Radio-Assisted Human Detection

Chengrun Qiu, Dongheng Zhang, Yang Hu, Houqiang Li, *Fellow, IEEE,* Qibin Sun, *Fellow, IEEE,* and Yan Chen, *Senior Member, IEEE*

Abstract—In this paper, we propose a radio-assisted human detection framework by incorporating radio information into the state-of-the-art detection methods, including anchor-based one-stage detectors and two-stage detectors. We extract the radio localization and identifer information from the radio signals to assist the human detection, due to which the problem of false positives and false negatives can be greatly alleviated. For both detectors, we use the confidence score revision based on the radio localization to improve the detection performance. For two-stage detection methods, we propose to utilize the region proposals generated from radio localization rather than relying on region proposal network (RPN). Moreover, with the radio identifer information, a non-max suppression method with the radio localization constraint has also been proposed to further suppress the false detections and reduce miss detections. Experiments on the simulative Microsoft COCO dataset and Caltech pedestrian datasets show that the mean average precision (mAP) and the miss rate of the state-of-the-art detection methods can be improved with the aid of radio information. Finally, we conduct experiments in real-world scenarios to demonstrate the feasibility of our proposed method in practice.

Index Terms—Human detection, radio localization, two-stage detector, anchor-based one-stage detector.

I. INTRODUCTION

Human detection is a fundamental problem in computer vision, which can power many other vision tasks such as instance segmentation [1], [2] and pose estimation [3], [4]. Generally, existing detection methods first extract the features from the images and then output the bounding boxes together with the corresponding confidence scores. Compared with the traditional methods [5], [6], deep learning based detection methods can achieve much better performance [1], [7] since the deep neural backbone can extract more useful features from the images.

However, problems still exist in the existing detection methods. As shown in Figure 1, there are mainly two problems: false positives and false negatives. The false positives stand for the bounding boxes that do not correctly cover an object or cover an object which has been already detected, *e.g.*, the green bounding boxes, while the false negatives represent the miss detections for the objects, *e.g.*, the blue bounding boxes. Note that most of the state-of-the-art detection methods use the confidence score to filter the detections, *i.e.*, if the confidence score is larger than a pre-defined threshold, the detection will be displayed in the image. Thus, the false positives are those

Chengrun Qiu and Dongheng Zhang are with School of Information and Communication Engineering, University of Electronic Science and Technology of China, E-mail: {cr_qiu, eedhzhang}@std.uestc.edu.cn.

Yang Hu and Houqiang Li are with School of Information Science and Technology, University of Science and Technology of China, Email:{eeyhu,lihq}@ustc.edu.cn.

Qibin Sun and Yan Chen are with School of Cyberspace Security, University of Science and Technology of China, Email:{qibinsun,eecyan}@ustc.edu.cn.
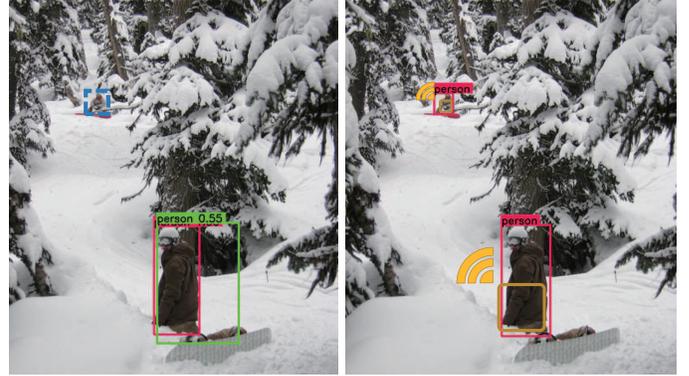
Fig. 1: The false positives (false detections) and false negatives (miss detections) still exist in the state-of-the-art detection methods (left figure). The false positives stand for the bounding boxes that do not correctly cover a person or cover a person which has been already detected, *e.g.*, the green bounding boxes, while the false negatives represent the miss detections for the persons, *e.g.*, the blue bounding boxes. With the aid of radio information, *e.g.*, the orange regions, the proposed method can well alleviate the false positives and false negatives (right figure).

"incorrect" bounding boxes with confidence scores larger than the pre-defined threshold, while the false negatives are those "correct" bounding boxes but with confidence scores smaller than the pre-defined threshold.

On the other hand, with the development of wireless communications and internet of things, radio signals are nowadays pervasive in our daily life. For example, people always carry smart phones, and more and more objects are attached with RFID to track their statuses [8], [9], [10]. Such radio signals can provide us a unique identifier and localization information for each person. The identifier can be the MAC address or the RFID information, while the localization can be estimated through the angle of arrival (AoA) and time of flight (ToF) information extracted from the channel state information (CSI) of the radio signals. The question now is how to utilize the identifier and localization information to improve the detection performance?

In this paper, we propose a radio-assisted human detection framework by incorporating the identifier and localization information obtained from the radio signals. We first align the localization (*i.e.*, the AoA and ToF) information with the image according to the camera's effective focal length (EFL) and field of vision (FOV), and obtain an initial estimation of each person. For both the anchor-based one-stage detectors and two-stage detectors, we generate different ways to revise the

detection's confidence score based on the localization results. For the two-stage detectors, we further propose to replace the region proposal network (RPN) with multi-scale anchors generated from the initial estimation as proposals.

Moreover, we utilize the number of identifiers as a constraint for the number of persons to be detected. Based on the specific constraint, we propose an improved non-maximum suppression (NMS) to guarantee that the detections are born from different localizations, and the number of the detections will not exceed the number of the localizations. It is worth pointing out that, with the proposed NMS, there is no need to use the confidence threshold to process the visible detections. Thus, the proposed framework can greatly alleviate the problems of false positives and false negatives in final detections.

We evaluate the proposed method on both simulative datasets and real-world scenarios. The simulative datasets are constructed based on the COCO and Caltech pedestrian datasets. Specifically, we reshape the ground-truth bounding boxes into squares and randomly shift the positions and sizes to simulate the initial estimation from the radio localizations. The experimental results show that, compared with the state-of-the-art detection methods, the proposed method can achieve better detection results in terms of both the mAP metric and the miss rate (MR) versus false positives per image (FPPI) metric. The experimental results on the real-world scenarios also demonstrate that the proposed method can greatly alleviate the problems of the false positives and false negatives, *i.e.*, the false and miss detections.

II. RELATED WORK

Object detection: The state-of-the-art object detectors can be classified as either two-stage detectors [1], [11], [12], [13], [14] or one-stage detectors [7], [15], [16], [17], [18], [19]. For two-stage detectors, the first stage is to filter out the background anchors to generate a sparse set of object proposals, while the second stage is to classify the proposals as well as refining the bounding boxes to obtain the final detections. Early two-stage detectors such as R-CNN [11] select regions of interest with traditional algorithms to construct proposals, while the latest two-stage detectors are equipped with RPN for more efficient and accurate proposals [1], [12]. Although the development of two-stage detectors may differ in multi-scale features [14], [20], [21], regression layers [13] and balance training [14], [22], their region proposals are all similar, which can be replaced by the initial estimations from the radio signals.

