Improving CT-image universal lesion detection with comprehensive data and feature enhancements

Zhe Liu[1*], Kai Han[1], Kaifeng Xue[1], Yuqing Song[1], Lu Liu[2], Yangyang Tang[1] and Yan Zhu[3]

[1]School of Computer Science and Communication Engineering, Jiangsu University, Zhenjiang, 212013, China.
[2]School of Informatics, University of Leicester, Leicester, LE1 7RH, UK.
[3]Department of Imaging, Affiliated Hospital of Jiangsu University, Zhenjiang, 212013, China.

*Corresponding author(s). E-mail(s): 1000004088@ujs.edu.cn;

Abstract

As a crucial task in Computer Vision, object detection has substantially improved in recent years, with the aid of deep learning and increasingly abundant datasets. However, compared with natural image detection, medical CT images require more precision due to the obvious clinical implications. Detecting multiple lesions or clusters with relatively few training samples and indistinctive feature representation is extremely problematic. In this paper, we propose comprehensive improvements to the original YOLOv3, such as data augmentation, feature attention enhancement and feature complementarity enhancement to increase general lesion area detection performance. Ablation studies use the open DeepLesion dataset to validate these improvements and confirm the effectiveness of each amendment. Comparisons between state-of-the-art counterparts demonstrated that the proposed lesion object detector has enhanced salient accuracy (under two commonly-used metrics) and an exceptional speed-accuracy trade-off. The proposed model achieved 57.5% mAP and 85.07% sensitivity at 4 False Positives (FPs) per image, while running at reliable 35.6 Frames Per Second (FPS). These findings indicate that the proposed detector is more practicable than other currently available Computer Aided Diagnostics (CAD).

Keywords: Computer Aided Diagnostics, Computer Vision, Deep Learning, Object Detection

1 Introduction

Cancer is a heterogeneous disease with generally poor prognosis and extremely high mortality [1]. While other rarer lesions cause substantial morbidity and mortality, the prevalence of lung and breast cancers account for more than 20 percent of all recorded cases. Unfortunately, many of those diagnosed with a type of lung cancer or breast cancer are diagnosed at a late (metastatic) stage, which substantially reduces the number of effective treatment options. In the metastatic stage, cancer cells proliferate, migrate and colonize substrate within adjoining organs. At this point, the surgical resection is no longer an option at best and any intervention only inhibits further metastatic action. Thus, accurate diagnosis increases the likelihood of receiving a curative intervention and the chance of survival.

Computed Tomography (CT) plays an essential role in cancer screening. It utilizes different Hounsfield Unit (HU) window levels and widths to assist radiologists to identify lesions by analyzing correlations between slices along the z-axis. However, this routine procedure is both time-consuming and labor-intensive, which creates a backlog of CT scans and slows the diagnostic process. Missed opportunities for early diagnosis are not extraordinary in cancer screening because CT scans have relatively low specificity and sensitivity for various lesions. However, the Computer Aided Diagnosis (CAD) system has the potential to identify suspected cases earlier and to reduce the ratio of human errors. In addition, an accurate and efficient CAD system will undoubtedly be a foundational element, combined with the Internet of Things (IoT) for an intelligent medical system. However, Computer Aided Detection (CADet), a component of CAD, remains the most challenging due to heterogeneous lesions in terms of shape, size and location.

Object detection is a vital task in Computer Vision (CV), which is relatively sophisticated compared to other tasks in CV, such as classification and localization. To be specific, the detector aims to simultaneously determine both the object's location and the corresponding confidence score. The former is usually encoded using a pair of coordinates with left-top and right-bottom corner orientation while the latter is the probability of object presence to certain class for all labeled classes in specific dataset.

Over the past few years, the detection frameworks based on deep convolutional networks have become more commonplace and have yielded distinguished results in terms of speed and accuracy within natural image domains[2][3][4][5]. These pioneering works have encouraged researchers to consider applying this knowledge to object detection in medical domain [6][7][8][9][10][11][12]. However, there are methodological differences and therefore relatively mature detectors need to be adapted. For example, there is a

vast difference between natural RGB images and medical gray-level images generated by CT scan technology. Besides, the detection frameworks applied in medicine are data-driven, task-specific and individualized.

The aim of this study is to comprehensively enhance the lesion detection performance, which is different from detection in the natural scenario. The main contributions of our proposed framework can be summarized as:
(i) A new data augmentation strategy;
(ii) Incorporating feature enhancement modules into the feature extraction stage;
(iii) Our approach has achieved a better tradeoff between speed and accuracy on DeepLesion.

2 Related Work

In this section, we introduce several representative deep object detectors with their characteristics. We then illustrate methods to improve the backbone by employing multi-scale feature fusion. Finally, we list more targeted detection frameworks for medical imaging and analyze the respective drawbacks.

2.1 One-stage Object Detectors

Due to their robust and accurate performance, convolution operations could be implemented to exploit deep representative features in images belonging to almost all domains[13]. These domains vary from natural RGB images[14][15] to gray-level medical CT images[7], making the application of convolution operations very diverse[16]. For example, they could be implemented for tasks such as 2D/3D detection, pose estimation, and so forth; Meanwhile, as deep learning becomes prevalent and thrives, almost all successful state-of-the-art object detection frameworks proposed in recent years are based on convolutional neural networks. The experimental results apparently indicate that convolution-based detectors (also named deep object detectors) outperform the traditional detectors in both speed and accuracy, bearing great promise. In fact, pioneering deep object detectors, such as Region-based Convolutional Neural Network (R-CNN)[17], have achieved comprehensive advances. Since then, researchers have proposed a series of representative deep object detectors, which can be categorized as one-stage and two-stage detectors.

