MOD: Benchmark for Military Object Detection

Xin Yi*, Jiahao Wu*, Bo Ma†, Yangtong Ou, Longyao Liu

***Abstract*—Object detection is widely studied in computer vision filed. In recent years, certain representative deep learning based detection methods along with solid benchmarks are proposed, which boosts the development of related researchs. However, there is no object detection benchmark targeted at military field so far. To facilitate future military object detection research, we propose a novel, publicly available object detection benchmark in military filed called MOD, which contains 6,000 images and 17,465 labeled instances. Unlike previous benchmarks, objects in MOD contain unique challenges such as camouflage, blur, inter-class similarity, intra-class variance and complex military environment. Experiments show that under above chanllenges, existing detection methods suffer from undesirable performance. To address this issue, we propose LGA-RCNN which utilizes a loss-guided attention (LGA) module to highlight representative region of objects. Then, those highlighted local information are fused with global information for precise classification and localization. Extensive experiments on MOD validate the effectiveness of our method.**

***Index Terms*—Object detection, benchmark, representative region highlight.**

I. Introduction

OBJECT detection is a fundamental problem in computer vision, which can be applied in instance segmentation, scene understanding, pose estimation, image captioning and multiple objects tracking (MOT), to name a few. Given an arbitrary image, the goal of object detection is to determine the presence of the predefined categories and locate them in this image. Recently, with the development of convolutional neural network, learning based object detection methods have achieved remarkable progress beyond the traditional detection methods. Meanwhile, in order to train and evaluate the performance of different detection models, certain solid benchmarks for object detection have also been proposed by researchers.

In the modern military filed, object detection plays an important role in military intelligence and has received increasing attention. However, to our best knowledge, no specific benchmark designed for military object detection is proposed so far. Thus, due to this restriction, previous military object detection methods mainly focus on sensor information processing [1], transfer learning [2] while the method of directly using deep learning has been studied very little. To promote the development of this research field, we propose a benchmark for military object detection (MOD) in this work, which consists of 6,000 images with 17,465 labeled instances.

The authors are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. (Email: yixin@bit.edu.cn; wujiahao@bit.edu.cn; bma000@bit.edu.cn)

*Equal contribution.

†Corresponding author.

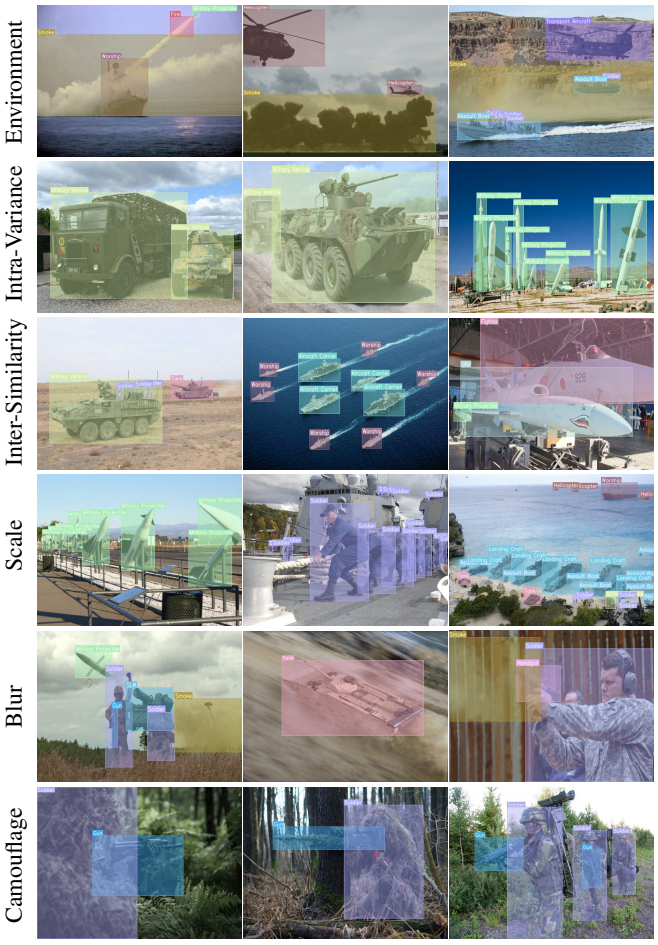


Fig. 1: Samples of our proposed MOD dataset, which contains various aspect challenge including complex military environment, intra-class variance, inter-class similarity, scale, blur and camouflage.

Military surroundings are much more complicated, where objects are likely to be in camouflage, occlusion or high-speed state. In order to restore those situation in the military field as much as possible, we propose MOD from abundant real-world military images. Therefore, MOD contains some unique challenges compared with other object detection benchmarks. As depecited in Figure 1, those challenges include (1) Complex environment: objects such as smoke and flames usually appear in military environment and thus military targets would be retreated behind; (2) Intra-class variance: the appearance of the same category like military vehicle could be quite different; (3) Inter-class similarity: the appearance of the different categories like military vehicle and tank could be quite similar; (4)

Scale: like commom datasets, objects at different distances would generate scale differences; (5) Motion blur: objects like projectile in military field are usually in motion; (6) Camouflage: objects like soldier and tank are usually decorated with camouflage. Therefore, part of existing object detection methods suffer from undesirable performance in MOD dataset (more details can be found in V).