Some one-stage detectors are anchor-based methods [7], [15], whose idea is to classify and refine each cell's multi-scale anchors and output the final detections with NMS. For example, YOLO[7] divides the image into multiple cells and directly refines the pre-defined multi-scale anchors located at each cell as the detection bounding boxes. For the one-stage detector, the anchor stands for the proposal boxes. Recently, anchor-free detectors have raised more and more attention, some of which rely on keypoint detection to output the bounding boxes [17], [18], while others mainly tackle the problem by dense prediction to predict the center of each object [16], [19].

There have been works focusing on object detection tasks with extra information. For example, [23] and [24] use extra context from the image to improve the detection performance. The authors in [25] utilize the depth image as extra information while the authors in [26] propose to utilize the extra thermal data. Generally, the performance can be improved when the extra information is carefully utilized.

Because the state-of-the-art detectors use confidence score (classification score) to estimate the detection, for the average precision (AP) metric, if the confidence scores of the false positives are lower than those of the true positives, the detection is recognized as a good result. However, in such a case, false positives with enough confidence displayed in the image will still bother the users with their observation. By simply rising the confidence score threshold to process the result, some correct detections with lower score will be removed instead. Therefore a detector performs well in COCO dataset with the mAP evaluation metric may not provide clear visual result.

Radio localization: The radio localization methods can be classified as received signal strength indication (RSSI) based approaches [27], [28], [29], AoA based approaches [30], [31], [32], and ToF based approaches [33], [34], [35]. All these approaches measure the RSSI, AoA, or ToF from the target at multiple antennas and localize the target through triangulation. The RSSI based approaches measure the RSSI from the target at multiple access points and locate the target by combining the RSSI via triangulation with a propagation model. With RSSI data of multiple access point, the system can work out the distance of the object to each antenna and finally achieve the coordinate of the object. The ToF localization is similar as the RSSI localization, which also uses the distance of the object to each antenna to get the object's destination. The AoA based approaches work out the AoA of the direct path of each receiver to the target. Similar as the RSSI based methods, the AoA based approaches require multiple estimations at different access points and use the triangulation to localize.

The joint AoA-ToF estimation is considered in [36], [37], [38], [39], [40], where the localization of the target can be estimated with the AoA and ToF from one single antenna. If the environment is equipped with a single access point, we can provide the accurate localization with the joint AoA-ToF estimation. The position of the object can be worked out with the angle and the distance of the object from the antenna receiver.

Radio-video fusion: Several literatures have attempted to combine radio signals with video data for tracking/localization tasks [41], [42], [43], [44], [45], [46], [47], [48]. Ishihara *et al.* in [41] proposed a network for BLE signals and integrated it with the PoseNet. [42] focused on minimizing the root mean squared error of the predicted tracking. Their method shows better accuracy for the camera's indoor position and rotation. In [43], the authors fused the vision and wireless modalities and improved the localization and tracking of individuals. With the received signal strength (RSS) from individual's mobiles, individual's tracking and localization can be accomplished with the RGB view and a corresponded wireless ring image. Zhao et al. implemented an easy-to-deploy system with multi-

modal data for indoor localization [46], which achieved 92-percentile error within 0.2m for indoor targets. Wang et al. performed localization with images and RFID tags and achieved 6.23cm localization error [47], while Bai et al. proposed visible light communication-assisted indoor localization that achieved less than 3cm positioning error[48]. Zhao *et al.* in [44] trained the RF encoder and decoder networks for human pose estimation with the image-based keypoint method as supervision, while Li *et al.* in [45] proposed a neural network to predict the human's pose with the input of both RGB image and radio heatmap.

Different from the aforementioned radio-video fusion work, our work focuses on utilizing the identifier and localization information extracted from the radio signals to improve the performance of the state-of-the-art object detection methods.

III. OUR METHODS

A. Radio localization and imaging

Joint AoA-ToF estimation. We assume that each person to be detected is equipped with an RF component that can emit radio signals. This assumption is reasonable given the development of wireless communications, *e.g.*, people nowadays always carry smart phone. The receiver is equipped with a vertical and horizontal antenna array, which is capable of estimating vertical and horizontal AoA of the person. The signal from a person with a specific AoA-ToF can be expressed as

$$P(\theta, \tau) = \sum_{m=0}^{M} \sum_{k=0}^{K} s_{m,k} e^{j2\pi f_k \frac{mdcos\theta}{c}} e^{j2\pi k\Delta f \tau}, \quad (1)$$

where θ and τ denote the AoA and ToF, respectively, m and k are the indexes of antenna and frequency, d is the inter-element space of antenna array, c is the speed of light and Δf is the frequency interval. Then, the AoA-ToF can be estimated by

$$(\hat{\theta}, \hat{\tau}) = \arg\max_{\theta, \tau} |P(\theta, \tau)|, \quad (2)$$

where $(\hat{\theta}, \hat{\tau})$ denotes the estimated AoA-ToF. In other words, we could estimate the AoA-ToF by pick the peak with the highest amplitude shown in Figure 2.

Radio imaging. With the estimated AoA-ToF, the 3D locations of the person to the camera can be obtained with the camera imaging model. Specifically, with the estimated ToF, we first derive the distance of the person to the camera, as shown in Figure 3. Then, we calculate the point-to-plane distance L with the horizontal and vertical AoA, α and β. With the ratio of L to the camera's focal length l, the sizes of the estimated regions in the image can be calculated. The coordinate of the person can then be known with the horizontal angle, the vertical angle and the focal length by using the tangent. In this way, we can obtain the initial estimate of the area occupied by a person.

B. Radio localization aware detectors

While the existing detection methods have achieved state-of-the-art detection performance, the false positives and false negatives problems still commonly exist in the final detection
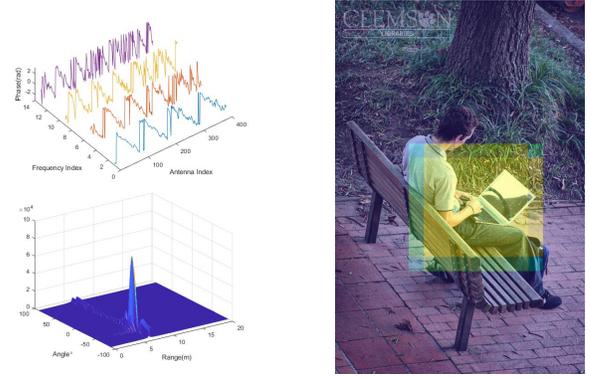

