GSDet: Object Detection in Aerial Images Based on Scale Reasoning

Wei Li, Wei Wei, *Senior Member, IEEE*, and Lei Zhang, *Member, IEEE*

***Abstract*— Variations in both object scale and style under different capture scenes (*e.g.*, downtown, port) greatly enhance the difficulties associated with object detection in aerial images. Although ground sample distance (GSD) provides an apparent clue to address this issue, no existing object detection methods have considered utilizing this useful prior knowledge. In this paper, we propose the first object detection network to incorporate GSD into the object detection modeling process. More specifically, built on a two-stage detection framework, we adopt a GSD identification subnet converting the GSD regression into a probability estimation process, then combine the GSD information with the sizes of Regions of Interest (RoIs) to determine the physical size of objects. The estimated physical size can provide a powerful prior for detection by reweighting the weights from the classification layer of each category to produce RoI-wise enhanced features. Furthermore, to improve the discriminability among categories of similar size and make the inference process more adaptive, the scene information is also considered. The pipeline is flexible enough to be stacked on any two-stage modern detection framework. The improvement over the existing two-stage object detection methods on the DOTA dataset demonstrates the effectiveness of our method.**

***Index Terms*— Object detection, aerial images, ground sample distance, reasoning.**

I. Introduction

AS AN increasing number of aerial images become available, object detection, which aims at recognizing and locating the objects in high-resolution aerial images, has become a key task in aerial image analysis [1], [2]. Due to the capture conditions, such as the bird's-eye view perspective, object detection in aerial images is more challenging than in natural images [3], [4]. More specifically, the diverse ground sample distance (GSD) and variation in captured scenes lead to huge appearance variations among each category, *i.e.*, objects of the same category exhibit substantial difference in local appearance or scale, while objects of different categories look similar, shown as in Figure 1. In order to handle those difficult-to-classify samples (referred to as hard samples in this study), higher semantic and global scene information is considered generally. Reference [5] adopts a top-down pyramid structure to pass contextual information from high-level features, while [6] first utilizes the global information of the whole image to improve the classification accuracy of target. Reference [7] uses a multi-scale fusion feature to supply additional surrounding information. Though the above works demonstrate the importance of auxiliary features and scene information, these methods focus only on how to get higher-quality proposal feature representations (*i.e.*, more matching feature scales and higher semantic information), while neglecting the targets' physical meanings. By contrast, humans can identify objects in complex situations even weakly supervised, (*i.e.*, only image-level annotation is provided for detection), [8]–[12], owing to the help of commonsense knowledge. Inspired by this, it is realized that the key to improve object detection performance in complex scenes is to make full use of the characteristics of aerial images and combine them with human common sense, with which the object detection performance can be further improved.

Recent works on utilizing commonsense knowledge in detection tasks can be categorized into those methods that rely on the relationships between objects and those that rely on the attributes of objects and scenes. For instance, [13]–[15] combine a base detection network with spatial relation information to obtain better performance for hard samples. However, due to the isolation of objects in aerial images, (*i.e.*, the limited spatial affiliation between objects), the spatial relationships among different categories are not as significant as those in natural scenes such as [16], [17]. Therefore, these methods only propose to contain intensive objects relying on clustering information alone, which limits the usage of such methods in aerial images. Other methods take advantage of objects' physical attributes and the characteristics of the scene. [13], [18] design several properties explicitly, (*e.g.*, color,

Manuscript received May 21, 2020; revised November 14, 2020; accepted March 29, 2021. Date of publication April 22, 2021; date of current version April 29, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 62071387 and Grant 61671385 and in part by the Science, Technology and Innovation Commission of Shenzhen Municipality under Grant JCYJ20190806160210899. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shuicheng Yan. *(Wei Li and Wei Wei contributed equally to this work.) (Corresponding author: Wei Wei.)*

Wei Li is with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xian 710072, China (e-mail: liw@mail.nwpu.edu.cn).

Wei Wei is with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xian 710072, China, also with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xian 710072, China, and also with the Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518031, China (e-mail: weiweinwpu@nwpu.edu.cn).

Lei Zhang is with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, School of Computer Science, Northwestern Polytechnical University, Xian 710072, China, and also with the National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xian 710072, China.

Digital Object Identifier 10.1109/TIP.2021.3073319

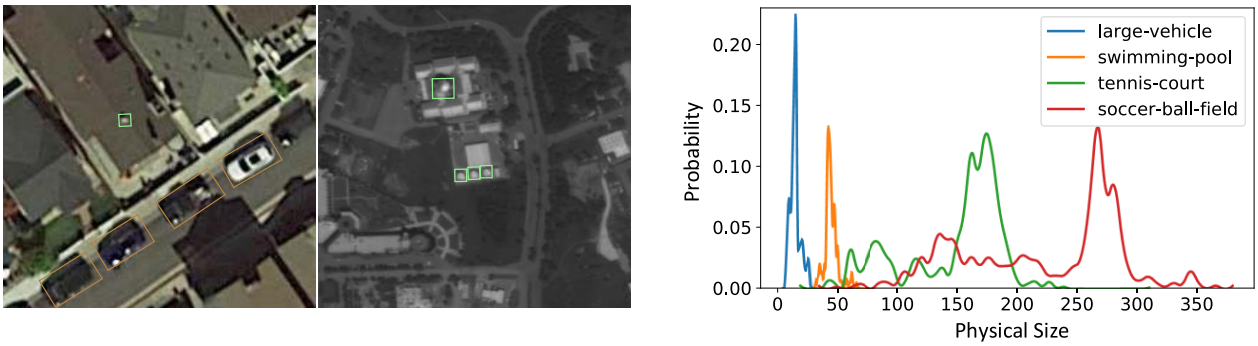


Fig. 1. Left images give the detection results of the detection model without using GSD information. Although the appearance of the target in the left image looks like a storage tank, human can easily identify it is not based on the physical size of object in the image. Right image is the statistical result of physical size relative to categories, which shows great diversity for different categories.

status), then conduct a knowledge graph with a large number of categories to improve the performance of classes with rare instances. However, such attributes describing such patterns have limited representation abilities when applied to aerial images with complex backgrounds and various capture scenes. In addition, the scarcity of categories hinders the applicability of such methods. Therefore, how to appropriately use attributes suitable for aerial images based object detection (especially for the reasoning in the detection) remains a challenging problem.