In this work, we propose a Loss-Guided Attention RCNN (LGA-RCNN) to tackle those challenges by highlighting representative region of military objects. We find that in dense detection framework, RoI module can generate almost all features of foreground objects and the bottleneck of performance lies in the classification of RoI features. Thus, we append a LGA module behind RoI feature layers, which predicts k Gaussian masks on RoI feature maps to seek discriminative parts of objects for more accurate classification. In addition, an extra classification loss is imposed on masked RoI feature maps to ensure that those Gaussian masks converge to optimal locations. Compared with common attention modules like CBAM [3] which only focus on contextual information (rather than global information), our method makes full use of global information to mine representative local parts. Besides, time and memory consumption of our method are also better than global-range methods like non-local [4]. Extensive experiments demonstrate the effectiveness of our proposed LGA module.

Our contributions can be summaried as follows.

(1) We propose the first specific benchmark for military object detection (MOD) which brings various challenges including complex military environment, intra-class variance, inter-class similarity, scale, blur and camouflage.

(2) We benchmark a set of representative algorithm on MOD. Abundant experimental results are evaluated and compared on the same experiment settings. We further analyze the difficulties of the existing methods.

(3) We build a solid baseline on MOD dataset using LGA-RCNN which utilizes a LGA module to highlight representative regions for performance improvement.

II. Related Works

A. Datasets

Datasets play a very important role in the history of learning-based object detection methods. Previous detection datasets can be divided into single-category object datasets and multi-category object datasets (general object datasets). Single-category object dataset only contains one specific category of object such as face [5], [6], [7], [8], pedestrian [9], [10], [11], vehicle [12], apple [13], etc. Multi-category object dataset contains multiple types of objects such as person, bicycle or car. Previous representative works of multi-category object datasets include ImageNet [14], PASCAL VOC 2007 [15], PASCAL VOC 2012 [16], MS COCO [17] and Open Images [18]. Specifically, the detailed information of each dataset is listed in Table I.

Although those datasets show their effectiveness under the verification of numerous algorithms, they are collected for generic object detection, in which the types of objects are broad but not specialized. The dataset for a specific field is necessary because the characteristics of objects in different fields are quite different. And detection methods in specific field need to be improved to adapt to these characteristics, such as apple detection using enhanced YOLO-v3 [19]. Thus, we propose a benchmark for military object detection (MOD) in this work, which fully demonstrates the characteristics of military objects.

TABLE I: Comparison of Object Detection Benchmarks

Dataset	Specific Field	Categories	Boxes/Images
Pascal VOC [16]	Not Specific	20	2.4
ImageNet [14]	Not Specific	200	1.1
COCO [17]	Not Specific	80	7.3
OpenImages [18]	Not Specific	600	9.3
MOD	Military Field	20	2.9

B. Methods

According to whether to utilize region proposal, object detecion methods can be divided into two mainstreams, two-stage methods and one-stage methods.

1) Two-Stage Methods: Similar to tranditional object detection methods, two-stage object detection methods utilize a region porposal stage to generate sufficient candidate regions.

Inspired by selective search [20], Girshick [21] proposes RCNN in 2014 for generic object detection. However, repetitive feature extraction in RCNN causes slow operation. Thus, He et al. [22] propose SPPNet to reduce calculation time by obtaining proposals from the whole feature maps rather than the whole source image. Besides, Fast RCNN [23] is proposed with a Region of Interest (RoI) pooling layer to generate proposals of the same scale. Networks behind RoI layer become end-to-end so that detection speed is accelerated. Moreover, Ren et al. [24] replace selective search with Region Proposal Network (RPN) in Faster RCNN, which sets k anchors with different aspect raito in feature maps to generate proposals.

Recently, more two-stage methods [25], [26], [27], [28], [29] are proposed to enhance speed and performance. However, due to the existence of RoI, the speed of the two-stage method is still slow and cannot meet the requirements of real-time detection. Thus, one-stage methods are proposed.

2) One-Stage Methods: Unlike two-stage methods, one-stage methods achieve object detection without a distinct region proposal stage. According to whether to utilize anchor, they can be further devided into anchor-based methods and anchor-free methods.

Anchor-based one-stage methods apply anchors to classify object category directly rather than to generate region proposals. Liu et al. [30] propose a fully convolutional network SSD, which sets anchors in features with multiple scale to achieve detection on objects with different size. Then, Kong et al. [31] propose enchanced SSD algorithm, RON, that adds multiple deconvolutional layers to improve the detection capability in small objects. Lin et al. [32] propose RetinaNet with 9 anchors in each FPN scale. This work also introduces the focal loss to

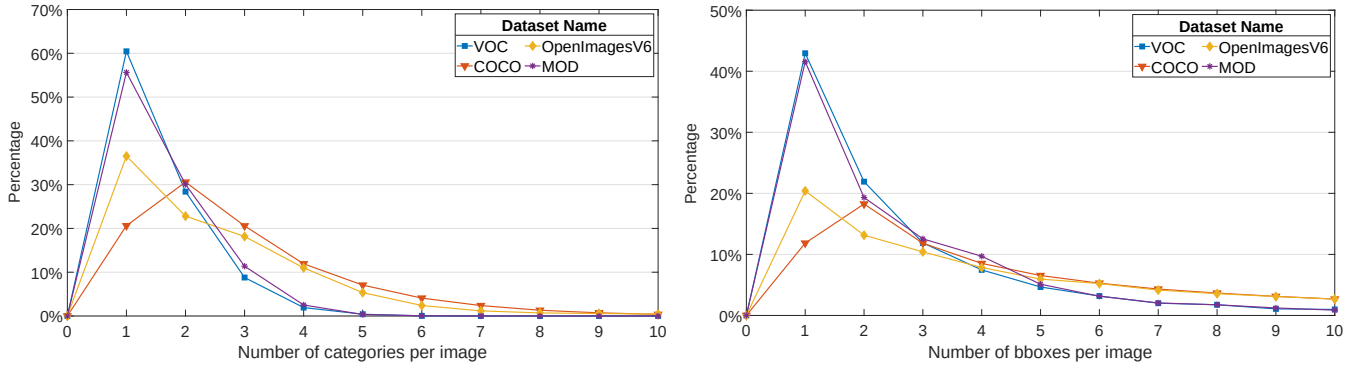


Fig. 2: Quantitative comparsion of the number of the categories per image. Our dataset MOD is similar with VOC 2012 [16].