Fig. 2: Joint AoA-ToF estimation from radio signals. **Left:** The raw CSI is sampled on different frequencies and antennas shown in the figure above. It is further transformed into AoA-ToF domain as shown in the figure below. The peak corresponds to the device location with highest confidence. **Right:** An initial estimated region with the radio information.
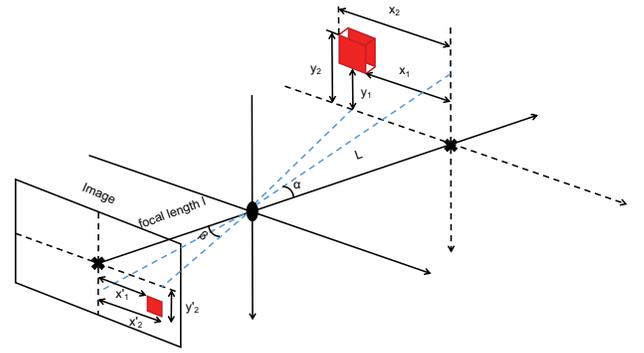

Fig. 3: Radio imaging to the image plane with the estimated AoA and ToF.

results. In this subsection, we discuss how to incorporate the radio localization information into the detectors to alleviate the false positives and false negatives, the details of our proposed radio localization guided methods will be introduced, respectively. The pipeline of the whole proposed radio-assisted human detection is shown in Figure 4.

Method 1: Radio confidence revision. We utilize the localization information from radio signals to revise the confidence scores of the detection results before NMS. In general, if the bounding box correctly covers the target, its confidence score should be larger than those of the boxes which only partly include the target. Thus, we introduce a decay factor, $\gamma \in [0, 1]$, for each detection bounding box. For anchor-based one-stage detectors, given a radio-assisted region in an image, γ is defined as the normalized intersection between the radio region and the detection's corresponding divided cell in the backbone over the cell, *i.e.*,

$$\gamma = \frac{area(localization) \bigcap area(cell)}{area(cell)}. \quad (3)$$

For two-stage detectors, we adjust each detection's confidence score using the intersection between the bounding box
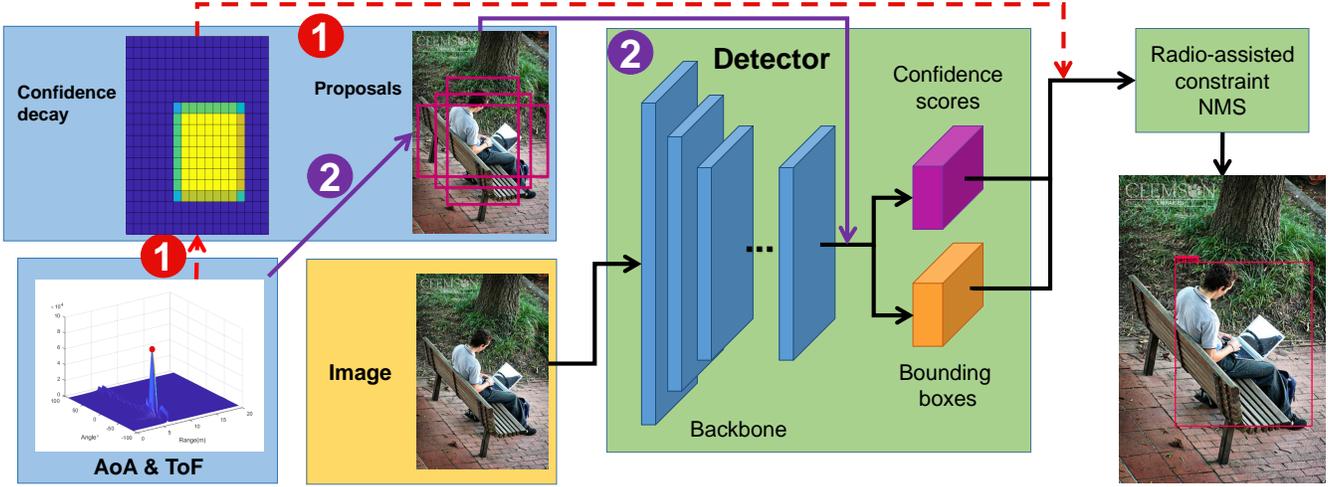
Fig. 4: Pipeline of the radio-assisted human detection. **Method 1:** The pipeline of the detectors with radio-assisted confidence revision. **Method 2:** The pipeline of the two-stage detector with radio-assisted proposals.

and the localization area over the localization area, *i.e.*,

$$\gamma = \frac{area(bbox) \bigcap area(localization)}{area(localization)}. \qquad (4)$$

Then, the new confidence score is obtained as follows

$$S_{new} = (1 - \lambda + \lambda * \gamma) * S, \qquad (5)$$

where λ is a pre-defined constant defined by the expected accuracy of the radio localization, *i.e.*, a larger λ means a higher expected accuracy. The case where $\lambda = 0$ reflects that the method is only determined by the traditional detection framework, while the case where $\lambda = 1$ means the revised score is processed by directly multiplying with the decay factor.

Method 2: Radio region proposals for two-stage detectors. For the two-stage detectors, we can directly replace the RPN with the radio region proposals. The RPN in two-stage detectors is in charge of selecting and refining proposals from multi-scale anchors to filter out the proposals of the background. For our detectors, the radio localization provides the multi-scale anchors as region proposals whose centers and sizes are determined by the radio imaging discussed in the previous subsection. If we only focus on the single-class person detection, the anchor shape and ratio can be designed in advance based on the people's potential postures and general size. With the proposed method, by refining the anchors only once, the detector is capable of outputting accurate detections. Note that by replacing the RPN with the radio region proposals, our method is capable of speeding up the two-stage detectors.

C. NMS with radio region constraint

An NMS method with the number constraint has been proposed in [49], where the constraint is not from the radio localization. In [49], the authors choose the final detections by maximizing the sum of their scores with the constraint of the scores' count and the Intersection over Union (IoU) of the detections. IoU specifies the amount of overlap between the

---

Algorithm 1: Radio-assisted constraint NMS

Input: Detection boxes sorted with confidence
$\mathcal{B} = \{b_1, b_2, ..., b_N\}$,
sorted confidence scores $\mathcal{S} = \{s_1, s_2, ..., s_N\}$,
radio-assisted regions $\mathcal{W} = \{w_1, w_2, ..., w_M\}$,
root regions of the boxes $\mathcal{R} = \{r_1, r_2, ..., r_N\}$,
IoU threshold $\mathcal{T}$

begin
  $\mathcal{R}_b \leftarrow \{\}, \mathcal{R}_s \leftarrow \{\}, \mathcal{R}_w \leftarrow \{\}, i = 0$
  while $i < |\mathcal{B}|$ *and* $|\mathcal{R}_w| < |\mathcal{W}|$ **do**
    $i = i + 1$
    if $r_i \in \mathcal{R}_w$ *or* $iou(\mathcal{R}_b, b_i) \geq \mathcal{T}$ **then**
      | continue
    end
    $\mathcal{R}_b \leftarrow \mathcal{R}_b \bigcup \{b_i\}, \mathcal{R}_s \leftarrow \mathcal{R}_s \bigcup \{s_i\},$
    $\mathcal{R}_w \leftarrow \mathcal{R}_w \bigcup \{r_i\}$
  end
  if $|\mathcal{R}_w| < |\mathcal{W}|$ **then**   Supplied for radio
    $\overline{\mathcal{W}} = \mathcal{W} - \mathcal{R}_w$ proposal detectors
    for w_i *in* $\overline{\mathcal{W}}$ **do**
      | $\mathcal{R}_b \leftarrow \mathcal{R}_b \bigcup \{b_{w_i}\}, \mathcal{R}_s \leftarrow \mathcal{R}_s \bigcup \{s_{w_i}\}$
    end
  end
  return $\mathcal{R}_b, \mathcal{R}_s$
end