One-stage detectors directly predict bounding boxes and generate corresponding confidence scores without the region proposal stage. The representative YOLO[18] is a typical one-stage detector, which is designed with a simpler network and achieves approximately one order of magnitude faster speed than R-CNN, making the real-time application more realistic. Unlike YOLO, SSD[19] adopts anchors with different scales and aspect ratios on each feature. Therefore, SSD conducts further predictions based on multiple scales. While one-stage detector accuracy is relatively low, many researchers are considering how to improve and refine networks or other components so as to pursue the performance of the two-stage detectors.

Our proposed method is improved over YOLOv3[4], which is an incremental improvement to yolo9000[20] and achieves a better trade-off between detection speed and accuracy.

2.2 Multi-scale Feature Fusion

VGG[21], widely implemented for classification tasks, is modified as backbone in earlier version deep detectors, such as Faster Region-based CNN (Faster R-CNN) and SSD. After several down-sampling operations, the smallest feature maps at the end are finally fed into the trailing detection head, i.e. classification and regression. These single-scale features have the larger reception field and are therefore more appropriate for detecting relatively large objects. However, small or even tiny objects, which account for the majority of some difficult datasets, such as COCO[15] and DeepLesion[7], become barely visible given such scale of the feature map.

To tackle this problem, SSD makes the final predictions on feature maps with different scales. FPN[3] expands down-sampling stages by adding extra up-sampling operations, such as bilinear (or nearest) interpolation and deconvolution. In this way, each feature scale can be seen as one pyramidal level. Lateral connection is then used on each level to concatenate original features through the down-sampling phase with up-sampled features. It forms final feature representations from which contextual information is derived. Inspired by FPN, YOLOv3 and DSSD[5] improve small objects' detection accuracy in YOLOv2 and SSD respectively by modifying backbones. Rather than concatenating each levels features to make final predictions, Feature-Fused SSD[22] selects two of the most appropriate feature levels to combine. FSSD[23] firstly fuses several features with different scales at once, then generates pyramidal features from such fused features. In our proposed method, the scenario of employing FPN is different from that of the frameworks described.

2.3 Object Detection in Medical Field

Motivated by mature detection frameworks based on natural image scenario, researchers generalize the object detection in the natural area to that in medicine by adjusting these frameworks. However, due to the large difference in medical images, object detection in medical field would be more individualized and task-specific: In order to yield higher mean Average Precision (mAP), GSSD[6] improves upon the SSD baseline by intercalating four-phase enhanced liver lesion CT images in combination with group convolution. The key 1×1 convolution is also imposed before each SSD detection head to fuse multi-phase features. Even though the input data becomes more abundant and could potentially compensate for the over-fitting problem, injecting contrasting medium to form such kind of dataset may do much more harm to patients than plain scanning. In addition, the dataset used is exclusive, which may be deemed less persuasive in comparison with those methods conducted on a publicly available dataset.

In previous works, Ding et al.[11] improved the pulmonary nodule detection framework over standard CT imaging by implementing 3D convolutions into the False Positive Reduction (FPR) phase. The entire detection pipeline firstly introduces RPN and ROI pooling to generate 2D detection candidates, as proposed in Faster R-CNN. This then expands such candidates to 3D patches and intercalates a trailing FPR module to further increase the detection precision. Dou et al.[12] also employed 3D CNNs to extract more discriminative feature representations. However, 3D convolutions are more resource-consuming and computationally-intensive than 2D convolutions. This way requires very large graphical memory and strong GPU computational capability, which can not be conditioned in common laboratories. Likewise, Chiao et al.[10] introduced Mask R-CNN[24] for ultrasound imaging so as to build an automatic breast cancer diagnostics system. This model intercalates both detection and segmentation of breast lesions. It is worth noting that both [10] and [11] are two-stage based methods, their detecting speeds are relatively slow.

The above frameworks are all aimed at detecting specific types of targets, such as nodules, lesions or benign masses. However, these frameworks may be lacking in generalizability. A universal medical detector that identifies almost all lesions in various organs and tissues. Fortunately, there is a large-scale medical object detection dataset, known as DeepLesion[7], which is open-access and provided by NIH in 2018. DeepLesion contains eight different types of lesions and it will be described further in Section 4.

In the original paper on DeepLesion, Faster R-CNN[2] was implemented and yielded great performance, but it was still suboptimal under the strict evaluation metrics. Continuous investigations have been conducted by researchers since the DeepLesion dataset was publicly accessible. For example, ULDor[8] adopted Mask R-CNN[24] to develop a universal lesion detector. This model constructs pseudo-masks to train Mask R-CNN feasibly and achieves better performance.

Likewise, improving upon Region-based Fully Convolutional Network (R-FCN), 3DCE[9] increases detection sensitivity by a substantial margin compared to the original Faster R-CNN. This model utilizes a central key slice with neighboring slices as input images, then passes those through a feature extractor separately and concatenates features to fetch 3D contextual information. The backbone is fine-tuned and extra layers are added. However, we would suggest that the training scheme may rely heavily upon pre-trained weights because the input channel of the first convolution of the backbone is cautiously designed i.e., 3-channels per image. Also, pre-trained weights commonly used in object detection is achieved through training on huge scale ImageNet[25] classification tasks, which is a considerably time-consuming process.

As such, we propose the following framework to comprehensively improve one-stage object detector YOLOv3, expecting to achieve a reasonable detection precision and recall compared with two-stage detectors. Our primary aim is to develop a universal detector, which can identify eight types of lesions in

CT scans from chest to abdomen. Therefore, it tends to be more robust and practical than detectors designed specifically for one type of lesions.

3 Methods

In this study, we aim to design an accurate and comprehensively improved one-stage medical universal detector by elaborating three key components. As has been mentioned, medical imaging is quite different from the natural image. Therefore, we implement a new data augmentation strategy as a substitute for randomized contrast in natural scenarios. To refine and strengthen extracted features from the backbone, we design two feature enhancement modules. Anonymized lesions in medical CT datasets are generally small, therefore an effective way is to set more anchors with smaller sizes and more aspect ratios. The overall proposed network architecture is shown in Fig 1.

