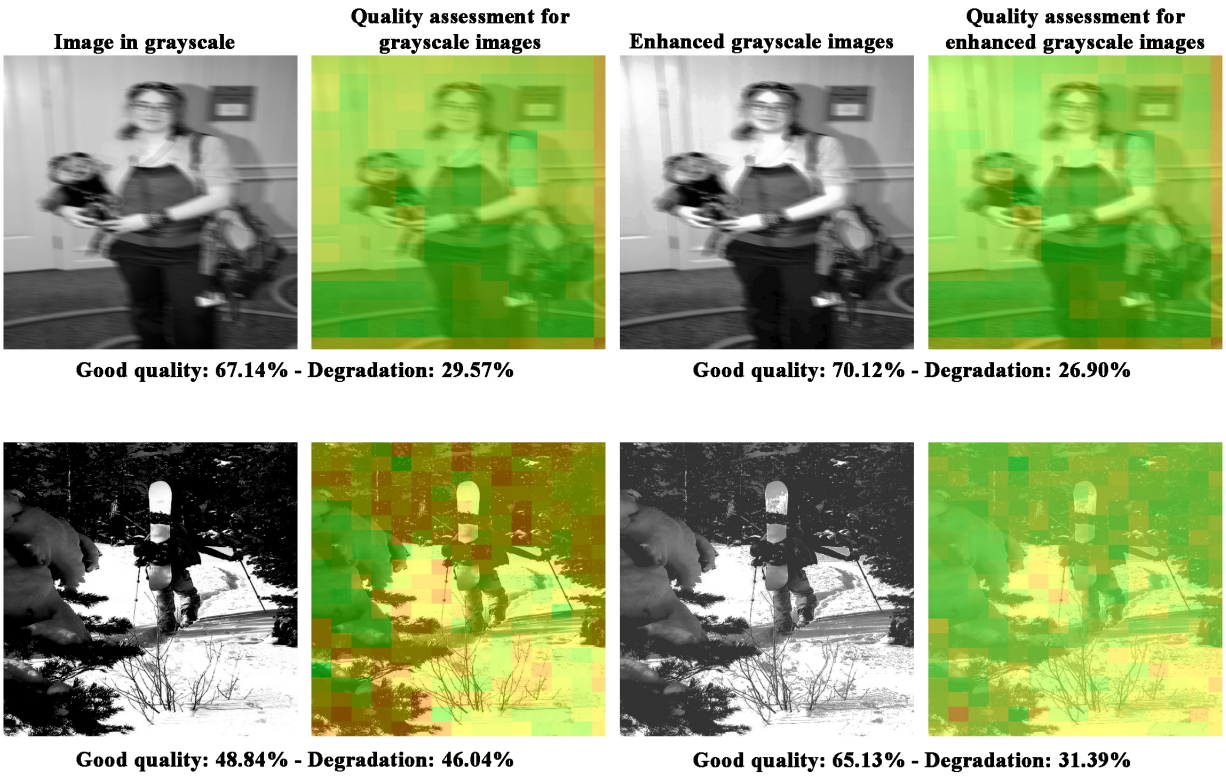

Fig. 3. Quality assessment [19] for grayscale images directly extracted from the CD-COCO dataset images versus the enhanced grayscale images used as fourth channel. Closer to red indicates poor quality, whereas closer to green indicates better quality.

and give a better feature representation than using traditional convolution layers. The feature map extracted by this block will be concatenated with the feature map extracted by the model backbone to create the final feature map that will be sent to the model neck.

III. Experimental results

To demonstrate the effect of our idea, two assessment steps are performed. First step is to see if the image quality has really improved or not. To accomplish this, we used our previous work for image quality assessment metric in [19]. The metric is end-to-end multi-score model that assess the quality and degradation of a given image. We compare the images before and after degradation as follows: one is the grayscale image retrieved straight from the noisy RGB image, and the other is the enhanced grayscale image O. Both images are fed into the model proposed in [19], and two scores are produced for each image: $quality\ score$ and $degradation\ score$. The higher $quality\ score$, the better. Figure 3 depicts the outcomes of two examples. Overall, the $degradation\ score$ before and after enhancement is reduced by up to 10%, where the $quality\ score$ improved by up to 20%. The quality metric divides the input image into 32 by 32 patches, assigning quality/degradation values to each patch. The scores are normalized (min-max normalization) then categorized into colors, with close to green indicating high quality and close to red indicating poor quality. The second assessment step is to see if those enhancements will boost the accuracy of the $YOLO$ model. To demonstrate this, two YOLO versions are selected: $yolov5$ and $yolov6$ are selected, note that the version selected of $yolov6$ [20] is the current state of the art and the one that performs better than $yolov8$. First, all the images are resized to (640,640,3) where 3 is the RGB channels. Then, the enhanced grayscale image O generated and concatenated on the channel axis to the resized image to get a tensor of shape (640,640,4). Finally, this tensor is fed as an input to the $yolov5$ model (same for yolov6). The models are trained on 95k images and evaluated on the 5k validation images. On the other hand both yolo versions are trained on regular RGB images with distortion from the CD-COCO dataset, and they will represent our reference. The evaluation metric used is the Mean Average Precision(mAP).

TABLE I
The results of the evaluation of YOLOv5 and YOLOv6 after and before enhancement.