Fig. 5: NMS with the one-on-one constraint of the radio-assisted region and the detection bounding box. The procedure marked red is only feasible for detectors with the radio region proposal input.

predicted and ground truth bounding box, which is an important metric to evaluate the performance of object detection. In this paper, we consider the constraint that each bounding box should be born from different radio localizations. The radio localization constraint could be recognized as a reinforced constraint of the number of persons because each localization finally will be matched with at most one bounding box. The details of the proposed radio region constraint NMS algorithm is illustrated in Figure 5.

In the proposed NMS algorithm, the first loop is a similar process as the common NMS. The main improvement of the first loop is that one radio localization will at most output one bounding box. The condition $iou(\mathcal{R}_b, b_i) \geq \mathcal{T}$ represents that the IoU of b_i and any bounding box in $\mathcal{R}_b$ is not smaller than the threshold $\mathcal{T}$. The condition $r_i \in \mathcal{R}_w$ refers that the radio region r_i of the current detection b_i has already existed in the result set, which is set to satisfy the constraint that one radio localization at most matches one detection bounding box. With the first procedure, it can be guaranteed that the neighbourhood boxes all lie out of the pre-defined overlap and different bounding boxes are born from different radio localizations.

For the two-stage detectors with the radio region proposal input, it is easy to know each detection comes from which proposal and belongs to which radio region. For the one-stage detectors with radio confidence mask, it is difficult to know the responsible radio region because the detections are the end-to-end results and one cell/detection may be covered by multiple radio regions. In such a case, we formulate the relation by the IoU of each single radio region and the detection box.

The second loop in Figure 5 is enabled if the number of the detections is smaller than the number of the radio localizations. This may happen when some correct detections are suppressed by unreasonable IoU threshold. With the second loop, the algorithm will output the detection generated from a pre-defined anchor of the missing radio region. This loop is implemented only for the two-stage detectors with the radio proposal input. For the one-stage detectors, it cannot be guaranteed that every region of radio localization is allocated at least one detection with the IoU relation. In such a case, we skip the second process and the proposed NMS can only ensure that the number of the final detections will not exceed the number of the radio localizations.

IV. EXPERIMENTS

Extensive experiments on simulative datasets and real-world scenarios are conducted to verify the effectiveness of our method by comparing with the state-of-the-art detection methods. For experiments on detection datasets, we simulate the wireless localization region from the ground truth and conduct detections. Firstly, we study the influence of the localization deviation (AoA and ToF) on the detection precision. We then compare the performance of existing detectors with our radio localization aware detectors on these datasets. Note that as shown in section 3.2, we have proposed two radio localization aware detectors: one is radio confidence revision, denoted as "**Proposed Method 1**"; the other is radio region proposal,

denoted as "**Proposed Method 2**". Specifically, for two-stage detectors equipped with radio region proposal, we choose the region in the image mapped from the radio localization and design three scales of the anchor to match the human potential postures. For the radio confidence revision, based on the localization region, we adjust each detection's confidence score as (3) and (4). Finally for real data, the wireless localization results together with the detection results in a real-word scenario are presented and the detections of the existing detectors are also illustrated for comparison.

A. Experimental Setup

Datasets: The simulative datasets are constructed based on COCO [50] and Caltech pedestrians [51] datasets. For COCO, we conduct detection tasks for 80 categories and person category and evaluate the detectors on the *val2017* set. For Caltech pedestrian, we implement human detections on every 30 frame in the video sequences in validation *set06* to *set10*.

In practice, the AoA and ToF estimated from the radio localization system may deviate from the ground truth due to the multipath interference in the environment. Since we do not have the prior knowledge about the shape of the person, we first reshape the ground-truth bounding box into a square region whose edge length L equals $\min(H, W)$. Then, to imitate the ToF estimation errors in radio localization, we manufacture independent Gaussian distributed noise ζ to the edge length L of each square region as below

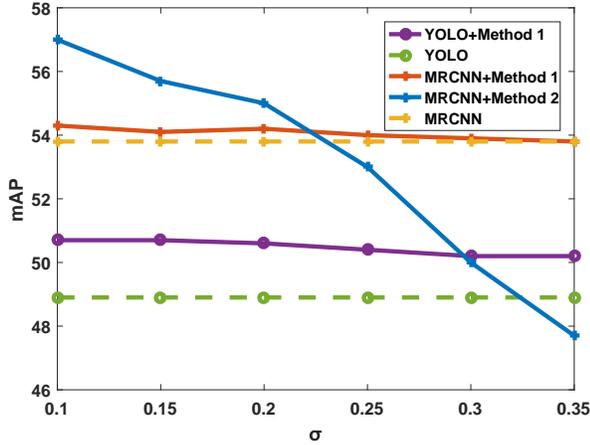$$L' = L \times \zeta, \qquad \zeta \sim \mathcal{N}(1, \sigma), \qquad (6)$$

where σ is a pre-defined standard deviation, $\mathcal{N}$ denotes the Gaussian distribution. Finally, to imitate the AoA estimation errors, we give the center of each square region a random shift related to its edge length L' as

$$x' = x + \xi_1, \qquad y' = y + \xi_2,$$
$$\xi_1 \sim \mathcal{N}(0, k_1 L'), \qquad \xi_2 \sim \mathcal{N}(0, k_2 L'), \qquad (7)$$
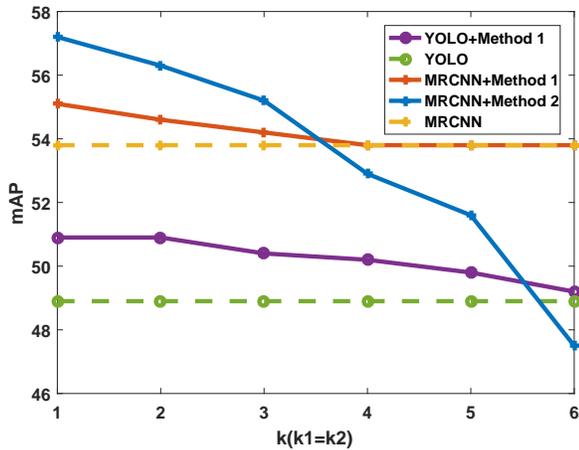
where ξ_1 and ξ_2 are the random shifts along the x and y direction, respectively, k_1 and k_2 are pre-defined standard deviations.