Despite the drawbacks (such as special capture scenes), there are also some unique characteristics for object detection in aerial image. Due to the consistency in capture height, the metrics (e.g., area, perimeter) of objects in images can reflect their actual physical sizes. In addition, there is obvious diversity among the physical sizes of objects in different categories, as shown in Figure 1. Such attribute description is unique to aerial images, compared with natural scene images. Thus, utilizing the GSD, which is related to the capture height, can naturally be used to improve the object detection performance. Furthermore, objects appear in a specific scene always has obvious regularity, *e.g.*, vehicles are usually located in urban areas, while boats are usually located in ports, which can also provide rich prior information. To take advantages of those rich prior knowledges contained in aerial images (e.g., GSD), we propose the first object detection network to incorporate the GSD and scene information into object detection modeling.

For this purpose, we build our method based on a two-stage network incorporating both the GSD prior knowledge and the scene prior knowledge. First, we propose a novel GSD identification network based on multi-branch network with different dilated rate for each branch, with which we can estimate the GSD information for each image from its shallow features. Since GSD measures the physical size of each pixel in the captured image, we can calculate the actual physical size of a object by multiplying the bounding box enclosing it in the image with the GSD. The obtained physical size is then mapped into the class-wise prior distribution, which is served as the GSD prior knowledge for each proposal (*i.e.*, the probability of a given RoI belonging to each category). However, distribution regarding size is prone to be close between categories with similar physical sizes, (*e.g.*, large-vehicle and ship), which will influence the performance of the proposed method. Then, We accordingly utilize the scene information to address the above problems. More specifically, we adopt a squeeze-excitation structure to produce category-wise attention (*i.e.*, the scene prior knowledge), which is capable of handling those samples that are difficult to be distinguished using GSD alone. To realize the reasoning process with the scene prior, we create a feature collection to store high-level semantic representations. Finally, based on the GSD prior knowledge and scene prior knowledge mentioned above, feature collection is soft-mapped to produce the final enhanced features for each given proposal feature. It should be noted that the enhanced feature will be concatenated with the original proposal features to improve the classification performance in an end-to-end manner. The main contributions of this paper can be summarized as follows:

- We propose to leverage the GSD to improve the object detection performance in aerial images. **To the best of our knowledge, this is the first paper to attempt the incorporation of GSD into the** object detection method.
- We propose a novel GSD-identifying network based on dilated convolution in order to effectively utilize the GSD, thereby achieving better object detection performance.
- Our method incorporates both GSD prior knowledge as well as scene prior knowledge, and is flexible enough to be stacked on any two-stage detection framework. We conduct our experiments on the DOTA [19] dataset, and report the best object detection performance, compared with other state-of-the-art methods.

II. Related Work

A. Object Detection in Aerial Images

Current CNN-based object detection methods can be divided into anchor-based and anchor-free detectors. The former can be further subdivided into one-stage methods [20], [21] and two-stage methods [22], [23]. Both of them preset a large number of fixed reference anchors in different sizes and positions, then predict the category and refine the coordinates of these anchors one or several times, before finally outputting these refined boxes as detection results. Representative two-stage methods include R-CNN [24], Fast R-CNN [25],

Faster R-CNN [22] and so on. The first stage of these methods produces numbers of candidate boxes, after which the second stage classifies these boxes into either foreground or background classes. R-CNN [24] extracts CNN features from the candidate regions and applies SVM as the classifier. To obtain higher speed, Fast R-CNN [25] proposes a novel ROI-pooling operation that extracts feature vectors for each candidate box from a shared convolutional feature map. Faster R-CNN [22] integrates proposal generation with the second-stage classifier to form a single convolution network. Recently, one-stage detectors like SSD [20] and YOLO [26]–[28] have been proposed that achieve real-time detection with satisfactory accuracy. Because two-stage methods refine anchors several times more than one-stage methods, the former achieves more accurate results while the latter has higher computational efficiency. Anchor-free detectors directly find objects without preset anchors in two different ways, *i.e.*, keypoint-based methods (Cornernet [29], ExtremeNet [30]) and center-based methods (FCOS [31]). The former first locate several pre-defined or self-learned keypoints, then bound the spatial extent of objects. The latter use the center point or region of objects to define positives, then predict four distances from these positives to the object boundary. Although anchor-free detectors are able to eliminate those hyperparameters related to anchors and have achieved similar performance to anchor-based detectors, well-designed preset anchors provide a more stable training process. Therefore, state-of-the-art results on common detection benchmarks are still achieved by anchor-based detectors.

One significant difference between the detection on aerial images and natural ones is that the former aims at regressing an oriented bounding box, while the latter needs a horizontal one. To apply the natural image detection model to aerial images, FR-O [19] and R2CNN [32] supply an angular regression term in the second stage, which can be adaptively combined with two-stage detection model. RRPN [33] generates an inclined proposal with orientation information. Different with the existing aerial image based object detection methods, we attempt the incorporation of the characteristics of aerial image such as GSD information for detection, and propose a first object detection network which incorporates both GSD prior knowledge as well as scene prior knowledge.