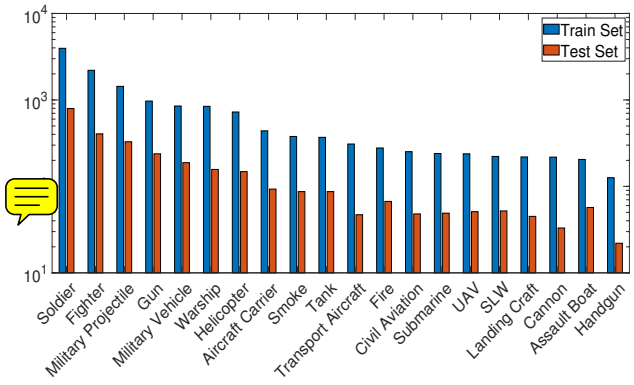


Fig. 3: Number of instances of MOD train set and test set.

solve the imbalance between positive sample assignment and negative sample assignment.

Those anchor-based one-stage methods are dependent on the setting of the anchor parameters to a large extent and unreasonable configuration prevents the anchor box from matching the target box well, resulting in performance drop. Thus, anchor-free one-stage methods are proposed [33], [34], [35], [36]. Specifically, YOLO [33] regards the object detection problem as the regression problem, where the feature map is split into $S \times S$ grid cells and each cell is responsible for preditcing objects centered at this cell. CornerNet [34] and CenterNet [35] convert object detection problem into a keypoint detection problem. Besides, ExtremeNet [36] utilizes the labeled data in the segmentation dataset to predict the boundary points and the center point of the object. The boundary points are guaranteed to fall into foreground area, so they are easier to detect than corner points. However, this algorithm needs to be trained with the mask annotation, increasing the acquisition cost.

III. Dataset

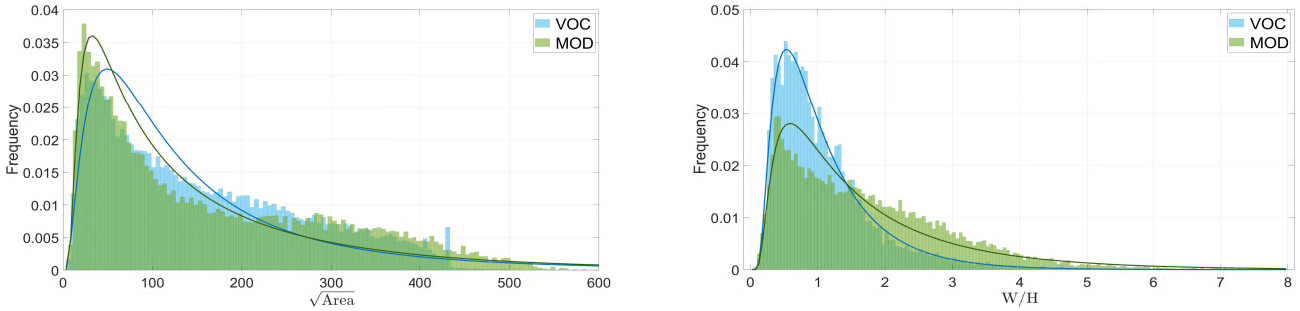
In this section, we introduce the process of dataset collection and dataset properties respectively.

A. Dataset Collection

1) Image Collection: The images in MOD are collected from website. In order to cover different aspects of military objects for practical use, we design a series of keywords as queries to crawl numerous military images from web search engine. Images that contain multiple categories of objects and rich semantic information can be retrieved by using keywords like "military exercise". These images account for a large part of our dataset and brings some unique challenges like camouflage or motion blur. Then, as a supplement, we query some particular military keywords like "helicopter" to get images with respective objects. Futhermore, we also collect small part of military images from ImageNet [14] by searching the relevant class in the word tree.

2) Image Filtering: After finishing data collection, we do the data cleaning and filtering manually to discard images which lack the reliability for the detection task. Considering the convenience for annotation, we filter those images with an extreme size and resize the rest of them into a corresponding size. To be specific, the larger side of the image is turned into 640, corresponding to COCO [17], while the aspect ratio of image remains unchanged. More unsuitable images are discarded with the following problems: (1) Images with extreme viusal problem caused by fire, smoke, high-speed and other unfavorable factors. (2) Images without military objects. (3) Images with very dense military objects. (4) Images containing severely deformed objects which are hard to recognize. (5) Duplicated images caused by multiple crawling. In total, we discard 25209 images from 31209 images to guarantee the quality of MOD.

3) Image Annotation: Subsequently, the military objects of interest in the collected images are annotated with rectangular boxes and corresponding categories. In MOD, the categories of military objects are divided into common military objects like tank, helicopter, as well as two special objects, smoke and flame that play an important role in high-level semantic perception and military situation awareness. The position of the military object is labeled with an orthogonal bouding box which is as close as possible to the real outline of the object. For special objects such as occluded objects, the criterias of labeling are: (1) densely label the densely distributed military objects; (2) only label the visible part of the occluded objects; (3) label all the visible parts of the high-speed objects with motion blur.