Metrics: We adopt two widely used metrics for detection performance: the mean average precision (mAP) defined in COCO [50] and the miss rate (MR) versus false positives per image (FPPI) curve defined in Caltech pedestrian [51]. The mAP evaluates the mean value of the AP with the requirement of the IoU interval [.5:.05:.95] between the output and ground-truth bounding boxes. The MR vs FPPI curve draws the MR curve under different FPPI and finally obtains the average MR to evaluate the detectors. The runtime is also evaluated to demonstrate that our method could improve not only the detection performance but also the detection efficiency.

Implementation details: We apply our method on various two-stage detectors including Mask R-CNN, Libra R-CNN with the backbone ResNet-101 and Grid R-CNN with the ResNext-101-32x4d backbone. We adopt the model weights in [52]. The confidence score threshold and the IoU threshold of the NMS in Mask R-CNN and Libra R-CNN are 0.05

(a) Effect of ToF estimation errors (person category).

(b) Effect of AoA estimation errors (person category).

Fig. 6: Effect of radio localization error.

and 0.5, respectively. In Grid R-CNN, the NMS's confidence threshold is 0.03 and the IoU threshold is 0.3 based on [53]. We also apply our method on the anchor-based one-stage detector YOLOv3 on Keras with the model weight in [54]. The backbone is Darknet-53 in [54] with the 608×608 input image size and the NMS's confidence threshold and the IoU threshold being 0.05 and 0.5 respectively. We perform the experiments on a single Geforce GTX 1080Ti GPU to measure the runtime.

B. Effect of radio localization error

We first analyze the effect of radio localization error on the mAP when the simulative COCO dataset is used. The influences of scale to the region size and shift to the region center are illustrated in Figure 6(a) and Figure 6(b) respectively. When conducting the analysis of σ or $k(k_1 = k_2)$ in Eqn. (6) or (7), the other parameter is set as $k = 0.1$ or $\sigma = 0.2$. By comparing Figure 6(a) and 6(b), we observe that the AoA error has a more significant influence on the mAP than the ToF error. Therefore, it's more important to provide a precise coordinate of each person's center by accurate AoA estimation. From the figure, we can also see that with

higher accuracy, the radio-assisted radio proposal in two-stage detectors performs better than the confidence adjustment. While on the other hand, the radio-assisted confidence revision is more robust than the radio-assisted region proposal at the lower localization accuracy. This phenomenon may be due to the fact that two-stage detectors crop features in the proposals to the classification and regression layers, and thus the region proposal requires higher localization precision.

C. Ablation study

In this subsection, we conduct the ablation experiments to show that both the radio region proposal/confidence revision and our proposed NMS are effective. The parameters of radio localization errors are set as: $\sigma = 0.2$, $k_1 = k_2 = 0.1$.

COCO mAP evaluation. Table I and Table II show the ablation results on COCO for person category. For the human detection as shown in Table I, we can see that with the help of radio-assisted confidence revision, the mAP performance for person category can be improved about 0.4%, 1.2%, 0.6% and 1.4% for Mask R-CNN, Grid R-CNN, Libra R-CNN, and YOLOv3, respectively. With radio localization region proposal, the performance can be raised for about 2% and 0.6% for Mask R-CNN and Libra R-CNN, although the Grid R-CNN's performance may be slightly worse.

We also compare the mAP performance in human detection while we constrain the number of the detections by the count of the ground truth instead of using confidence threshold. In this case we can conduct a fair performance evaluation on our proposed NMS method. The results are shown in Table II. In the results, we find that if the number of the detections is limited, the improvement of the radio-assisted methods is much more obvious. Furthermore, we can see that equipped with our proposed NMS, the detectors' performance can even increase about 2% for confidence revision and 3% for localization's region proposal. In Table II, we also make performance comparison with a mentioned NMS in [49], which is an NMS for weakly supervised localization with object's count constraint. And we find that their NMS doesn't perform well in our person category detection tasks for it only optimizes the specific case where detected bounding boxes are loose and contain two or more object instances.

MR vs FPPI evaluation.

To give enough penalty on the false positives, we employ the MR v.s. FPPI metric [51]. The MR v.s. FPPI curves of human detection on the COCO dataset are shown in Figure 7(a) to evaluate the ability of the proposed method on suppressing false positives while conducting correct detections, where the numerical percentages in the legend stand for the average miss rates. The results show that the average miss rate and the FPPI at low miss rate are both improved with the proposed method. Compared with the mAP metric in Table II, the improvement of the average miss rate in Figure 7(a) can be explicitly observed.
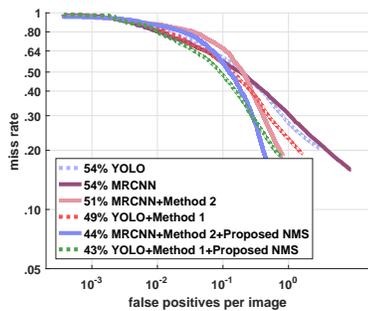
We also evaluate the Caltech pedestrian dataset with the 'reasonable' experiment and the 'all' experiment, respectively. In the 'reasonable' experiment, the ground truth only contains people of height more than 60 pixels and occlusion ratio

TABLE I: Ablation studies of the radio-assisted detectors on COCO *val-2017* for *person* category.
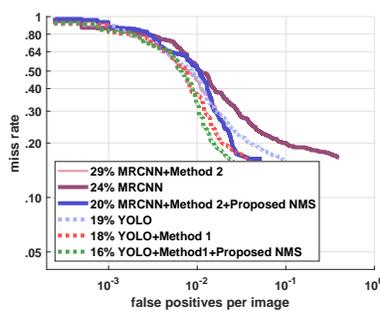
Methods	radio localization	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
Mask R-CNN		53.8	83.3	58.4	36.3	**61.8**	**70.1**
	proposed method 1	54.2	86.0	58.8	39.7	61.7	67.8
	proposed method 2	**55.8**	**87.0**	**61.2**	**41.1**	61.3	68.6
Grid R-CNN		56.5	82.8	60.7	38.4	63.3	**74.4**
	proposed method 1	**57.7**	**85.4**	**62.2**	**42.2**	**64.0**	73.7
	proposed method 2	56.5	83.0	61.0	40.7	61.8	71.2
Libra R-CNN		54.7	83.3	59.6	36.3	62.0	**71.6**
	proposed method 1	**55.3**	**86.4**	60.1	**40.3**	**62.2**	69.6
	proposed method 2	**55.3**	84.5	**61.0**	38.9	61.5	69.6
YOLOv3		48.9	82.4	52.2	30.7	56.4	66.2
	proposed method 1	**50.3**	**84.6**	**53.7**	**33.3**	**58.0**	**66.6**

TABLE II: Ablation studies of the radio-assisted detectors on COCO *val-2017* for *person* category with detection's count constraint.