B. Enhancement Method via Reasoning

Contextual and prior information can be utilized for reasoning accompanied with object detection, with which better object detection results can be obtained. Inspired by this, a number of detection methods have been proposed to utilize contextual and prior information for reasoning. Reference [16] reweights the classification loss function according to the spatial information to focus on the proposal located in the specific area. Reference [13], [34] directly modify the proposal feature to incorporate context. Reference [13] adopts two GRUs to store the learned knowledge, while [34] collects classifier parameters to generate a class-wise semantic pool. Since we aim to provide proposal-wise prior classification probabilities, we adopt a similar feature enhancement strategy to that proposed in [34]. Within a two-stage detection framework, we collect parameters from the classifier in the second stage as the raw category-wise knowledge and then reweight them as enhanced features, which avoids the computational burden relative to traditional methods [35]. Note that the difference between the proposed method and [34] is that we do not adopt any additional classifier to handle the enhanced features, *i.e.* the mapped parameters are fed iteratively back to the classifier, which can effectively simplify the class reasoning process.

C. Dilated Convolution

Dilated convolution [36] enlarges the convolution kernel with original weights by performing convolution at sparsely sampled locations, thus increasing the receptive field size without additional cost. Dilated convolution has been widely used in semantic segmentation to incorporate large-scale context information [36]–[39]. In the object detection field, DetNet [40] designs a specific detection backbone network to maintain the spatial resolution and enlarge the receptive field using dilated convolution. Deformable convolution [41] further generalizes dilated convolution by adaptively learning the sampling location. Moreover, POD [42] learns a stable global scale for each layer and transfers it to a fixed integral dilation using scale decomposition. In our work, we take advantage of the property that different dilation rates have different sensitivity to targets of the same size. In addition, a larger GSD of images is accompanied with smaller object pixel sizes. Inspired by these, we estimate the GSD information by fusing feature maps from filters with different dilation rates.

III. Methods

A. Overview

In this paper, we propose a two-stage detection framework that incorporates the GSD and scene information into the modeling process. The proposed GSDet is illustrated in Figure 2. The proposed detection network consists of one basic detection network module, a GSD subnet and a scene subnet. Basic detection network module accomplish the basic detection functions including RPN, RoI alignment, classification, and box regression, etc., as well as combine two kinds of prior probability information (i.e., physical size and scene information) for detection. The GSD subnet is exploited to predict the GSD information, from which the physical size prior can be provided for detection. Scene subnet is utilized to improve the reasoning ability among categories with similar physical sizes.

Utilizing ResNet as the feature extractor, low-level features from the base ResBlock is passed into the GSD subnet to predict the GSD probability distribution in logarithmic space. The probability distributions regarding the physical sizes of RoIs can be obtained by moving the distribution of GSD a specific distance, which is obtained via enforcing logarithm on the sizes of RoIs. From which, strong class-wise prior probability related with GSD can be extracted for object detection. Another subnet utilizes high-level features to obtain scene contextual information and provides class-wise attention for feature collections, which complements the discriminability of the GSD subnet on categories with similar physical size.

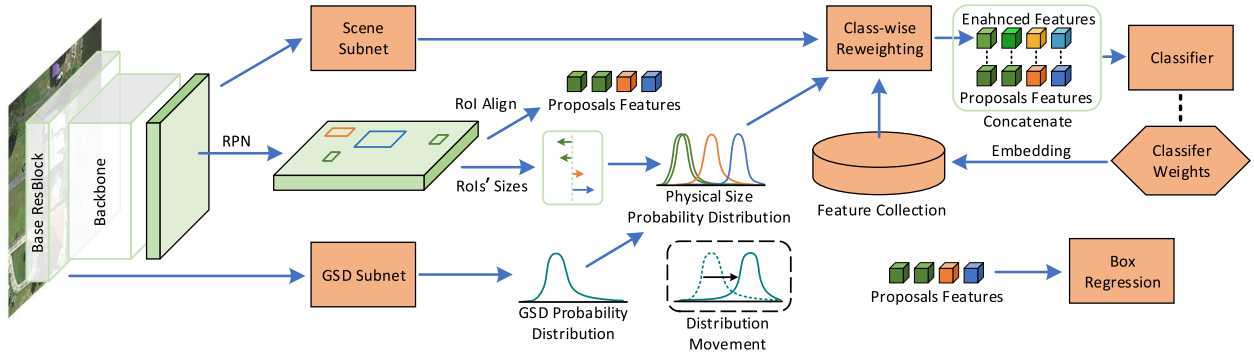


Fig. 2. Overview of the proposed GSDet. We build GSDet based on a two-stage detection framework. Shallow features from the base ResBlock are passed into the GSD subnet to predict the GSD probability distribution in logarithmic space. To obtain the probability distributions regarding physical size of RoIs generated by RPN, we enforce logarithm on the size of RoIs to obtain the expected movement distance relative to the GSD probability distribution. Furthermore, high-level features are passed into the scene subnet to extract scene information. Based on the information of scene and RoIs' sizes, we obtain class-wise reweighting factors for soft mapping feature collection from the weighting of classifier into enhanced features. Concatenated features are then utilized to improve the performance of classifier.

Finally, class-wise weighting factors obtained by combining the upper two subnets are used to soft-map the embedded contextual features into discriminable enhanced features, thereby further improving the object detection performance.

B. GSD Subnet

The prediction of GSD is essentially a regression task. With the true numerical GSD denoted as D, GSD regression aims to predict a value d that approximates D. It should be noted here that GSD regression focuses on regressing into a specific value utilizing the low-level texture features, which differs from previous regression tasks in object detection methods, (*e.g.*, bounding box regression depends on the response location on the feature map, while image depth estimation depends on the relative values of the feature map between areas). However, with high numbers of objects and scene styles within images, accurate regression of GSD is difficult and not robust, *i.e.*, the calculation of physical size is sensitive to the numerical value of GSD. We accordingly treat the regression problem as a classification problem, which is much easier to incorporate into neural networks and provides much richer probability information.