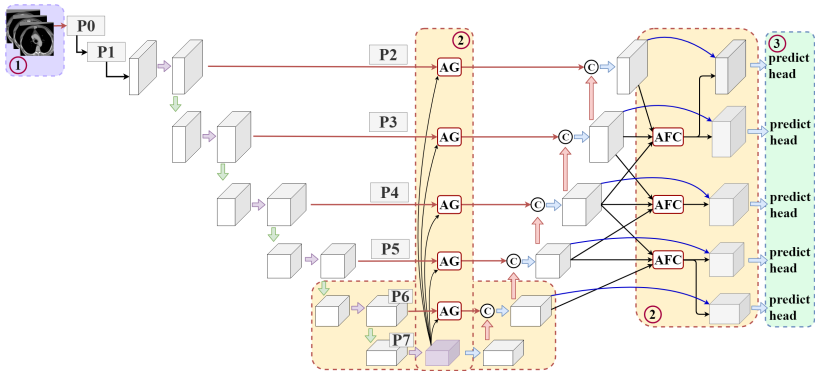

Fig. 1 Overall network architecture. Modifications are highlighted in rounded rectangles with different colors. Numbers from 1 to 3 indicate data augmentation, feature strengthening and anchors adjustment & expansion, consequently. This is described and elaborated upon in Section 3 (below). Best viewed in color.

3.1 Data Augmentation

A crucial factor of affecting detection accuracy is how to pre-process input data. Incorporating effective data augmentation strategy could alleviate over-fitting problem caused by the relatively insufficient training data to some extent.

During radiologists' routine diagnosis, another crucial basis is the views of CT scans, which are called HU window level and window width. Under different HU window levels, radiologists could focus on the lesions in different locations(e.g. lung, liver, pelvis, soft tissue, etc). In addition, under different window widths, the contrast between the focused location and its surrounding part is different. In this work, we initially do HU restriction in accordance with the window level provided by expert radiologists, which means fixing

the window level. Then we impose randomized increment on both the upper and lower bounds of such window level, implementing randomized window width, which could also be deemed as the contrast randomization in medical CT-image domain. Note that such randomization is not arbitrary. Once the increment exceeds 500, the contrast may become extremely low in some cases, thus making some lesions scarcely visible. See Fig 2 for visualization.

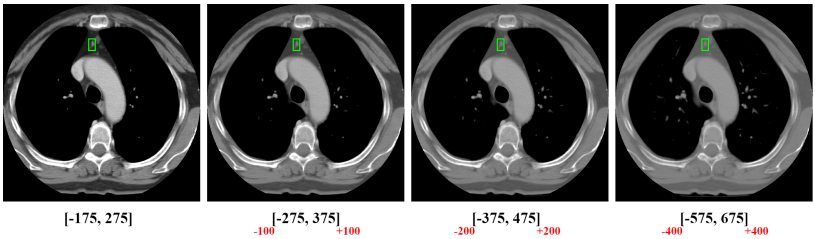

Fig. 2 CT scans through different HU window widths when the window level is fixed. Intervals below each scan are HU windows. Each increment imposed upon the bounds is highlighted in red. The lesion is marked with a green bounding boxes. Best viewed in color.

Besides the aforementioned new strategy of data augmentation, several commonly-used affine transformation strategies are also incorporated into training, including randomized translation, rotation, shear, scale, vertical and horizontal flipping. The parameters set in them are 10%, 10°, 10°, 10% and 0.5 probability respectively. Note that no adjacent slice and data augmentation is used during testing and detecting.

3.2 Feature Enhancement

Incorporating residual blocks and multi-scale feature predictions into detection network, YOLOv3[4] achieves great detection results in the natural scenario. However, we hypothesize that such results are still suboptimal for medical lesion detection in terms of insufficient feature representation, due to the indistinguishable contrast between lesions and non-lesions. Besides, the lesion targets pending to be detected are relatively medium-scale or small-scale.

We address this potential problem by initially deepening the backbone network. Our proposed detector also incorporates FPN to implement multi-scale prediction. One common principle of designing such prediction is that the lower or shallower pyramidal level is suitable for detecting smaller-sized objects, while the higher or deeper level for detecting larger-sized objects[3][4][5]. Following this principle, initially, we append two extra downsampling stages in encoding part by expanding the original Darknet[4]. In this way, we add two extra upsampling stages in decoding part correspondingly, which means that two more pyramidal levels are generated. Hence, we could enable the network to extract semantically stronger features, and impose two more prediction layers to make detector more robust to objects with various sizes. For clarity, (P0, P1, ..., P7) stands for each pyramidal level from shallow to deep hereinafter.

According to the aforementioned expansion, we set prediction layers after P3-P7. However, the objects in DeepLesion dataset are generally small-scaled. To enhance the detection accuracy, we adjust prediction layers by shifting them from P3-P7 to P2-P6. Note that P7 is kept to maintain the overall pyramidal network structure and the capability of extracting deep features, as depicted in Fig 3.

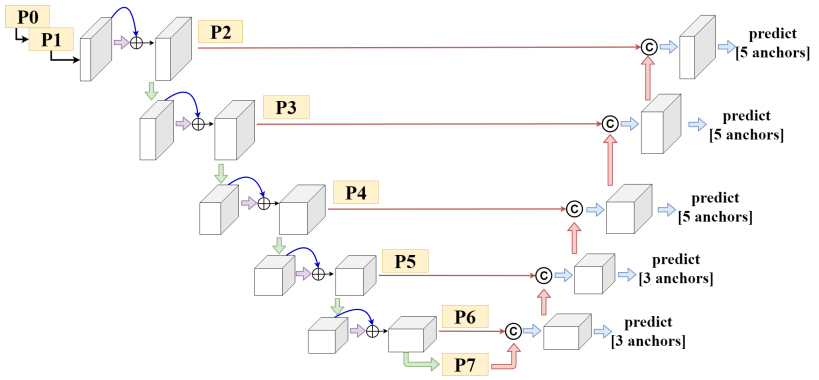

Fig. 3 Network architecture after deepening and adjusting

Based on the aforementioned preliminary network refinement, we further propose two feature enhancement modules embedded in network. We will elaborate them in subsequent part of this section.