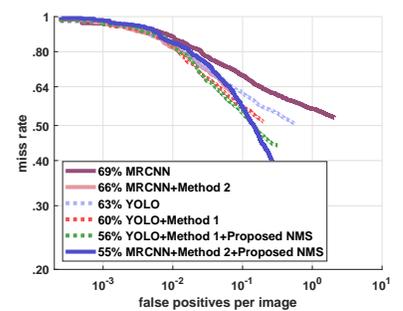| Methods	radio localization	NMS	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
Mask R-CNN			49.3	74.3	54.6	30.8	57.2	68.4
	proposed method 1		50.1	80.6	54.0	35.4	57.4	63.7
	proposed method 1	GRS[49]	49.5	77.9	54.1	35.6	57.3	61.3
	proposed method 1	proposed	52.0	82.7	57.1	37.7	58.9	65.5
	proposed method 2		54.5	85.1	59.3	39.8	59.8	68.1
	proposed method 2	GRS[49]	52.0	80.1	56.8	36.1	58.7	64.9
	proposed method 2	proposed	**58.3**	**93.2**	**63.0**	**45.3**	**63.4**	**70.2**
Grid R-CNN			52.4	75.1	56.9	33.0	59.1	72.7
	proposed method 1		54.8	81.2	59.3	38.5	61.2	71.0
	proposed method 1	GRS[49]	51.5	75.1	56.0	36.9	58.9	65.7
	proposed method 1	proposed	56.3	82.7	61.1	40.6	62.3	72.5
	proposed method 2		56.3	83.1	61.1	41.4	61.3	70.8
	proposed method 2	GRS[49]	53.6	79.7	57.9	38.4	59.7	66.8
	proposed method 2	proposed	**59.1**	**88.4**	**63.1**	**43.4**	**64.3**	**74.0**
Libra R-CNN			50.6	75.3	56.6	31.3	57.8	69.9
	proposed method 1		51.4	80.8	56.0	36.2	58.6	65.9
	proposed method 1	GRS[49]	51.1	78.4	56.2	36.7	58.6	64.0
	proposed method 1	proposed	53.1	82.7	58.1	38.3	60.2	67.1
	proposed method 2		55.1	83.9	61.0	39.5	61.2	68.8
	proposed method 2	GRS[49]	51.6	79.5	56.4	34.8	58.7	64.9
	proposed method 2	proposed	**58.9**	**92.0**	**64.0**	**44.7**	**64.3**	**71.9**
YOLOv3			46.8	77.9	50.5	28.0	54.2	65.6
	proposed method 1		48.9	81.4	53.1	31.6	56.3	66.0
	proposed method 1	GRS[49]	47.1	79.9	49.5	30.0	54.7	63.5
	proposed method 1	proposed	**50.8**	**86.4**	**54.1**	**34.6**	**58.0**	**66.0**

(a) MR v.s. FPPI on COCO.

(b) MR v.s. FPPI on Caltech Pedestrian's 'reasonable' experiment.

(c) MR v.s. FPPI on Caltech Pedestrian's 'all' experiment.

Fig. 7: Miss rate versus false positives per image.

Fig. 8: Scenarios of the adopted validation dataset.

TABLE III: Performance of the radio-assisted detectors on our validation dataset.

Methods	Localization	Proposed NMS	FP & FN per image	True Detections Ratio(%)
Mask R-CNN			1.152	46.44
	Method 1	✓	**0.990**	**52.18**
	Method 2	✓	1.001	49.04
Grid R-CNN			1.003	48.28
	Method 1	✓	**0.984**	**52.42**
	Method 2	✓	1.088	46.25
Libra R-CNN			1.564	42.03
	Method 1	✓	**0.990**	**52.82**
	Method 2	✓	1.004	48.71
YOLOv3			0.499	69.87
	✓	✓	**0.329**	**75.03**

less than 35%, while in the 'all' experiment, detectors should detect people of height more than 20 pixels and occlusion ratio less than 80%. The results are shown in Figure 7(b) and Figure 7(c). Similar to the results on the COCO dataset, both the 'reasonable' and 'all' experiments reflect that the proposed method can efficiently reduce the average miss rate. Furthermore, it can be observed from the figures that with our proposed NMS, the FPPI is significantly decreased at low miss rates, which demonstrates that our proposed NMS can reduce the number of the false positives.

D. Experiments in real-world scenarios

Finally, we conduct experiments in the real-world scenarios to verify the feasibility of our method in practice.

To make detection evaluation with real localization data, we conduct our validation dataset with synchronized image and localization data. We captured videos in three scenarios with two cameras and the total length of videos is about 15 minutes. We collected the synchronized localization data with one localization device to work out the coordinates of the people. The camera we used is the camera of HUAWEI honor8lite with the horizontal FOV 64° and the vertical FOV 52°. The sizes of the captured videos are 1280×720 pixels and the pixel-measured focal length approximately equals to 3000. We divided the videos into nearly 1000 frames and some of them are shown in Figure 8. The environment contains laboratory and open spaces in overcast and evening, and people's activities include walking, jogging, hugging and jumping. In order to increase the difficulty of detection, we chose to capture the image data in the scenarios where the brightness is not efficient or some obstacles (*i.e.* chairs and tripods) exist in the image which cover half part of one person at most (*i.e.* head, legs or body). In one frame there are at most three people and a data frame may also be empty, *i.e.*, it does not include any person.

To collect the localization data, we used the TI board of MMWCAS-RF-EVM to collect the RF signals, and adopted Equation (1) to work out the coordinates of the people in the localization space. Finally we calculated the length and the angle of each person to each camera.

To show the main advantage of our method, we use two other evaluation methods in our dataset. Because our methods focus more on the final visual performance in detection tasks, the evaluation methods consider all of the detection bounding boxes without sorting them in descending order
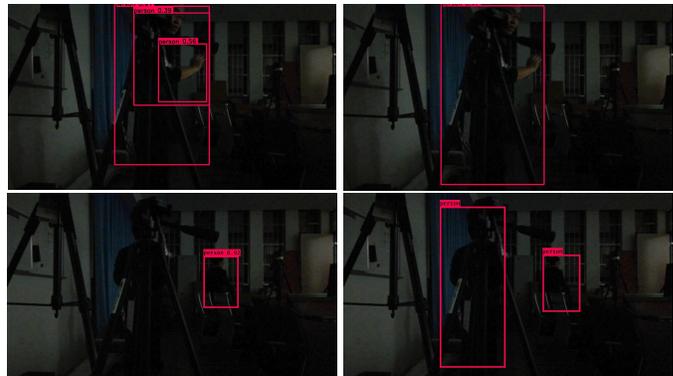

Fig. 9: Detection results comparison of our dataset. The left column is the detection of the original detector. The right column is the detector with our proposed methods.

by the confidence score. One evaluation is to calculate the false positives and false negatives per image, the other we called true detection ratio is to calculate the ratio between true positives to the sum of true positives, false positives and false negatives, *i.e.*

$$\frac{true\ positives}{true\ positives + false\ positives + false\ negatives}. \quad (8)$$

Figure 9 shows the performance comparison of our dataset. The above row shows the result comparison of mask R-CNN detector while the below row shows the result comparison of YOLO detector. Observing the above row, we can find it is the situation of false positive. The sum of the false positives and negatives in the left is 2 (two boxes on the same person) and the true detection ratio is $1/3$. The below row stands for the situation of false negative. The sum of the false positives and negatives in the left is 1 (one person's detection is missed) and the true detection ratio is $1/2$. Both of our adopted evaluation methods will be influenced by redundant detection and missed detection without considering confidence score. And with our proposed methods, we find the evaluation will be improved while the visual performance gets better as well.