Considering that the magnification gaps of GSD are more representative than numerical gaps among images, (*e.g.*, the difference between paired images with 0.1 and 0.2 GSD is larger than that between 1.0 and 1.3 as the magnification of the former pair is 2.0 while the latter is 1.3, even though the numerical gap of the former pair 0.1 is smaller than the latter pair 0.3), we take the logarithm of GSD and obtain $\tilde{D} = \ln D$ to generate classification labels. In more detail, we numerically divide the logarithm space into K equal intervals, denoted as $\{I_i\}, i \in K$, and label the interval in which $\tilde{D}$ lies as positive while the others are deemed negative. By doing this, we convert continuous probability distributions into discrete ones. To guarantee good detection results, the following two aspects are utilized when we discretize the continuous probability distributions. First, GSD of aerial dataset always has clustering property, due to different capture altitude of the satellites. Taking DOTA dataset as an example, the images are captured from the Google Earth, the satellite of JL-1 and the satellite of GF-2. Thus, we can avoid the problem that GSD values fall at the junction of two intervals by utilizing such a clustering characteristic. Second, due to the existence of minor variance within the interval, converting a numerical probability distribution into a one-hot classification label will cause some distribution information to be sacrificed. This is because one-hot labeling encourages the output scores to be dramatically distinctive. As a result, we adopt label-smoothing regularization [43] to represent the labels. This procedure can be expressed as

$$q_i = \begin{cases} 1 - \varepsilon & \text{if } \ln \tilde{D} \in I_i, \\ \varepsilon / (K - 1) & \text{otherwise,} \end{cases} \tag{1}$$

where q_i is the probability for interval I_i and ε is the smooth factor. Supervised by this smooth label, the probability distribution learning can be more accurate and stable.

To accurately estimate the GSD information, we need to design an effective structure. For aerial images, a larger image GSD is accompanied with smaller object pixel sizes, which will also result in a smaller texture response scale. By contrast, the texture response area in images with smaller GSD will be larger. Low-level features are appropriate for handling such tasks and can make the inference independent of scene information contained in high-level features. In addition, considering object detection task and GSD identification task only has a weak relation, high-level features will enforce the relation of two tasks during learning, which will in turn deteriorate the performance of object detection as well as GSD identification. For above reasons, we utilize the low-level features output from the first block (i.e., base ResBlock) as the input of the GSD subnet in this study.

Inspired by the above analysis, we identify GSD by measuring the scale of the texture response area. Dilated convolutions [36] with different dilated rates are highly suitable to handle such a task. A higher dilated rate leads to a larger receptive field. As a consequence, convolutions with a higher dilated rate are insensitive to small-scale responses compared to those with a lower one. This results in a larger percentage of responses lost in images with higher GSD. Therefore,

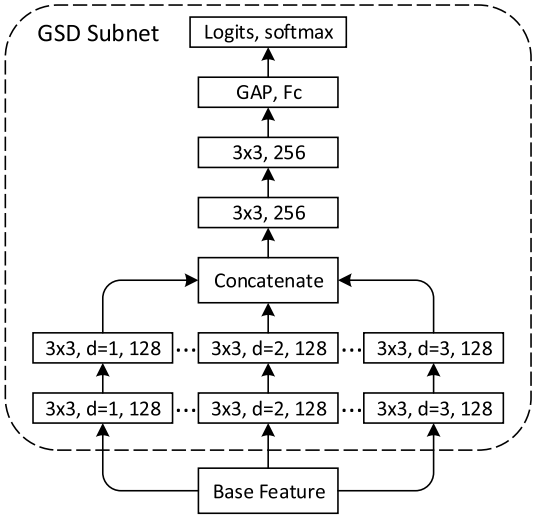


Fig. 3. The architecture of the GSD subnet. Dashed lines indicate parameter sharing. Kernel size, dilated rate and number of channels are indicated in the box. Each branch contains two dilated convolution blocks (consists sequentially of convolution with kernel size 3×3, ReLU, batch normalization and max pooling with 2×2 kernel size). Generated features are concatenated in channel level and passed into two serial convolution blocks. Global average pooling, fully connected layer and softmax operation are then applied on the resultant features sequentially to obtain the predicted discrete GSD distribution.

the diversity between responses of different branches can reflect the GSD information. We accordingly fuse features with different convolution dilated rates, enabling the GSD information to be extracted via learning process. As shown in Figure 3, the multiple branches with different dilation rates share the same parameters. Each branch contains two dilated convolution blocks (consisting sequentially of convolution with kernel size 3×3, ReLU, batch normalization and max pooling with 2×2 kernel size). Generated features are concatenated at channel level and passed into two serial convolution blocks. Global average pooling, fully connected layer and softmax operations are then sequentially applied to the resultant features to obtain the predicted discrete GSD distribution $\mathbf{q} = \left[q'_0, q'_1, \cdots, q'_K\right]$, where q'_i is the predicted probability for GSD that lies in the i-th interval I_i. We use the following cross-entropy function as the loss function

$$\mathcal{L}_{GSD} = -\sum_{k=1}^{K} \log\left(q'_k\right) q_k \tag{2}$$

C. Distribution Movement

Suppose that RPN samples N_r region proposals in one minibatch. We calculate the sum of the length and width of each proposal and obtain $\mathbf{s} = \left[s_1, s_2, \cdots, s_{N_r}\right] \in \mathbb{R}^{N_r}$. The reason we utilize the sum of the length and the width instead of area for calculation is given as follows. Considering an object enclosed by a square box may has the same area as the one enclosed by a rectangle box (with large aspect ratio), the area is not suitable to discriminate these kinds of objects. For this purpose, we utilize the sum of length and width instead the area in this study. After the calculation, the physical sizes of the proposals can be formulated as $\tilde{\mathbf{s}} = D\mathbf{s}$. In logarithmic space, $\ln\tilde{\mathbf{s}} = \ln\mathbf{s} + \tilde{D}$. Since the predicted probability distribution $p\left(\tilde{D}\right) \simeq \mathbf{q}$, the relation between $\tilde{\mathbf{s}}$ and $\mathbf{s}$ is equivalent to move $\mathbf{q}$ on the x-axis. For this purpose, we set a reference value δ to align the centers of original distribution $\tilde{\mathbf{s}}$ with the target one $\mathbf{s}$, and take $\mathbf{x} = \ln\mathbf{s} - \delta$ as the moving distance. The probability distribution regarding the physical size of regions can be formulated as

$$p\left(\tilde{\mathbf{s}}\right) = p\left(\tilde{D} + \mathbf{x}\right) \tag{3}$$