3.2.1 Feature Attention Enhancement

The attention mechanism is initially proposed in the machine translation domain and has become a key element of Natural Language Processing (NLP)[26]. Recently, the attention mechanism has more been adopted for image classification tasks[27][28] and appears to enhance accuracy. There are two main approaches to implement the attention mechanism in such tasks, i.e. channel-wise attention and pixel-wise attention. In our preliminary exploration, we initially attempted to use channel-wise attention in the backbone. However, the final mAP drops substantially, which indicates that this way is not suitable for the entire detection framework. Therefore, we hypothesize that the global pooling operation causes some deviations, since the input of detection task is the whole image and certain objects may appear in an arbitrary region. Hence, we adopt pixel-wise attention and embed an Attention Gate (AG) module into the backbone.

Similar to the correlation calculations between query and key in the original scenario to which the attention mechanism is applied, this module finally generates an attention map. Then pixel-wise multiplication with input features is conducted. This process strengthens the salient features in foreground regions while suppressing those in the non-pertinent surrounding background,

making our detection framework focuses on lesion areas. Besides a series of operations, the gating feature is also determinant to generate an attention map. The features on P7 are chosen as global gating features since they have the largest reception field and strongest semantic information, enabling them to guide each AG set on P2-P6 before lateral connection. The detailed structure of AG is depicted in Fig 4. The process for generating attention maps is described elsewhere 1.

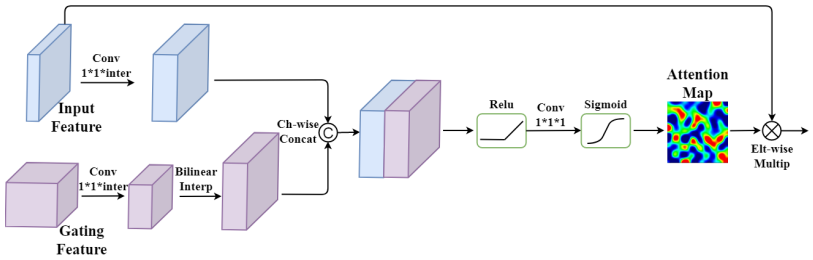

Fig. 4 Detailed AG structure. Input features derived from each P2-P6 layer before lateral connection, while gating feature is fixed.

Note that our detector could also be regarded as a pixel-wise task, in which several predictions are generated on each grid cell of feature maps. The strengthening and suppression functions of AG directly affect objectness score, which is one value within the prediction vector. Meanwhile, since there is only one class in the lesion detection task, the objectness score also determines confidence score, which could be considered as an indicator. When the confidence value is greater than the predetermined threshold, the prediction is selected as a positive sample, otherwise it is selected as a negative sample. Take one false positive sample for example, its confidence score will decrease and may become a true negative under the suppression of AG. Hence, the precision will be improved to some extent.

$$map = \sigma\{C_{1*1*1}[relu[C_{in}(F^{l}_{input}) \oplus up[C_{gate}(F_{gate})]]]\}, l \in [P2, P5] \quad (1)$$

Where $\oplus$ and σ are channel-wise concatenation and sigmoid function, respectively. C, F, and up represent Convolution, Features and upsample, respectively.

3.2.2 Feature Complementarity Enhancement

When designing a FPN-based detector, it is necessary to abide by a common principle. The shallower or lower pyramidal level, the more suitable for detecting small-sized objects while the deeper or higher levels are more appropriate for large-sized objects. However, we assume that this concept is a little unclear. There may be no guarantee that each level is absolutely suitable for detecting

objects from specific pre-defined anchors. In other words, features before each prediction head seem to be relatively independent and remain to be locally fine-tuned to optimally match the scale of objects pending to be detected. Hence, we try modifying the network architecture right before each prediction head, enabling it to fetch complementary information from adjacent pyramidal level, both spatially and semantically.

The embedded module Adjacent Feature Complementation (AFC) could be competent for this task, which is divided into two phases, i.e. feature fusing and feature restoring. In the former phase, features from three sequential levels are initially resized to the same scale, then these are concatenated together across the channel axis. The subsequent two 1×1 convolution operations play the crucial role of feature fusion and local fine-tuning. During the latter phase, adopting interpolation or pooling to restore the same spatial scale. Then the convolution operation is conducted to restore channel scale as the original channel scale. At this point, each feature level is restored to its original scale. Finally, adding the restored feature and the original feature. To stabilize unanticipated circumstances, we also employ the shortcut structure as proposed in residual block[29]. The aforementioned process is clearly depicted in Fig 5.

3.3 Anchors Adjustment & Expansion

The detection head is the key component in object detection, which connects both the output of CNN backbones, i.e. the extracted feature, and the prediction of the whole object detector. Hence, the detail concerning the amount of predicted boxes regressed from anchors (also named prior boxes or default boxes) is also non-negligible.

In the original YOLOv3, the anchor point set in the detection head is specific to the dataset, rather than fixed[19]. Specifically, they are generated by the K-means algorithm based on the width and height distribution of the overall ground truth (GT) box in the COCO dataset. Conversely, the dataset we use is the medical lesion dataset, which is very different from a natural image data set in terms of data distribution. In the Deeplesion dataset, there are more small targets and in order to improve the sensitivity for detecting small lesions, we use more anchor points in the prediction layer after the shallower P2, P3 and P4 levels. Therefore, we re-clustered the DeepLesion dataset to generate specially designed anchor points, which is shown in Table 1.

4 Experiments

In this section, we initially introduce the training schedule, the datasets and evaluation metrics used in our experiments. Then, we analyze the effectiveness of our methodology by ablation study and the comparison with the state-of-the-arts. All experiments are conducted on GeForce GTX 1080Ti GPU with 11 GB graphical memory, Intel Core i7-7700K CPUs with 8 threads and 15.6 GB physical memory.