Table III shows the performance improvement of our methods on our dataset and the confidence thresholds of the original detection methods are both 0.3. We can find that with the localization and the proposed NMS, the false positives and false negatives can be restrained, while the ratio of the true positives can be increased. Different from the results in simulated COCO dataset, we find that in this case the
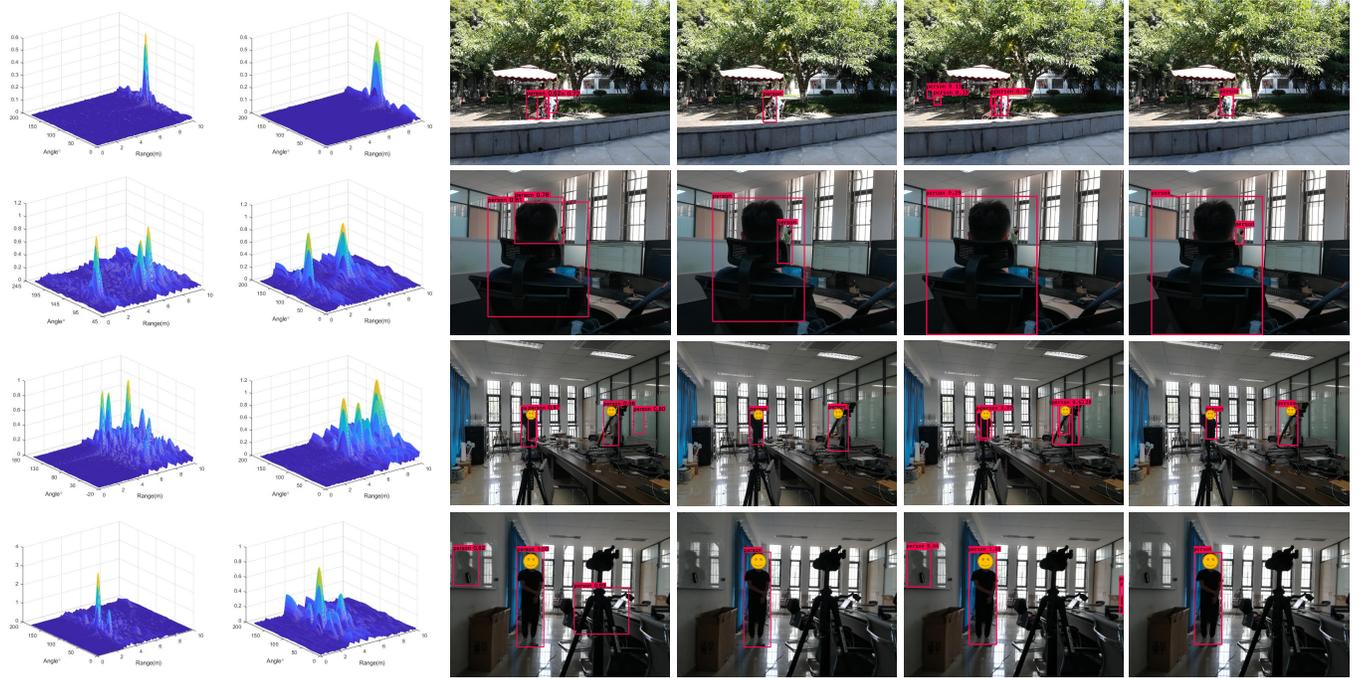
Fig. 10: Detection results in real-world scenarios. **First column:** Estimated horizontal AoA-ToF heatmap. In real-world environment, the heatmap would be disturbed by multi-path interference. Nevertheless, the final detection would not be affected since the AoA-ToF is only utilized as te initial estimate. **Second column:** Estimated vertical AoA-ToF heatmap. Similar to the horizontal heatmap, the vertical heatmap is also disturbed by multi-path interference. **Third column:** Results of Mask R-CNN. **Fourth column:** Results of Mask R-CNN equipped with proposed method 2. **Fifth column:** Results of YOLOv3. **Sixth column:** Results of YOLOv3 equipped with proposed method 1.

proposed method 1 performs better. This is because in real world the localization precision may not be as accurate as the localization result simulated in COCO dataset.

Figure 10 illustrates more results in real-world scenarios. Similar to previous experiments, we conduct the real-world detection with the original Mask R-CNN detector, the original YOLOv3 detector, and our proposed detectors. For Mask R-CNN, we set the confidence threshold to 0.5 to make the detection results more visible. The peaks which are the potential object locations will be chosen if their magnitudes exceed half of the highest magnitude. The results clearly demonstrate that with radio information, our proposed detectors can well address the problems of false positives and false negatives, compared with the original Mask R-CNN and YOLOv3.

V. CONCLUSION

In this paper, we proposed a human detection framework with the aid of radio information for anchor-based one-stage detector and two-stage detectors. Systematically, based on the radio signals, we first estimated the localization of each person in the image with its angle and distance from the camera. Then, we proposed two ways to utilize the radio localization information for anchor-based one-stage detector and two-stage detectors. With the radio identifier and localization information, we also proposed a non-maximum suppression with an extra constraint that the radio localizations and the detections should be one-on-one matched. Experiments on simulative datasets and real-world scenarios showed that our proposed

methods could improve the performance of state-of-the-art detectors and alleviate the problem of false positives and false negatives.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 386–397, 2020.

[2] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[3] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "Openpose: Realtime multi-person 2d pose estimation using part affinity fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172–186, 2021.

[4] H. S. Fang, S. Xie, Y. W. Tai, and C. Lu, "Rmpe: regional multi-person pose estimation," in *IEEE International Conference on Computer Vision*, 2017.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[6] P. Viola and M. J.Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[7] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[8] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of things for smart cities," *IEEE Internet of Things Journal*, vol. 1, no. 1, pp. 22–32, 2014.