We denote the target distribution as $\mathbf{T} \simeq p\left(\tilde{\mathbf{s}}\right)$, $\mathbf{T} = \left[\mathbf{t}_1, \mathbf{t}_2, \cdots, \mathbf{t}_{N_r}\right], \mathbf{t} \in \mathbb{R}^L$, where $L = K + 2E$. E is the number of extended intervals on the left and right side to accommodate the distribution after the movement. To determine E, we first calculate the value of δ, which is used to minimize the total movement distance $|\mathbf{x}|$ based on the train dataset. Then, we round the upper bound of $\mathbf{x}$, i.e., $\lceil \max\left(\mathbf{x}\right) \rceil$ and take the result as E. In line with the above, we formulate the process from $\mathbf{q}$ to the physical size of the r-th proposal $\mathbf{t}_r$ as

$$t_i^r = q'_{i+\lfloor x_r \rfloor} + \left(x_r - \lfloor x_r \rfloor\right) q'_{i+\lfloor x_r \rfloor - 1} \tag{4}$$

$t_i^r \in \mathbf{t}_r$ is the value in i-th interval and $x_r \in \mathbf{x}$ is the moving distance for the r-th proposal. Note here that when the movement larger than E happens, the values of the distribution, which excess the extended intervals, will be accumulated to the marginal interval. From above operations, we explicitly obtain the probability distribution of physical size $\mathbf{t}$ for each proposal. For clarification, the distribution movement is illustated in Figure 4.

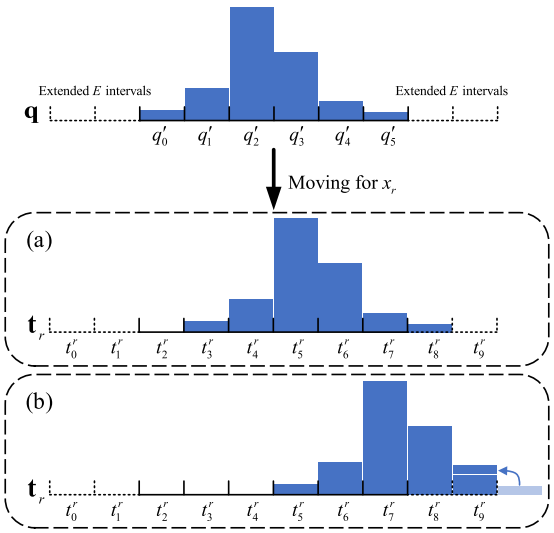


Fig. 4. Diagram of distribution movement. (a) shows the resultant distribution without excessing the extended intervals. (b) shows the resultant distribution when excessing the extended intervals.

With the obtained probability distribution of physical size $\mathbf{t}_r$ for each proposal, we first assemble those from all N_r proposal as $\mathbf{T} \in \mathbb{R}^{N_r \times L}$, then map $\mathbf{T}$ into the probability of classification for all C categories. Specifically, we adopt a fully connected layer $\mathbf{TW}_g$ to establish the mapping, where $\mathbf{W}_g \in \mathbb{R}^{L \times C}$ is shared for all proposals. The proposal-wise classification scores $\alpha \in \mathbb{R}^{N_r \times C}$ then can be obtained by $softmax\left(\mathbf{TW}_g\right)$.

D. Semantic Pool

Within a two-stage detection framework, we collect parameters from the classifier in the second stage as the raw category-wise knowledge to avoids the computational burden, and utilize them to generate the semantic pool. For this purpose, we collect the classifier parameters to store class-wise knowledge. Let $\mathbf{M} \in \mathbb{R}^{C\times N}$ denote the raw parameters for the classifier, where N is the number of parameters for each category. Despite the rich contextual information contained in raw parameters, they are still too redundant and sparse to be utilized directly as enhanced features. To address this problem, we adopt a transformation matrix $\mathbf{W}_c \in \mathbb{R}^{N\times d'}$ to embed the raw parameters into a low-dimensional space as our feature collection.

E. Scene Subnet

To improve the reasoning ability among categories with similar physical sizes, we take semantic pool from scene information into account to further re-weight the feature collection. Different from GSD information, the identification process requires high-level semantic features. Therefore we take the entire image feature $\mathcal{F} \in \mathbb{R}^{W\times H\times D}$ out of the backbone as the input of scene subnet. More specifically, features are squeezed into half their former size by convolution (with 3×3 kernels and $D/4$ output channels) and global average pooling layer. The output $\mathbf{z} \in \mathbb{R}^{D/64}$ is further combined with category knowledge by means of a matrix multiplication $\mathbf{z}\mathbf{W}_s\mathbf{M}^T$, where $\mathbf{W}_s \in \mathbb{R}^{D/64\times N}$. A soft-max function is then applied to obtain the category-wise attention $\beta = softmax\left(\mathbf{z}\mathbf{W}_s\mathbf{M}^T\right)$, $\beta \in \mathbb{R}^C$.