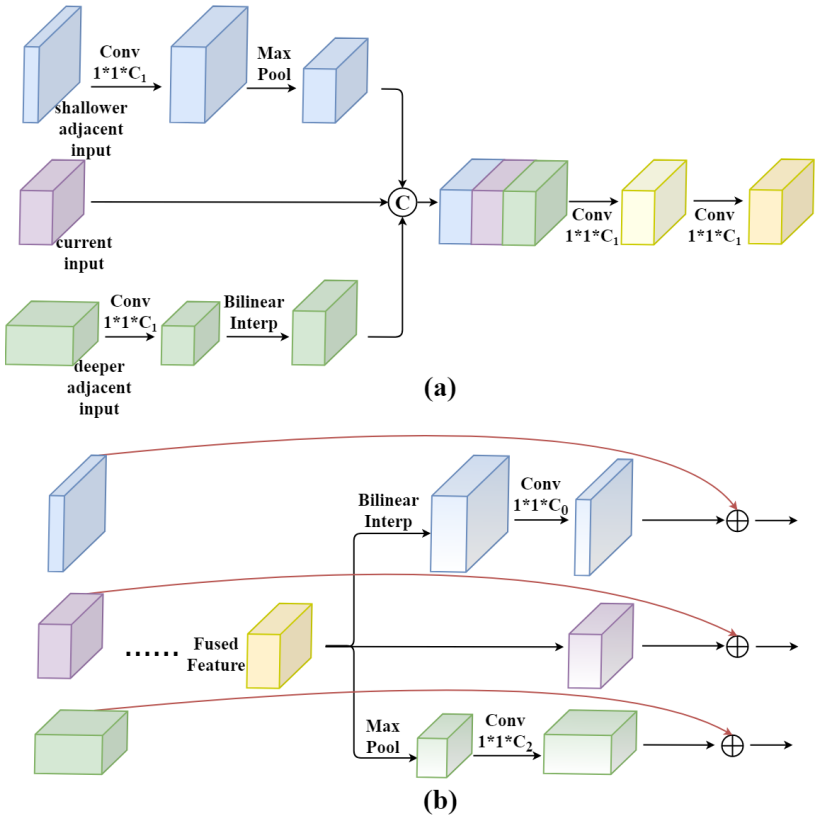

Fig. 5 The detailed structure of AFC. (a) feature fusing phase; (b) feature restoring phase. The input derive from the current and its two adjacent levels of features before prediction heads. Best viewed in color.

4.1 Dataset

We conducted the experiments over publicly available DeepLesion dataset. In this dataset, the lesion types are very diverse, including bone, abdomen, mediastinum, liver, lung, kidney, soft tissue and pelvis. It is large-scale because the number reaches up to 32,725 lesions on 32,120 axial CT slices from 10,594 studies of 4,427 unique patients. The official standard splits the overall 32,120 CT images with the proportion of 70%, 15% and 15% to generate training set, validation set and test set, respectively. The crucial labeled information comprises the bounding boxes of lesions, z-axes spacing and HU windows, which are indispensable for object detection task itself and our proposed methodology. In addition, the labeled lesion types are only given in the testing set to help to visualize and evaluate the detection results of each type, which means the detection is actually a single-class task.

Table 1 Anchor configuration in each prediction layer following certain pyramidal level after adjusting detection heads.

Pyramidal level	Anchors Configuration
P2	(15,15),(20,20),(28,22),(20,14),(14,20)
P3	(24,28),(30,29),(30,36),(30,20),(20,30)
P4	(41,33),(39,48),(58,43),(44,29),(29,44)
P5	(52,62),(73,60),(69,93)
P6	(102,85),(124,132),(193,168)
Total number of anchors: 5*3+3*2=21	

4.2 Training Schedule

Our detector was trained with PyTorch v1.1.0 deep learning framework. Due to limited graphical memory, the batch size was set to 8 for all experiments. During data loading, the number of workers was set to 8, which is exactly the number of CPU threads. Stochastic Gradient Descent (SGD) was employed as the optimizer, with weight decay of 0.0005 and momentum of 0.9. The initial learning rate was 0.001, and then be decreased by 10 times when the current training process reaches 80% and 90% of total epochs, which was set to 80 for training DeepLesion. The training will eventually converge after several epochs when learning rate is 1e-5.

4.3 Evaluation Metrics

In this section, the main metrics are introduced to evaluate our detector in subsequent ablation studies and comparisons.

4.3.1 Mean Average Precision

Average Precision (AP) is generally defined as the approximate area under the precision-recall (PR) curve of a certain class. Whereas, mean Average Precision (mAP) is the mean value of APs added up by each class. For the single-class detection task, mAP is identical to AP. To scatter and finally draw PR curve, we should dynamically calculate the progressive recall value and its corresponding precision value, according to all positive predicted boxes with sorted confidence scores from high to low. Such calculating process is also named all-points interpolation method (See (3)), which is suggested by PASCAL VOC 2012 standard and is extensively adopted. The paired recall and precision value is calculated from the current number of true positives (TPs) and false positives (FPs), see 2.

$$precision = \frac{TPs}{TPs + FPs} = \frac{TPs}{All\ Detections}$$

$$recall = \frac{TPs}{All\ Ground\ Truths} \tag{2}$$

$$with\ P = \begin{cases} TP,\ if\ IOU(p, GT) \geqslant threshold \\ FP,\ if\ IOU(p, GT) < threshold \end{cases}$$

Where $IOU(p, GT)$ stands for the Intersection over Union between one specific predicted box and corresponding ground truth box, while the threshold is set to 0.5 by default.

$$AP = \sum_{r=0}^{1} (r_{n+1} - r_n) * p_{interp}(r_{n+1})$$

$$with\ p_{interp}(r_{n+1}) = \max_{\hat{r}:\hat{r} \geqslant r_{n+1}} p(\hat{r}) \tag{3}$$

$$mAP = \frac{1}{N_c} \sum_{c=0}^{N_c - 1} AP_c$$

Where the middle equation indicates searching for the precision envelop at the right side of recall point, to gradually obtain the final estimated area under PR curve.