(a) area of bboxes (b) aspect ratio of bboxes

Fig. 4: Statistics of MOD and PASCAL VOC 2012 [16].

B. Dataset Properties

MOD is designed for military object detection, thus we assess the similar statistic with refer to those common object detection dataset, and make the comparison with the most widely-used dataset. At the meantime, it is worth noting that our MOD encounters unique challenges for the real-world use. The detailed properities are summaried as follows.

1) Overall: Our dataset MOD contains 20 categories from different military scenes and 6000 images in total, with 17465 object bounding boxes annotated. The whole dataset is divided into training set and testing set, where the ratio of training set to testing set is 5 to 1. The specific categories and number of instances in training set and testing set are shown in Figure 3.

2) Density: Other properties are also summaried and compared with mainstream object detection datasets (VOC 2012 [16], COCO [17] and OpenImages V6 [18]). We calculate the number of categroies per image and the number of bboxes per image. The quantitative results of are depicted in Figure 2. Our dataset MOD is similar with VOC 2012 and tends to have more percentage of images with single category or single box than COCO or OpenImages V6.

3) Scale: Since MOD has similar distributions of category and box with VOC dataset, we further compare more detailed properties of bboxes between MOD and VOC in Figure 4. The range of the annotations' scale in MOD is relative wilder than that of VOC. Figure 4(a) indicates that the middle scale bounding box proportion in MOD is less than VOC, whereas the small and large annotations in MOD surpass VOC. In fact, the large proportion of the small object brings challenges for the detection task, since the small object tend to cause severer detection failure than normal size.

4) Ratio: Meanwhile, as for the ratio of width to height of bboxes (Fig 4(b)), MOD possesses more instacnes between 1.6 and 8, less instacnes between 0.3 to 1.6 and equal instances between 0 and 0.3 than VOC, i.e., the bounding boxes in MOD tend to process a wilder aspect ratio compared with VOC. This phenomenon is caused by the special feature of the military target, for example, common military object such as cannon and gun tend to have a relative large aspect ratio. The extreme ratio would also bring challenges and cause performance drop for the traditional detection method, especially anchor-based algorithm.

5) Other Challenges: Beside the aforementioned challenges, other challenges also exist in the MOD. As depicted in Figure 1, the high speed of some objects like military projectile may cause heavy blur. Also it is common in battlefield that soldiers and other military objects camouflage themselves well, resulting in the challenge that the targets are extremely similar with their surrounding environment. Another challenge is that the heavy interference caused by the complicated environment. According to our statistics, 6.14% and 4.9% images in MOD training set are interfered with relative heavy smoke and fire, and 6.9% and 5.8% in MOD testing set, correspondingly.

IV. LGA R-CNN

After establishing the MOD dataset, we propose a solid baseline for military object detection. As illustrated in Figure 1, several challenges exist in military objects, e.g., occlusion, camouflage, and complex military environment, which cause a performance drop of previous methods to some degree. Thus, targeting at addressing those issues, we propose a LGA R-CNN for military object detection.

A. Overall

We build our method LGA-RCNN based on R-CNN framework and the whole pipeline is illustrated in Figure 5. Given an arbitrary image, RCNN detector utilizes backbone network and region proposal network (RPN) to generate feature maps with certain proposals. Then, RoI align is applied to crop RoI feature maps from the whole image feature maps. In such a dense detection framework, the bottleneck of performance lies on networks behind RoI features. Thus, besides the common classification and regression branches, we append auxiliary LGA module on RoI feature maps to predict and highlight representative regions for more accurate classification. Afterwards, those highlighted features are fused with the original RoI feature for preciser classification and regression.

B. LGA Module

The principal of designing LGA module is to mine and highlight those more representative and discriminative regions of the object, and reduce the adverse effect in potential region with occlusion, camouflage or other interference. To achieve this target, the proposed component should be able