F. Class-Wise Reweighting

After generating the proposal-wise classification scores α and category-wise attention β, the enhanced features $\mathbf{f}' \in \mathbb{R}^{N_r\times d'}$ can be denoted as

$$\mathbf{f}'=\alpha\left(\beta \otimes \mathbf{M}\right)\mathrm{W_c} \tag{5}$$

where $\otimes$ denotes element-wise multiplication. The generation of enhanced features $\mathbf{f}'$ is formally described in Algorithm 1. Finally, we concatenate $\mathbf{f}'$ with $\mathbf{f}$ (i.e., the feature of the original proposal) as $\left[\mathbf{f};\mathbf{f}'\right]$, which is then fed into the classifier to promote the classification performance. The only difference between the proposed detector and the traditional detector is that the traditional detector utilizes the feature $\mathbf{f}$ of the original proposal for training, while we utilize the concatenated feature to train the classifier. Note that the enhanced features are calculated by classifier parameters. We obtain the enhanced feature from the batch in the previous training stage and utilize it for the current stage, which is an iterative process. The experimental results demonstrate the effectiveness of such a reweighting method.

G. Training

The proposed detection network consists of one basic detection network module, a GSD subnet and a scene subnet. To obtain satisfactory detection results, proper loss functions need to be defined before training. Within the modeled network, three supervised tasks are involved. They are category classification, location regression as well as GSD identification task. The loss function for these three tasks are termed as $\mathcal{L}_{cls}$, $\mathcal{L}_{loc}$ and $\mathcal{L}_{GSD}$, respectively. In this study, cross-entropy loss is adopted for classification tasks $\mathcal{L}_{GSD}$ as well as $\mathcal{L}_{cls}$, and l-1 loss for the regression task $\mathcal{L}_{loc}$.

We then sum the GSD loss function $\mathcal{L}_{GSD}$, classification loss function $\mathcal{L}_{cls}$ and location loss function $\mathcal{L}_{loc}$ in the detection task together with the weighting factor λ to obtain the overall loss function as

$$\mathcal{L}_T = \lambda\mathcal{L}_{GSD} + \mathcal{L}_{cls} + \mathcal{L}_{loc}. \tag{6}$$

Then, the proposed network can be jointly trained via minimizing above function.

Among these module and subnets, both basic detection network module and scene subnet are supervised by the annotation information. Therefore, we train the scene subnet jointly with the basic detection network module without pretrain. As for the the GSD subnet, it is supervised by the GSD. In addition, considering large amount of parameters (i.e., the parameters within the base ResBolock) are shared with the basic detection network module and fixed during the training of GSD subnet, the GSD subnet is pretrained alone. In applications, we first pretrain the GSD subnet. Then we jointly train the basic detection network module as well as these two subnets. The detail settings of training can be seen from Subsection IV.B

Algorithm 1 The Generation of Enhanced Features $\mathbf{f}'$

Input: $\mathbf{z}$ denotes the squeezed feature of the image,
$\mathbf{s}$ denotes the size of each proposal,
$\mathbf{q}$ denotes the predicted discrete GSD distribution,
$\mathbf{M}$ denotes the raw parameters of the classifier,
$\mathbf{W}_g, \mathbf{W}_s, \mathbf{W}_c$ denote the transformation matrices
Output: $\mathbf{f}'$ denotes the enhanced features

begin
  Initialize $\mathbf{T} \leftarrow \{\}$;
  for $z_r \in \mathbf{z}, s_r \in \mathbf{s}, r$ denotes the r-th proposal **do**
    Obtain moving distance for each proposal $x_r \leftarrow \ln s_r - \delta$;
    Initialize $\mathbf{t}_r \leftarrow \{\}$;
    for $q'_i \in \mathbf{q}, t_i^r \in \mathbf{t}_r, i$ denotes the i-th interval **do**
      Obtain the value of i-th interval in $\mathbf{t}_r$ $t_i^r \leftarrow q'_{i+\lfloor x_r\rfloor} + (x_r - \lfloor x_r\rfloor)\, q'_{i+\lfloor x_r\rfloor-1}$;
      Update $\mathbf{t}_r \leftarrow \mathbf{t}_r \cup \{t_i^r\}$;
    Update $\mathbf{T} \leftarrow \mathbf{T} \cup \{\mathbf{t}_r\}$;
  Obtain the proposal-wise classification score $\alpha \leftarrow softmax\left(\mathbf{T}\mathbf{W}_g\right)$;
  Obtain the category-wise attention$\beta \leftarrow softmax\left(\mathbf{z}\mathbf{W}_s\mathbf{M}^T\right)$;
  Obtain the enhanced feature $\mathbf{f}' \leftarrow \alpha\left(\beta \otimes \mathbf{M}\right)\mathrm{W_c}$;
  return $\mathbf{f}'$

Fig. 5. Visualization of the detection results from GSDet+Res in DOTA.

TABLE I

COMPARISONS WITH STATE-OF-THE-ART DETECTORS ON DOTA. THE SHORT NAMES FOR EACH CATEGORIES CAN BE FOUND IN SECTION 4.1. +RES STANDS FOR REPLACING FULLY CONNECTED LAYER IN GSDET WITH A RESBLOCK