4.3.2 Sensitivity at various FPs per image

Different from the constant AP metric, another stricter metric we employed is the sensitivity (recall) at various FPs per image. As its name implies, it is a metric to evaluate the capability of detector under varying levels of strictness. To implement this, we should set different confidence threshold, which is used for distinguishing between positive and negative samples before doing NMS. For illustration, if confidence threshold is set from 0.01 to 0.001, a definite consequence would be the increase of recall and the decline of precision. This is because slightly more TPs and dramatically more FPs are introduced, see (3). Hence, by continuously changing this threshold, we can obtain recall values under different FPs per image. The metric fixes FPs per image (usually be 0.5, 1, 2, 4 and 8), in order to see whether the detector could find more true positives as much as possible under the same fault tolerance.

4.4 Ablation Study

In this section, we will gradually confirm the effectiveness of each component of improvement by conducting ablation study over DeepLesion dataset and giving quantitative analysis.

According to the two evaluation metrics, we conducted ablation over DeepLesion dataset. From Table 2 and Table 3, we observed: by adding each improved component, either Recall or mAP continuously increased, which demonstrates that each improvement could play a specific role in enhancing accuracy. Specifically, data-augmentation (DA) has two functions. On the one hand, conducting HU restriction and randomized increment can strengthen the difference between the lesion area and its sorrounding area, which makes it easier for the network to locate the lesion area. On the other hand, to alleviate over-fitting problem, several common affine transformation strategies are used to expand the dataset. Table 2 shows the effectiveness of the data-augmentation we proposed. Besides, Feature Srengthening(FS) also shows positive effect, which includes Attention Gate(AG) and Adjacent Feature Complementation(AFC) structure. The former strengthens the salient features in foreground regions while suppressing those in the non-pertinent surrounding background. The latter could fuse different scale complementary information from semantic and semantic aspect. Furthermore, feature fusing phase and feature restoring phase are integrated to Feature Complementation(AFC), where feature fusing phase could fuse different scale adjacent features to obtain more semantic and spatial information while feature restoring phase combines low-level semantic information and high-level semantic information to compensate the detail information during the process of upsampling. Finally, according to dataset distribution, Anchors Adjustments & Expansion(AE) module resets anchors to detect lesions, which also enhences the performance of our proposed approach.

Meanwhile, the detection speed drops continuously, which is because the backbone is more sophisticated. Then, the number of corresponding training parameters increase. In spite of this, our proposed detector is still a real-time one, which FPS is greater than 30. Since DeepLesion provides a coarse lesion type of each CT slice, we calculate AP of each lesion type, which is provided in Table 4. It could be found that lesions located in the bones, abdomen and soft tissue are relatively more difficult to detect, in comparison with other types. We generated a PR Curve and Free-response Receiver Operation Characteristic (FROC) curve to make results more intuitive, see Fig 6.

To further verify the effectness of our AG structure, several typical attention structions are compared. At first, channel attention(CA) is embeded into network and then final mAP drops, we infer that the global pooling operation causes some of the deviation since the input of detection task is the whole image and certain objects may appear in an arbitrary region. In addtion, other attention modules are compared in 5. We could find that our proposed AG module has obtained the best mAP score. Meanwhile, the Recall score is close to CBAM.

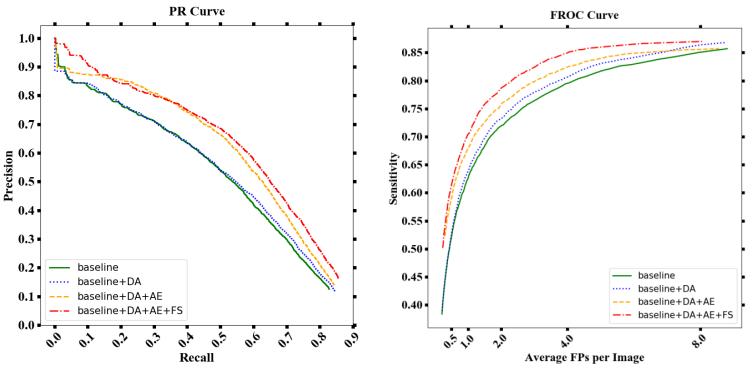

Fig. 6 Ablation w.r.t. PR Curve (left) and FROC Curve (right) on the official split test set of DeepLesion.

Table 2 Ablation w.r.t. Recall(%),mAP(%) and detecting speed (FPS) on the official split test set of DeepLesion.

Components	Recall	mAP	Detecting Speed
Baseline	82.8	47.9	**63.9**
Baseline+DA	84.7	48.7	63.9
Baseline+DA+AE	84.4	55.1	48.5
Baseline+DA+AE+FS(Proposed)	**85.8**	**57.5**	35.6

Where DA, AE and FS are short for Data Augmentation, Anchors Adjustments & Expansion, and Feature Strengthening, respectively.

Table 3 Ablation w.r.t. Sensitivity(%) at various FPs per image on the official split test set of DeepLesion.

Components	0.5	1	2	4	8
Baseline	52.02	62.55	71.89	79.51	85.14
Baseline+DA	52.63	63.81	73.29	80.63	86.38
Baseline+DA+AE	58.88	67.81	75.74	82.42	85.57
Baseline+DA+AE+FS(Proposed)	**61.27**	**70.43**	**78.67**	**85.07**	**87.01**

Detection results are visualized in Fig 7. As can be seen: (1) Some prediction boxes are better regressed and therefore achieve higher confidence scores compared to results from the baseline detector (first column); (2) Even as the baseline detector can identify one true lesion, there still exists a duplicate box for one GT or false positive somewhere else (second column); (3) The proposed detector can at least locate the lesion area, while the baseline finally generates nothing but the original image pending to be detected.