Model	mAP@0.5(%)	mAP@0.5:0.95(%)
YOLOv5	68.1	48.2
YOLOv6	67.6	50
YOLOv5++	**70.3**	**49.7**
YOLOv6++	**69.5**	**51.4**

The models are evaluated on the same machine, Dell Precision 5570 laptop equipped with an Intel i7-12800H CPU @ 4.80GHz processor and an NVIDIA Quadro RTX

A 2000 GPU. In terms of run-time speed, no significant change was recorded after adding our proposed method to both versions of YOLO during inference. The table I shows the improvement of the mAP for both models after adding our enhancement. Let us denote by YOLOv5++ and YOLOv6++ the enhanced YOLOv5 and YOLOv6 respectively. YOLOv5++ $mAp@0.5:0.95$ is 1.5% greater than that of YOLOv5, and YOLOv6++ $mAp@0.5:0.95$ is 1.3% greater than that of YOLOv6. The $mAp@0.5$ metric also shows an improvement of about 2% in both cases.

IV. Conclusion

In this paper we introduce a new method to enhance computer vision models in object detection tasks in distorted environments. The method uses general enhancements that do not require previous knowledge about the noise type. In addition, it extracts features in a depth-wise manner and then joins those features using a deformable convolution layer. The method is tested on YOLOv5 and YOLOv6 models and the experimental results show an improvement in the both model's mAP. Note that we have added just one block parallel to the model backbone that consists of three depthwise layers and one deformable convolution layer. Based on the results, We believe that it is worthwhile to add additional blocks. For future work, this method can be more investigated and evaluated on other types of computer vision tasks.

V. Acknowledgments

The authors would like to express their sincere gratitude to Bpifrance for their valuable support and funding. This work was conducted as part of the UDD project (Usines De Demain) in collaboration with Bpifrance, contributing to advancements in object detection under an uncontrolled acquisition environment.

References

[1] S Milyaev and I Laptev, "Towards reliable object detection in noisy images," *Pattern Recognition and Image Analysis*, vol. 27, pp. 713–722, 2017.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," 2017.

[3] Chao Yao, Shuo Jin, Meiqin Liu, and Xiaojuan Ban, "Dense residual transformer for image denoising," *Electronics*, vol. 11, no. 3, 2022.

[4] Kang Xu, Weixin Li, Xia Wang, Xiaoyan Hu, Ke Yan, Xiaojie Wang, and Xuan Dong, "Cur transformer: A convolutional unbiased regional transformer for image denoising," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 3, feb 2023.

[5] Achleshwar Luthra, Harsh Sulakhe, Tanish Mittal, Abhishek Iyer, and Santosh Yadav, "Eformer: Edge enhancement based transformer for medical image denoising," 2021.

[6] Tao Xue and Pengsen Ma, "Tc-net: transformer combined with cnn for image denoising," *Applied Intelligence*, vol. 53, no. 6, pp. 6753–6762, 2023.

[7] Geonsoo Lee, Sungeun Hong, and Donghyeon Cho, "Self-supervised feature enhancement networks for small object detection in noisy images," *IEEE Signal Processing Letters*, vol. 28, pp. 1026–1030, 2021.

[8] Maheep Singh, Mahesh C Govil, Emmanuel S Pilli, and Santosh Kumar Vipparthi, "Sod-ced: salient object detection for noisy images using convolution encoder–decoder," *IET Computer Vision*, vol. 13, no. 6, pp. 578–587, 2019.

[9] Long Chen, Feixiang Zhou, Shengke Wang, Junyu Dong, Ning Li, Haiping Ma, Xin Wang, and Huiyu Zhou, "Swipenet: Object detection in noisy underwater scenes," *Pattern Recognition*, vol. 132, pp. 108926, 2022.

[10] Kaiwen Ding, Xianjiang Li, Weijie Guo, and Liaoni Wu, "Improved object detection algorithm for drone-captured dataset based on yolov5," in *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2022, pp. 895–899.

[11] Panumate Chetprayoon, Theerat Sakdejayont, and Monchai Lertsutthiwong, "Object-based vehicle color recognition in uncontrolled environment," in *Proceedings of the 2023 6th International Conference on Machine Vision and Applications*, 2023, pp. 88–94.

[12] Shuangjiang Du, Pin Zhang, Baofu Zhang, and Honghui Xu, "Weak and occluded vehicle detection in complex infrared environment based on improved yolov4," *IEEE Access*, vol. 9, pp. 25671–25680, 2021.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," 2016.

[14] Ayman Beghdadi, Azeddine Beghdadi, Malik Mallem, Lotfi Beji, and Faouzi Alaya Cheikh, "Cd-coco: A versatile complex distorted coco database for scene-context-aware computer vision," in *2023 11th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 2023, pp. 1–6.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[16] Jocher Glenn., "Yolov5 release v6.1.," https://github.com/ultralytics/yolov5/releases/tag/v6.1, 2022.

[17] Lingyan Ran, Yanning Zhang, Wei Wei, and Qilin Zhang, "A hyperspectral image classification framework with spatial pixel pair features," *Sensors*, vol. 17, no. 10, pp. 2421, 2017.

[18] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10781–10790.

[19] Oussama Messai and Aladine Chetouani, "End-to-end deep multi-score model for no-reference stereoscopic image quality assessment," in *2022 IEEE International Conference on Image Processing, ICIP 2022,Bordeaux, France, 16-19 October 2022*. 2022, pp. 2721–2725, IEEE.

[20] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu, "Yolov6 v3.0: A full-scale reloading," 2023.