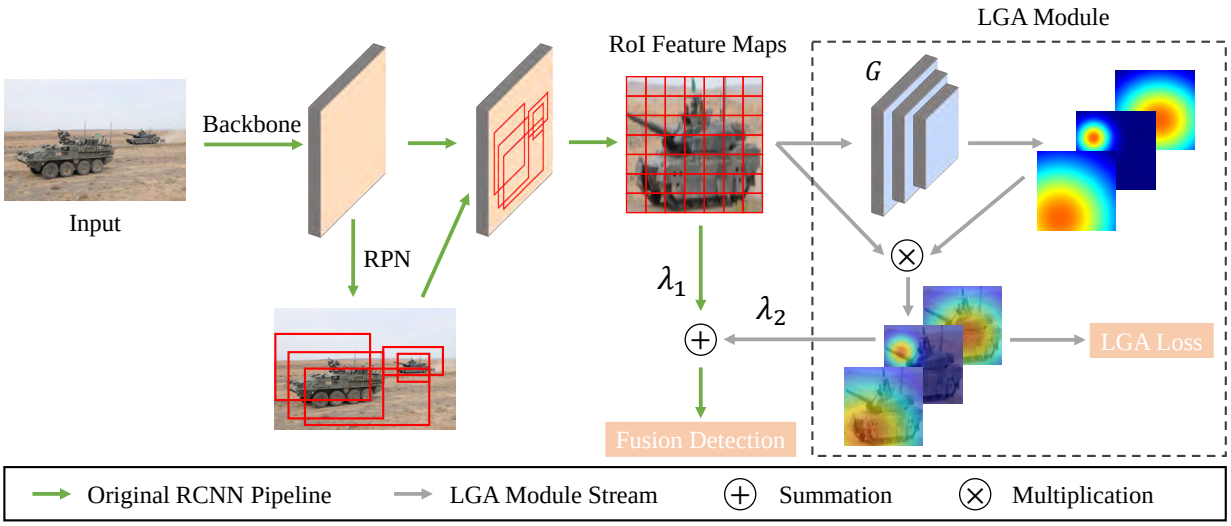


Fig. 5: Illustration of our proposed LGA R-CNN. Foreground object feature is extracted by Region Proposal Network (RPN). Then, we utilize a Loss-Guided Attention (LGA) module to predict several Gaussian maps from object feature and highlight discriminative part of feature map using those predicted Gaussian maps. LGA module is supervised and guided by classification loss under highlighted feature map. Furthermore, in order to achieve better regression performance, we fuse global information (original feature) and local information (highlighted feature) for final classification and regression.

TABLE II: Statictical evaluation metrics of previous detection methods and our proposed method. "LGA R-CNN" denotes the LGA module

Benchmark	Methods	BackBone	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MOD	SSD	VGG-16	40.200	65.800	41.300	7.100	27.200	47.200
	Faster R-CNN [24]	ResNet-50	42.910	70.286	44.913	15.865	34.217	49.003
	Retinanet [32]	ResNet-50	45.980	68.716	47.578	13.150	33.745	52.234
	FCOS [37]	ResNet-50	43.400	66.000	45.800	12.400	29.800	50.600
	Cascade R-CNN [38]	ResNet-50	49.634	68.308	53.258	16.250	37.016	56.621
	LGA R-CNN	ResNet-50	44.420	71.290	46.782	15.620	35.529	50.205
	Cascade LGA R-CNN	ResNet-50	50.806	72.160	54.721	16.595	37.460	58.017

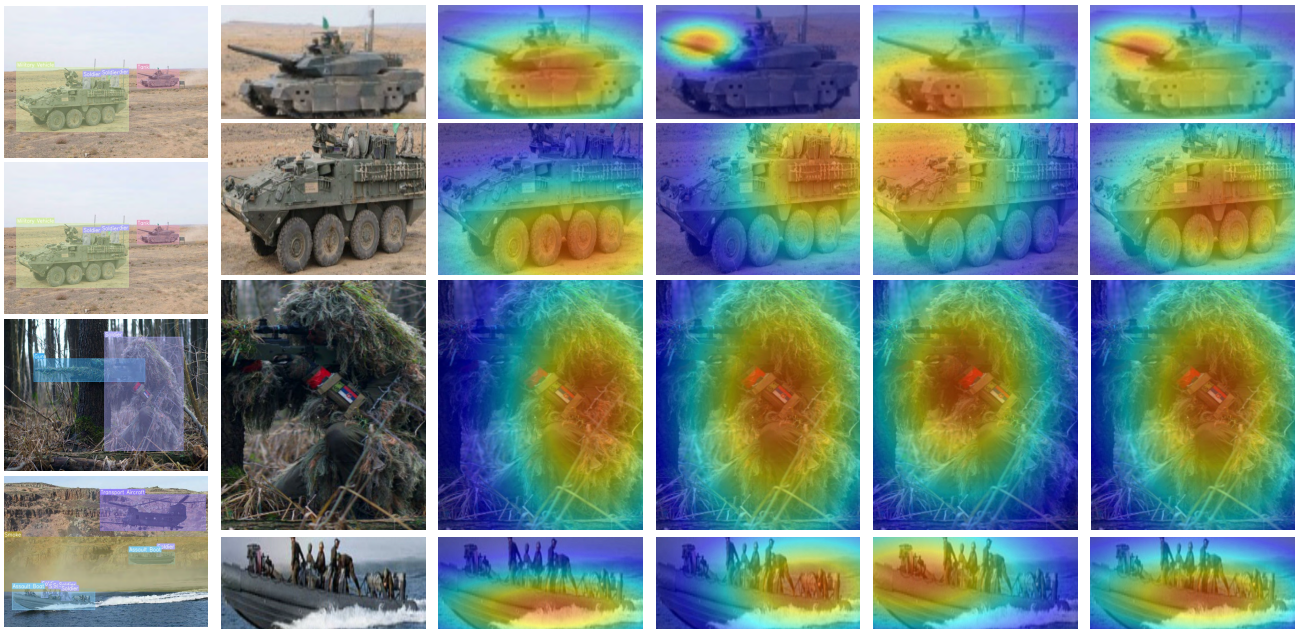
to sense the global information and seek the local region with more discriminative clues. Thus, we utilize a network to predict the Gaussian attention masks from the global RoI features. Assuming that those representative regions should be discriminative enough for a detector to do the classification, e.g., a solider's face is strong enough to be distinguished from other categories, we attach a classification loss to force LGA to learn a better attention. Furthermore, original global information need to be maintained for accurate locating and classification fine-tuning. Thus, we fuse those masked local-enhanced feature maps with those original feature maps for final detection heads.