Method	backbone	W/FPN	Plane	BD	Bridge	GTF	SV	LV	Ship	TC	BC	ST	SBF	RA	Harbor	SP	HC	mAP
FR-O	resnet101		79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.40	52.52	46.69	44.80	46.30	52.93
R2CNN	resnet101		80.94	65.75	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN	resnet101		88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
RoITransformer	resnet101		88.53	77.91	37.63	74.08	66.53	62.97	66.57	90.50	79.46	76.75	59.04	56.73	62.54	61.29	55.56	67.74
R-DFPN	resnet101	✓	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
[45]	resnet101	✓	81.25	71.41	36.53	67.44	61.16	50.91	56.60	90.67	68.09	72.39	55.06	55.60	62.44	53.35	51.47	62.29
ICN	dresnet101	✓	81.40	74.30	47.70	70.30	64.90	67.80	70.00	90.80	79.10	78.20	53.60	62.90	67.00	64.20	50.20	68.20
Baseline	resnet101		81.10	**76.37**	39.18	71.50	56.65	49.91	61.04	90.44	78.34	76.44	**62.94**	**58.53**	62.02	51.22	45.24	64.06
Baseline+GSD	resnet101		81.17	75.39	39.70	73.04	60.25	51.60	64.49	90.60	**79.27**	77.74	62.80	55.33	**63.30**	**60.39**	51.50	65.77
Baseline+Scene	resnet101		**88.35**	76.07	39.01	67.17	57.92	49.89	63.32	90.47	78.05	76.75	61.97	57.89	62.39	57.45	**52.17**	65.26
GSDet	resnet101		83.21	75.41	**40.03**	**76.36**	**62.17**	**54.42**	**65.51**	**90.70**	78.59	**79.73**	59.05	56.07	63.14	60.15	50.83	**66.36**
Baseline+Res	resnet101		79.20	73.18	37.37	73.95	61.14	57.07	71.71	88.65	77.23	76.38	62.09	63.19	64.32	57.17	47.80	66.03
Baseline+GSD+Res	resnet101		80.99	74.67	38.28	75.21	64.15	**60.14**	74.12	89.54	78.41	78.51	64.96	60.61	65.39	**58.75**	48.96	67.51
Baseline+Scene+Res	resnet101		**82.08**	73.91	38.71	73.60	62.82	59.81	72.66	88.68	77.83	77.42	63.12	62.91	65.22	56.75	47.98	66.58
GSDet+Res	resnet101		81.12	**76.78**	**40.78**	**75.89**	**64.50**	58.37	**74.21**	**89.92**	**79.40**	**78.83**	**64.54**	**63.67**	**66.04**	58.01	**52.13**	**68.28**

IV. EXPERIMENTS

A. Dataset

Considering that our method requires GSD labeling, we choose DOTA [19] as our experimental dataset. It contains 2,806 aerial images from different sensors and platforms. The image sizes range from around 800×800 to $4,000 \times 4,000$ pixels, while the images contain objects exhibiting a wide variety of scales, orientations, and shapes. There are 15 categories, including *Baseball diamond (BD), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Tennis court (TC), Baseball diamond (BD), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Swimming pool (SP) and Helicopter (HC)*. The fully annotated DOTA benchmark contains 188,282 instances, each of which is labeled by an arbitrary quadrilateral.

We use both the training and validation sets for training and leave the testing set for testing. Apart from horizontal image-flipping, random contrast, random saturation and random brightness, no other data augmentation strategy is utilized in this study. We crop a series of 800×800 patches from the original images with stride 400, resulting in 75,951 patches in total. For images without GSD labels, we ignore the calculation of GSD subnet loss $\mathcal{L}_{GSD}$. In the testing stage, we resize the image into two scales (1.0 and 0.5). Considering the influence of images scale on the scene subnet, we adopt the same patch size as that in the training stage, *i.e.* cropping 800×800 patches with stride 400.

B. Implementation Details

We build our method based on the oriented-head detector in FR-O [19] as well as ResNet101 as backbone. For comparison purpose, we term the model directly composed by FR-O and ResNet101 as baseline. With the pre-trained basic detection network module on ImageNet dataset, the parameters of base ResBlock within this network module are fixed. RPN is applied to propose regions with a horizontal bounding box under the supervision of fake horizontal ground truth. The fake ground truth is generated by enclosing a minimum horizontal rectangle on the oriented ground truth. It samples 256 proposals in a minibatch after NMS, each of which is positive if IoU>0.7 with the ground-truth regions, and negative if IoU<0.3. After gathering the sizes and RoI alignment, the features of the proposals are averagely pooled, fed into a fully-connected layer, and output 1024-dimension feature. The resultant features are then fed into two independent fully-connected layers with 1024-dimension output for the classification and regression tasks separately. Unlike horizontal location regression, FR-O needs to regress five values (x, y, w, h, θ) for the representation of the oriented bounding box. To verify the flexibility of the proposed method on different head, we also establish experiments by aligning the proposal features into 14×14 size and replacing the above fully-connected operation with ResBlock, which has stronger representation ability (termed as +Res). We then apply global pooling to obtain the 2048-dimension intermediate features $\mathbf{f}$.

Different with the baseline, GSDet generates 256-dimensional (*s.t.* $d' = 256$) enhanced features $\mathbf{f}'$ via the GSD subnet and the scene subnet first, and then we concatenate them with $\mathbf{f}$. In the GSD subnet, we divide GSD in logarithmic space into $K = 6$ intervals. We calculate E via the method mentioned in Subsection III.C, and obtain 2 as E for the DOTA dataset. Moreover, N is $1,025$, which is composed of $1,024$-dimension weight parameters and 1-dimension bias of the final classification layer. In our experiments, we set the smooth factor ε as 0.1 and the weighting factor λ as 1.0. We utilize SGD with momentum 0.9 to optimize the entire network. Note that we pretrain the GSD subnet alone for 1 epoch with a learning rate of 0.0003, which can provide more stable GSD information in the early training stage and accelerate the convergence process. Then the GSD subnet is trained jointly with the other networks for 32 epochs in total with an initial learning rate of 0.0003 that is decreased twice ($\times 0.01$) after 15 and 28 epochs.

C. Results

We compare our method with FR-O [19], R2CNN [32], RRPN [33] RoITransformer without FPN [45], R-DFPN [5] method in [44] and ICN [2].