Table 4 Ablation w.r.t. AP(%) of each lesion type on the official split test set of DeepLesion.

Components	BN	AB	ME	LV	LU	KD	ST	PV
Baseline	37.1	37.2	58.7	52.4	56.2	37.3	40.3	43.5
Baseline+DA	34.7	37.9	58.6	53.6	59.1	42.6	38.4	41.4
Baseline+DA+AE	46.3	43.7	65.0	60.2	**66.4**	49.0	41.8	46.0
proposed	**47.8**	**45.3**	**66.9**	**63.7**	**66.4**	**52.9**	**46.3**	**52.0**

Each abbreviation form of type from left to right indicates bone, abdomen, mediastinum, liver, lung, kidney, soft tissue and pelvis respectively.

Table 5 Ablation w.r.t. Recall(%) and mAP(%) on the official split test set of DeepLesion.

Components	Recall	mAP
Baseline	82.8	47.9
Baseline+CA[30]	83.5	49.3
Baseline+SE[27]	83.2	50.8
Baseline+RA[31]	82.6	48.4
Baseline+CBAM[30]	**84.3**	51.1
Baseline+AG	83.9	**52.2**

Where CA, SE, RA and AG are channel attention, SE module, residual attention module and attention gate, respectively.

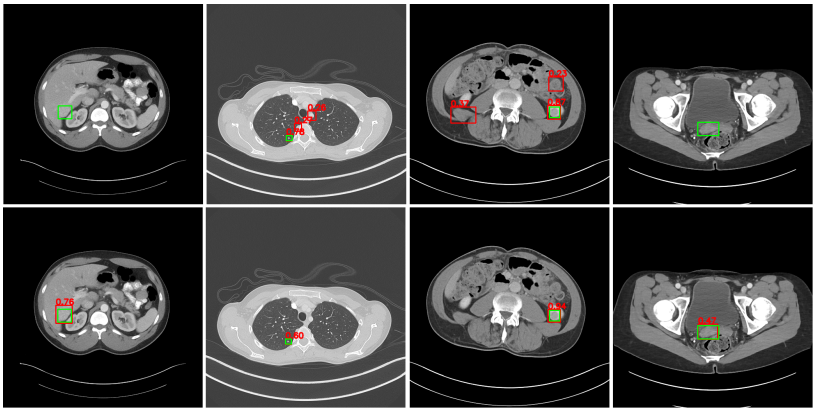

Fig. 7 Detection visualization of the comparison between baseline and proposed detector. The red and green bounding boxes represent predictions and ground truths. Each column from left to right shows the lesions on liver, lung, kidney and pelvis, respectively. Best viewed in color.

4.5 Comparison and Analysis

In this section, we compare our proposed detector with other state-of-the-art ones by comprehensively analyzing the trade-off between detection accuracy and efficiency over the official test set. Table 6 shows detection sensitivity at

various FPs per image of each method, while displaying corresponding detection speeds. The typical two-stage detectors, i.e. Faster R-CNN[2] and Mask R-CNN[24], are supposed to be more accurate than our one-stage detector, due to the coarse-to-fine pipeline guided by RPN. Conversely, their detection sensitivities at each FPs per image are completely inferior to ours. In addition, even though other one-stage detectors, i.e. RetinaNet[32], Efficientdet[33] and so on, are close (and in some cases superior) to ours in terms of detection efficiency, our detectors detection accuracy consistently outperforms them.

Table 6 Sensitivity(%) at various FPs per image and runtime of various methods on the official split test set of DeepLesion.

Methods	0.5	1	2	4	runtime
RetinaNet[32]	45.80	54.17	62.50	69.80	28 ms
YOLOv3[4]	52.02	62.55	71.89	79.51	**16 ms**
Faseter R-CNN[2]	56.90	67.26	75.57	81.62	32 ms
Mask R-CNN[24]	39.82	52.66	65.58	77.73	–
ULDor[8]	52.86	64.80	74.84	84.38	–
3DCE, 3slices[9]	55.70	67.26	75.37	82.21	–
3DCE, 9slices[9]	59.32	70.68	79.09	84.34	56 ms
3DCE, 27slices[9]	**62.48**	**73.37**	**80.70**	**85.65**	114 ms
Ours	61.27	70.43	78.67	85.07	28 ms

Table 7 mAP(%) and AP(%) of each lesion type of various methods on the official split test set of DeepLesion.

Methods	Total	BN	AB	ME	LV	LU	KD	ST	PV
Faster R-CNN[2]	48.4	52.4	39.1	51.2	54.9	58.2	41.9	43.6	36.8
RetinaNet[32]	51.0	53.9	43.0	55.5	52.4	61.2	42.4	45.5	42.1
Tan et al.[33]	56.8	52.4	42.8	60.6	57.4	**67.7**	50.3	43.1	40.7
Duan et al.[34]	52.1	**54.3**	44.7	51.0	55.2	65.7	47.4	40.9	37.5
Zhou et al.[35]	54.2	50.5	44.1	53.5	55.8	64.7	48.2	41.9	44.0
Wang et al.[13]	57.3	–	–	–	–	–	–	–	–
3DE, 3 slices[9]	50.6	43.4	42.4	52.2	54.3	63.3	42.6	42.1	42.3
3DE, 9 slices[9]	54.4	49.2	**46.8**	57.7	56.4	66.3	48.0	44.1	47.0
Ours	**57.5**	47.8	45.3	**66.9**	**63.7**	66.4	**52.9**	**46.3**	**52.0**

The most competitive counterparts are ULDor[8] and 3DCE[9], which also adopt the two-stage pipeline in the backbone network. ULDor uses pseudo masks and hard negative example mining strategy to enhance accuracy. Regretfully, this is still inferior to proposed detector. 3D contextual information is introduced during training and testing in 3DCE, making the whole detector much more sophisticated and therefore time-consuming. When the number of input slices is increased to 9, 3DCE is already much slower than our proposed detector, even as the majority of feature maps of slices could be cached and

reused for the next inference. 3DCE with 27 input slices achieves the best result among different methods. The sensitivity at 4 FPs per image (which is a commonly used standard for comparison) is 85.65% compared to our detector which was 85.07%, demonstrating that our detector is slightly inferior from this perspective. In spite of this, our detector runs in real-time (28ms, 35.6FPS) and is about three times faster than that with 114ms, 8.8FPS. To make the results of comparison more intuitive, we generated FROC curves of several methods at the top of Fig 8. The sensitivity and corresponding inference time of several methods are also scattered and depicted at the bottom of Fig 8.