1) Gaussian mask prediction: Common attention module such as CBAM [3] is implemented with channel-wise pooling and spatial-wise convolution, which thus leads to the lack of the global information. Non-Local methods are able to percept global information, but they are much more complicated and time-consuming. In LGA module, we construct a learnable mapping function to map the global features into a Gaussian parameters (μ and σ) then transfer those parameters into Gaussian masks. To be specific, given RoI feature x with 256 channels and 7×7 spational resolution, we first downsample the feature into a lower channel dimension to avoid high complexity by network f_d. Then, network f_c is applied on the downsampled feature to predict Gaussian parameters.

$$\begin{aligned} \mu &= S_{ratio} * Tanh(f_c(f_d(x))) \\ \sigma &= ReLU(f_c(f_d(x))) + 1 \end{aligned} \tag{1}$$

We utilize $Tanh$ and S_{ratio} to ensure that μ falls in the range of the spatial resolution of feature ($[0, 7]$ in this case) and $ReLU$ to ensure σ is no less than 1. Actually, Gaussian parameters are capable of representing some instance-level semantic information. For a RoI region of 7×7 size, the way we obtain gaussian parameters ensure that it can sense high-level semantic feature of the target instance. μ can be regarded as a position prediction on the discriminative region, while σ can be regarded as the scale of this region.

2) Loss-Guided Training: After initialized, different Gaussian masks pay attention to different regions, i.e., different local features are enhanced. We hope that those Gaussian masks would focus on more representative and discriminative regions. For example, when it comes to a picture with tanks



(a) Input (b) RoI (c) Mask1 (d) Mask2 (e) Mask3 (f) Fused Mask

Fig. 6: The visualization of RoI features and Gaussian masks. Note that the features and masks are actually with 7×7 spational resolution. Those Gaussian masks highlight the discriminative regions of objects accurately (e.g., the barrel of tank, the tires of military vehicle, the wrist of soldier in camouflage and the engine of the assault boat) which validates our hypothesis. An extra classification loss can guide the Gaussian attention to converge into more representative regions with less classification error.

and military vehicles, those unique parts such as barrel and caterpillar tread is more discriminative than similar parts like armor shell. To achieve this, we apply an extra classification loss on masked RoI feature maps for supervision. Assuming that common attention module do benefit the performance where they probably focuses on the armor shell, however the highlighted feature could be a disadvantage to distinguish tanks out of military vehicle. Loss-Guided training attention is designed to focus on a more discriminative region like barrel, which would not be a part of the military vehicle. With the supervision of the classification loss on Gaussian feature, the LGA module is forced to search for the aforementioned region to make the new-attached loss decline.

3) Feature Fusion: Although classification accuracy is improved by enhanced local informaion, part of global information is sacrificed in those highlighted RoI features. Therefore, inaccurate position regression would appear if we directly using highlighted features to locate the object. In order to maintain the accuracy and robustness of the bboxes regression process, we fuse masked RoI feature maps with original RoI feature maps to combine local information with global information. Then, we apply final detection on fused RoI features. Furthermore, part of Gaussian mask focuses on marginal region of the object. Thus, fused RoI features can sense more on the outline of the objects, which enhances the result of location.

V. Experiments

In this section, we introduce our experimental setup in V-A, results of previous methods in V-B and our method in V-C respectively, and ablation study in V-D.

A. Experimental Setup

To validate the effectiveness of our proposed benchmark MOD, we train certain state-of-the-art object detectors on MOD training set and evaluate the performance of those detectors in MOD testing set. Those methods include Faster R-CNN [24], SSD [30], RetinaNet [32], Cascade R-CNN [38]. For fairly comparing performance of previous methods in MOD dataset, we utilize official object detection frameworks Detectron2 [39] and mmDetecion [40] to train the models. As for our own detection model, we train it based on Deterctorn2 framework.

B. Results of Previous Methods

The statictical evaluation metrics of previous detection methods are listed in Table II. Among those methods, Cascade R-CNN [38] performs the best which achieves 49.634 mAP in ResNet-50 [41] backbone. The significant performance improvement from faster R-CNN [24] to cascade R-CNN [38] comes from classification and regression fine-tuning.

C. Results of Our Method

We implement our Loss-Guided Attention (LGA) on Faster-RCNN [24] network and Cascade-RCNN [38] network. As

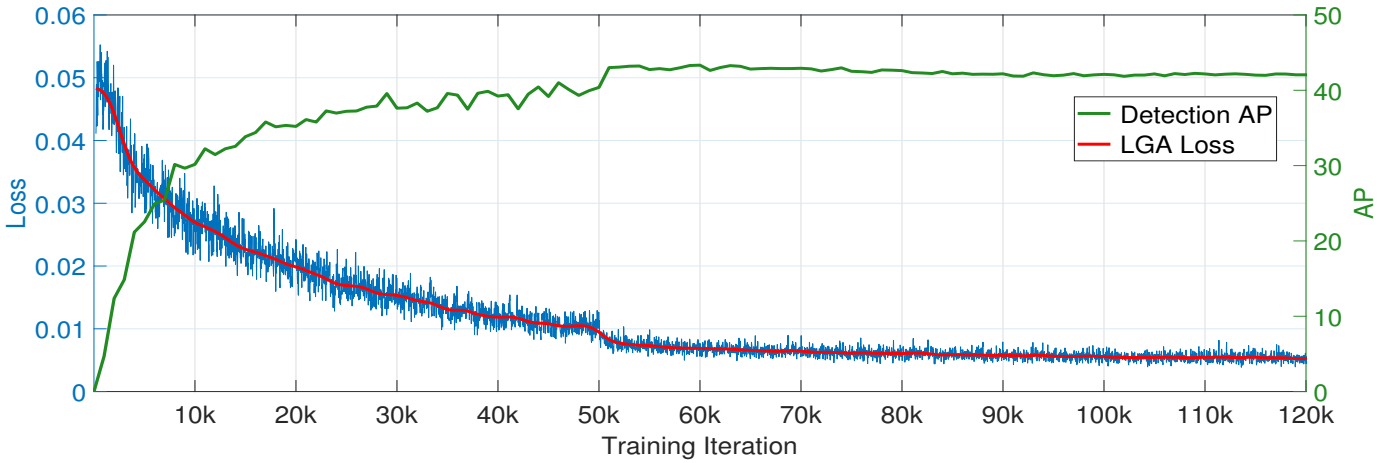


Fig. 7: Illustration of the convergence procedure of classification loss on masked RoI feature maps.