[9] A. J. Jara, M. A. Zamora-Izquierdo, and A. F. Skarmeta, "Interconnection framework for mhealth and remote monitoring based on the internet of things," *IEEE Journal on Selected Areas in Communications*, vol. 31, no. 9, pp. 47–65, 2013.

[10] S. Amendola, R. Lodato, S. Manzari, C. Occhiuzzi, and G. Marrocco, "Rfid technology for iot-based personal healthcare in smart spaces," *IEEE Internet of Things Journal*, vol. 1, no. 2, pp. 144–152, 2014.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[13] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan, "Grid r-cnn," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[14] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[15] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Europeon Conference on Computer Vision*, 2016.

[16] L. Huang, Y. Yang, Y. Deng, and Y. Yu. (2015) Densebox: Unifying landmark localization with end to end object detection. [Online]. Available: https://arxiv.org/abs/1509.04874

[17] X. Zhou, J. Zhuo, and P. Krähenbühl, "Bottom-up object detection by grouping extreme and center points," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[18] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Europeon Conference on Computer Vision*, 2018.

[19] Z. Tian, C. Shen, H. Chen, and T. He, "Fcos: A simple and strong anchor-free object detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[20] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[21] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2018.

[22] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[23] H. Qiu, H. Li, Q. Wu, F. Meng, L. Xu, K. N. Ngan, and H. Shi, "Hierarchical cntext features embedding for object detection," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3039–3050, 2020.

[24] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 944–954, 2017.

[25] N. Huang, Y. Liu, Q. Zhang, and J. Han, "Joint cross-modal and unimodal features for rgb-d salient object detection," *IEEE Transactions on Multimedia*, 2021.

[26] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 160–173, 2020.

[27] P. Bahl and V. N. Padmanabhan, "Radar: An in-building rf-based user location and tracking system," in *IEEE International Conference on Computer Communications*, 2000.

[28] V. Bianchi, P. Ciampolini, and I. D. Munari, "Rssi-based indoor localization and identification for zigbee wireless sensor networks in smart homes," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 2, pp. 566–575, 2019.

[29] M. T. Hoang, B. Yuen, X. Dong, T. Lu, R. Westendorp, and K. Reddy, "Recurrent neural networks for accurate rssi indoor localization," *IEEE Internet of Things journal*, vol. 6, no. 6, pp. 10 639–10 651, 2019.

[30] H. J. Shao, X. P. Zhang, and Z. Wang, "Efficient closed-form algorithms for aoa based self-localization of sensor nodes using auxiliary variables," *IEEE Transactions on Signal Processing*, vol. 62, no. 10, pp. 2580–2594, 2014.

[31] Y. Sun, K. C. Ho, and Q. Wan, "Eigenspace solution for aoa localization in modified polar representation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2256–2271, 2020.

[32] M. A. G. Al-Sadoon, R. Asif, Y. I. A. Al-Yasir, R. A. Abd-Alhameed, and P. S. Excell, "Aoa lcalization for vehicle-tracking systems using a dual-band sensor array," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 8, pp. 6330–6345, 2020.

[33] S. A. Golden and S. S. Bateman, "Sensor measurements for wi-fi location with emphasis on time-of-arrival ranging," *IEEE Transactions on Mobile Computing*, vol. 6, no. 10, pp. 1185–1198, 2007.

[34] A. Marcaletti, M. Rea, D. Giustiniano, V. Lenders, and A. Fakhreddine, "Filtering noisy 802.11 time-of-flight ranging measurements," in *International Conference on emerging Networking EXperiments and Technologies*, 2014.

[35] J. Wang, Q. Gao, Y. Yu, X. Zhang, and X. Feng, "Time and energy efficient tof-based device-free wireless localization," *IEEE Transactions on Industrial Informatics*, vol. 12, no. 1, pp. 158–168, 2016.

[36] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-level localization with a single wifi access point," in *13th USENIX Symposium on Networked Systems Design and Implementation*, 2016.

[37] F. Wen and C. Liang, "Fine-grained indoor localization using single access point with multiple antennas," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1538–1544, 2015.

[38] Z. Chen, G. Zhu, S. Wang, Y. Xu, J. Xiong, J. Zhao, J. Luo, and X. Wang, "M3: Multipath assisted wi-fi localization with a single access point," *IEEE Transactions on Mobile Computing*, vol. 20, no. 2, pp. 588–602, 2021.

[39] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "Breathtrack: Tracking indoor human breath status via commodity wifi," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 3899–3911, 2019.

[40] W. Gong and J. Liu, "Roarray: Towards more robust indoor localization using sparse recovery with commodity wifi," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1380–1392, 2019.

[41] T. Ishihara, K. M. Kitani, C. Asakawa, and M. Hirose, "Deep radio-visual localization," in *IEEE Winter Conference on Applications of Computer Vision*, 2018.

[42] S. Papaioannou, A. Markham, and N. Trigoni, "Tracking people in highly dynamic industrial environments," *IEEE Transactions on Mobile Computing*, vol. 16, no. 8, pp. 2351–2365, 2017.

[43] A. Alahi, A. Haque, and L. Fei-Fei, "Rgb-w: When vision meets wireless," in *IEEE International Conference on Computer Vision*, 2015.

[44] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-wall human pose estimation using radio signals," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[45] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, "Making the invisible visible: action recognition through walls and occlusions," in *IEEE International Conference on Computer Vision*, 2019.

[46] Y. Zhao, J. Xu, J. Wu, J. Hao, and H. Qian, "Enhancing camera-based multimodal indoor localization with device-free movement measurement using wifi," *IEEE Internet of Things Journal*, vol. 7, no. 2, pp. 1024–1038, 2020.

[47] Z. Wang, M. Xu, N. Ye, F. Xiao, R. Wang, and H. Huang, "Computer vision-assisted 3d object localization via cots rfid devices and a monocular camera," *IEEE Transactions on Mobile Computing*, vol. 20, no. 3, pp. 893–908, 2019.

[48] L. Bai, Y. Yang, M. Chen, C. Feng, C. Guo, W. Saad, and S. Cui, "Computer vision-based localization with visible light communications," *IEEE Transactions on Wireless Communications*, 2021.

[49] M. Gao, A. Li, R. Yu, V. I. Morariu, and L. S. Davis, "C-wsl: Count-guided weakly supervised localization," in *Europeon Conference on Computer Vision*, 2018.

[50] T. Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," in *Europeon Conference on Computer Vision*, 2014.

[51] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: a benchmark," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[52] 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. (2019) Mmdetection: Open mmlab detection toolbox and benchmark. [Online]. Available: https://arxiv.org/abs/1906.07155

[53] X. Lu, B. Li, Y. Yue, Q. Li, and J. Yan. (2019) Grid r-cnn plus: Faster and better. [Online]. Available: https://arxiv.org/abs/1906.05688

[54] J. Redmon and A. Farhadi. (2018) Yolov3: an incremental improvement. [Online]. Available: https://arxiv.org/abs/1804.02767