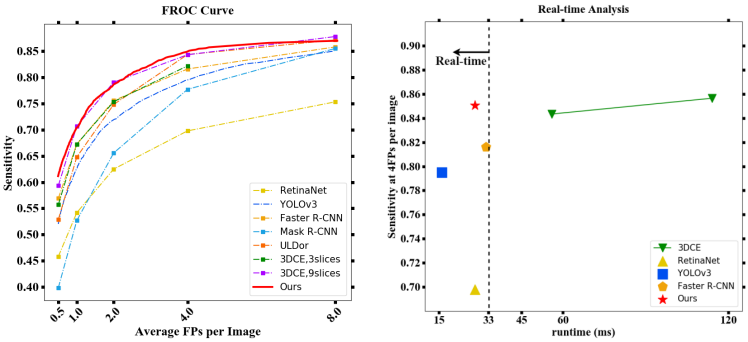

Fig. 8 FROC Curves of various methods (left); Real-time analysis of various methods (right) on the official split test set of DeepLesion.

Comparisons concerning mAP and APs are provided in Table 4, where the similar phenomena and trend could be discovered. In particular, Wang et al. [13] proposed a universal detector across almost all image domains (11 different datasets), establishing a universal detection benchmark. Although Wang et al.[13] sacrifices a huge amount of time and resources during training, the mAP associated with our detector is 57.5%, which is still comparable to 57.3%. Secondly, our detector outperforms others when detecting lesions of mediastinum and liver, with an improvement of more than 7 points in AP.

In summary, our detector achieves the best trade-off between accuracy and efficiency by comparing with other state-of-the-art detectors. Specifically, our proposed method not only achieves high accuracy in medical CT-image detection, but also maintains a relatively good detection speed.

5 Conclusion

In this paper, we propose an accurate one-stage universal lesion detector based on YOLOv3. In data preprocessing stage, a new data augmentation strategy is proposed to highlight the lesion area and avoid over-fitting risk. Besides, our detector could fuse more semantic and sptial complementary information, allowing the network to focus on the lesion area and obtain more accurate detection results. Experiments conducting with DeepLesion dataset confirm

the effectiveness of such contributions. By comparison with state-of-the-art detectors, we achieved comparable results with regard to Sensitivity at Various FPs per image and mAP, while maintaining ideal inference speed.

References

[1] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A.: Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **68**(6), 394–424 (2018)

[2] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28**, 91–99 (2015)

[3] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)

[4] Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)

[5] Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., Berg, A.C.: Dssd: Deconvolutional single shot detector. arXiv preprint arXiv:1701.06659 (2017)

[6] Lee, S.-g., Bae, J.S., Kim, H., Kim, J.H., Yoon, S.: Liver lesion detection from weakly-labeled multi-phase ct volumes with a grouped single shot multibox detector. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 693–701 (2018). Springer

[7] Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of medical imaging **5**(3), 036501 (2018)

[8] Tang, Y.-B., Yan, K., Tang, Y.-X., Liu, J., Xiao, J., Summers, R.M.: Uldor: a universal lesion detector for ct scans with pseudo masks and hard negative example mining. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 833–836 (2019). IEEE

[9] Yan, K., Bagheri, M., Summers, R.M.: 3d context enhanced region-based convolutional neural network for end-to-end lesion detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 511–519 (2018). Springer

[10] Chiao, J.-Y., Chen, K.-Y., Liao, K.Y.-K., Hsieh, P.-H., Zhang, G., Huang,

T.-C.: Detection and classification the breast tumors using mask r-cnn on sonograms. Medicine **98**(19) (2019)

[11] Ding, J., Li, A., Hu, Z., Wang, L.: Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 559–567 (2017). Springer

[12] Dou, Q., Chen, H., Yu, L., Qin, J., Heng, P.-A.: Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. IEEE Transactions on Biomedical Engineering **64**(7), 1558–1567 (2016)

[13] Wang, X., Cai, Z., Gao, D., Vasconcelos, N.: Towards universal object detection by domain attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7289–7298 (2019)

[14] Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010)

[15] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755 (2014). Springer

[16] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)

[17] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)

[18] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)

[19] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: European Conference on Computer Vision, pp. 21–37 (2016). Springer

[20] Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)

[21] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[22] Cao, G., Xie, X., Yang, W., Liao, Q., Shi, G., Wu, J.: Feature-fused ssd: Fast detection for small objects. In: Ninth International Conference on Graphic and Image Processing (ICGIP 2017), vol. 10615, p. 106151 (2018). International Society for Optics and Photonics

[23] Li, Z., Zhou, F.: Fssd: feature fusion single shot multibox detector. arXiv preprint arXiv:1712.00960 (2017)

[24] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)

[25] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

[26] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

[27] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)

[28] Li, X., Wang, W., Hu, X., Yang, J.: Selective kernel networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 510–519 (2019)

[29] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[30] Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: European Conference on Computer Vision (2018)

[31] Fe I, W., Jiang, M., Chen, Q., Yang, S., Tang, X.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)

[32] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)

[33] Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object

detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10781–10790 (2020)

[34] Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., Tian, Q.: Centernet: Keypoint triplets for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6569–6578 (2019)

[35] Zhou, X., Zhuo, J., Krahenbuhl, P.: Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 850–859 (2019)