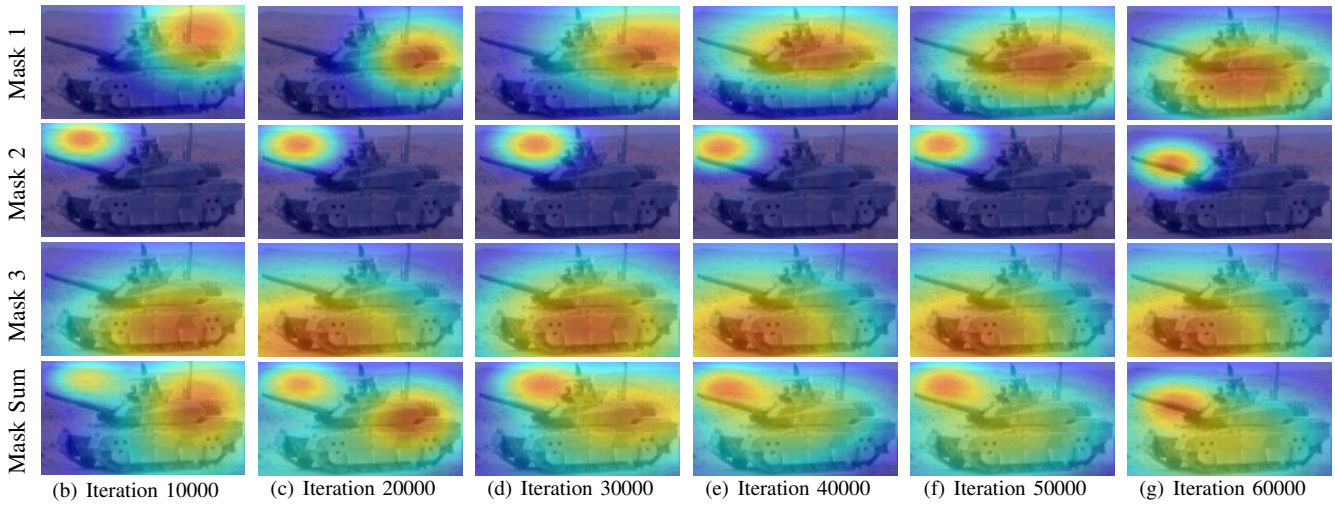


Fig. 8: Visualization of the movement of Gaussian masks. During training phase, Gaussian masks tend to converge to more discriminative region of object. For example, the barrel and caterpillar tread of tank is discriminative compared with military vechile. Therefore, mask 2 pays attention to barrel area and mask 3 pays attention to caterpillar tread area.

described in Section IV, we add auxiliary network after RoI feature maps to predict certain Gaussian maps and fuse those Gaussian maps with original RoI features. In order to ensure that those Gaussian maps highlight the discriminative parts of the object, we add classification loss on fused Gaussian maps.

The statistical evaluation metrics are shown in Figure II. After attaching LGA module, both Faster R-CNN and Cascade R-CNN achieves mAP gains.

To validate that those Gaussian masks in LGA module indeed pay attention to more discriminative regions, we further visualize gaussian maps along with RoI features. Since the spational resolution of RoI features is only 7×7 and it is difficult to recognize different parts of the object, we crop the RoI on the input image based on bbox prediction and utilize the prediction of μ and σ to draw Gaussian maps on it directly. The visualized results are shown in Figure 6. For objects with similar appearance, the center of Gaussian masks are able to fall in the regions with more discriminative local information (e.g., barrel and caterpillar tread in the tank, tires and rear seats in the military vehicle). For camouflaged objects, which lack sufficient apparent information, Gaussian masks are more concentrated on the only discriminative regions (e.g., wrist of the soldier). For occluded objects, the masks are able to focus on visible regions (e.g., three parts of the assault boat). Therefore, multi-region attention can be realized by fused masks.

Furthermore, to validate that the convergence of the Gaussian masks is guided by LGA loss, we plot the classification loss on masked RoI feature maps and visualize the Gaussian masks from iteration 10000 to iteration 60000, which is shown in Figure 7 and Figure 8, respectively. In training iteration 10000, the Gaussian masks focus on less discriminative region and the classification loss A is high. As the iteration time increases, the loss A decreases, and the Gaussian mask gradually moves to more discriminative regions.

TABLE III: Ablation study on number of Gaussian masks and "$0^{\dagger}$" denotes faster-rcnn baseline. We only evaluate the results with less than 5 masks, since more masks would be time-consuming and unnecessary. Among those experimental settings, the detector with 3 masks performs the best and achieves 1.51 mAP gains.

Mask Nums	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
$0^{\dagger}$	42.910	70.286	44.913	15.865	34.217	49.003
1	43.913	70.300	47.692	15.183	35.436	49.561
2	43.988	71.924	45.958	15.628	34.916	49.660
3	44.420	71.290	46.782	15.620	35.529	50.205
4	43.773	71.074	46.121	16.630	35.667	49.360
5	43.942	71.352	47.185	14.666	32.821	50.317

TABLE IV: Ablation study on LGA loss. "Loss" denotes extra classification loss supervision and "Cascade" denotes Cascade RoI heads. Experiments are conducted in 3 masks condition.

LGA	Loss	Cascade	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
			42.910	70.286	44.913	15.865	34.217	49.003
✓			44.144	71.981	46.668	15.435	34.145	49.866
✓	✓		44.420	71.290	46.782	15.620	35.529	50.205
			49.634	68.308	53.258	16.250	37.016	56.621
✓		✓	50.226	71.265	52.926	15.393	36.985	57.164
✓	✓	✓	50.806	72.160	54.721	16.595	37.460	58.017

D. Ablation Study

To demonstrate the effectiveness of different components in LGA-RCNN, we conduct ablation study on mask numbers and LGA loss.

The effect of different mask numbers on the final detection results is shown in Table III. LGA module with 3 Gaussian masks achieves the best performance because 3 two-dimensional Gaussian distributions are enough to fit complex planar distributions, i.e, discriminative regions at different locations can be highlighted by 3 Gaussian distributions. More Gaussian masks can also achieve this effect, however there are higher risks that Gaussian centers would overlap and the network would be overfitting. In addition, applying more Gaussian masks is not time-friendly.

The effect of LGA loss on detection results is shown in Table IV. Without LGA loss, performance drop appears. However, it still has higher mAP than baseline because final detection loss can also backpropagate to optimize the parameters of the Gaussian prediction network. Thus, this ablation study demonstrates that attaching the additional LGA loss is actually more effective than merely using the common final detection loss in Gaussian masks optimization process.

VI. Conclusion

In this work, we propose a military object detection benchmark (MOD) for evaluating object detection methods in military field. MOD comprises 6,000 images and 17,465 annotated instances in total. Compared with the generic object detection datasets, MOD contains rich military semantic information and brings certain unique challenges including camouflage, motion blur, compliated military environment, etc. In addition, we establish the benchmark of the previous detection methods on MOD and validate that those methods suffer from performance drop to some extent due to aforementioned unique challenges. Thus, we propose a Loss-Guided Attention RCNN (LGA-RCNN) to address those issues by adding LGA module in common R-CNN framework. LGA module utilizes a network to predict Gaussian masks from RoI features and force those masks to focus on representative regions of object by an extra LGA loss. Extensive experiments demonstrate the effectiveness of our method toward unique challenges in military object detection.

REFERENCES

[1] S. Astapov, J.-S. Preden, J. Ehala, and A. Riid, "Object detection for military surveillance using distributed multimodal smart sensors," in *2014 19th international conference on digital signal processing*. IEEE, 2014, pp. 366–371.

[2] Z. Yang, W. Yu, P. Liang, H. Guo, L. Xia, F. Zhang, Y. Ma, and J. Ma, "Deep transfer learning for military object recognition under small training set condition," *Neural Computing and Applications*, vol. 31, no. 10, pp. 6469–6478, 2019.

[3] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[4] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[5] V. Jain and E. Learned-Miller, "Fddb: A benchmark for face detection in unconstrained settings," University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[6] B. Yang, J. Yan, Z. Lei, and S. Z. Li, "Fine-grained evaluation on face detection in the wild," in *Automatic Face and Gesture Recognition (FG), 11th IEEE International Conference on*. IEEE, 2015.

[7] B. F. Klare, B. Klein, E. Taborsky, A. Blanton, J. Cheney, K. Allen, P. Grother, A. Mah, and A. K. Jain, "Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1931–1939.

[8] S. Yang, P. Luo, C. C. Loy, and X. Tang, "Wider face: A face detection benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 4, pp. 743–761, 2011.

[10] S. Zhang, R. Benenson, and B. Schiele, "Citypersons: A diverse dataset for pedestrian detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3213–3221.

[11] S. Zhang, Y. Xie, J. Wan, H. Xia, S. Z. Li, and G. Guo, "Widerperson: A diverse dataset for dense pedestrian detection in the wild," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 380–393, 2019.

[12] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[13] N. Häni, P. Roy, and V. Isler, "Minneapple: a benchmark dataset for apple detection and segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 852–858, 2020.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[15] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[16] M. Everingham, S. A. Eslami, L. Van Gool, C. K. 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.

[17] 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 *European conference on computer vision*. Springer, 2014, pp. 740–755.

[18] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale," *arXiv preprint arXiv:1811.00982*, 2018.

[19] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and electronics in agriculture*, vol. 157, pp. 417–426, 2019.

[20] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[21] 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*, 2014, pp. 580–587.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[23] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[25] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, pp. 379–387, 2016.

[26] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head r-cnn: In defense of two-stage object detector," *arXiv preprint arXiv:1711.07264*, 2017.

[27] G. Gkioxari, J. Malik, and J. Johnson, "Mesh r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9785–9795.

[28] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid r-cnn," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7363–7372.

[29] S. Beery, G. Wu, V. Rathod, R. Votel, and J. Huang, "Context r-cnn: Long term temporal context for per-camera object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 075–13 085.

[30] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[31] T. Kong, F. Sun, A. Yao, H. Liu, M. Lu, and Y. Chen, "Ron: Reverse connection with objectness prior networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5936–5944.

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

[33] 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*, 2016, pp. 779–788.

[34] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[35] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[36] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 850–859.

[37] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE international conference on computer vision*, 2019, pp. 9627–9636.

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

[39] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[40] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.