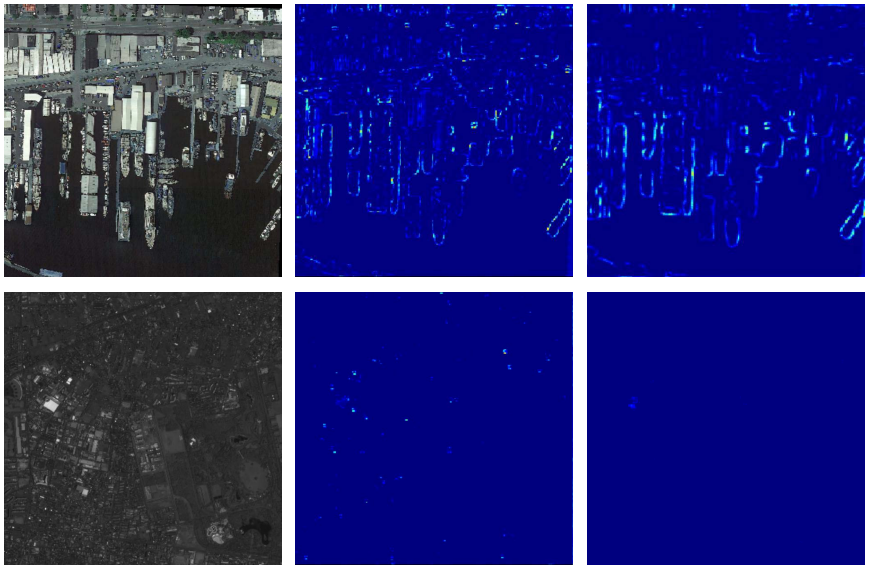


Fig. 6. Heat-maps visualizations of the intermediate feature maps from GSD subnet. Images in middle column correspond to the result from the first branch with dilated rate 1, and images in the right column correspond to that from the third branch with dilated rate 3. The third branch in top image with smaller GSD can keep large percentage of response compared with the first branch, while the third branch lost most percentage of response in bottom image with larger GSD. This result is consistent with the analysis above.

As shown in Table I, compared with the baseline, after incorporating the GSD and scene subnets, our method GSDet achieves 66.36% mAP, which is 2.3% higher than baseline. In addition, GSDet clearly improves the object detection results on several categories, *e.g.*, 5.52% for Small vehicle, 8.93% for Swimming pool and 5.59% for Helicopter. This shows that the introducing of GSD prior as well as scene prior can effectively improve the performance of object detection.

With ResBlock, which has stronger representation ability, GSDet+Res can achieve 68.28% mAP, which is 2.2% higher than baseline+Res and obtains state-of-the-art performance even when compared with the method with FPN. It also produces significant performance gains in several categories, *e.g.*, 3.6% for Baseball diamond, 3.36% for Small vehicle and 4.33% for Helicopter. The visualization of the detection results is also shown in Figure 5.

By summarizing the comparision results from GSDet and GSDet+Res, the effectiveness of our method stacking on different two-stage detection frameworks can be seen.

D. Interpretability

The heat maps on different branches in the GSD subnet is shown in Figure 6. It is obvious that a branch with high dilated rate will lose a larger percentage of response on images with higher GSD. Correspondingly, a branch with a high dilated rate can keep most response on images with smaller GSD. This result verifies the rationality we model the multi-branch GSD identification network with different receptive field for each branch.

Considering that the class-wise reweighting stage can learn a mapping from the distribution of physical size to that of category-wise probability for a given proposal, we visualize the learned transform matrix $\mathbf{W}_g \in \mathbb{R}^{L\times C}$ which contains the relation between the physical size and the prior classification probabilities, shown as Figure 7. The magnitude in the weight matrix reflects the degree of correlation between the variables on the two corresponding axes. The higher the brightness, the greater the value. The visualization shows that a category usually has larger and contiguous correlations on adjacent values of vertical axis related with physical size, which is consistent with the fact that the physical sizes of a category vary within a certain range. In addition, human common-sense have been learned explicitly to some extent. For example, ships usually have small physical sizes, and the matrix value shows that Ship mainly correlate with small scales in vertical axis. Furthermore, it is noticeable that the proposed method is prone to obtain better detection result when the number of highly related scales is small than that is large. For example,

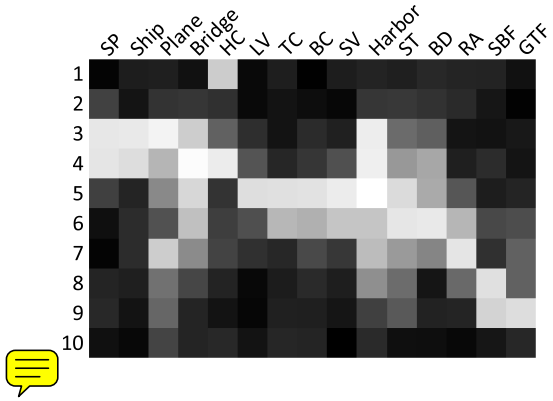


Fig. 7. Visualizations of transform matrix $\mathbf{W}_g$. The higher the brightness, the greater the value. Responses of a category to $\mathbf{t}$ (obtained from GSD) is contiguous on adjacent values of vertical axis obviously. The number 1 to 10 for the vertical axis represent different physical scales related with GSD, in which larger number is utilized for larger scale.

TABLE II
COMPARISONS OF DIFFERENT GSD SUBNET ARCHITECTURES

Opeation	branch	share	dilated rates	Acc@train	Acc@valid
Conv	1		1	92.9	85.5
Conv	3		1-1-1	99.2	89.3
Dilated Conv	3		2-2-2	97.5	89.5
Dilated Conv	3		3-3-3	97.2	89.0
Dilated Conv	3		1-2-3	99.1	92.8
Dilated Conv	3	✓	1-2-3	98.7	**95.9**

there is only two highly related scale for the category of Small vehicle, while the highly related scale for the category of Harbor is 7. In experiment, the detection result for Small vehicle is better than that for Harbor. This result is make sense since the small number of highly related scales means we obtain more deterministic estimation for scale, which provides more convincing prior knowledge for reasoning. The above experiments reveal that the proposed method can effectively incorporate the physical size knowledge, which ensures the accuracy of the obtained GSD prior information for object detection.

E. Ablation Studies

In this subsection, we demonstrate the superiority of the designed GSD subnet architecture with its five variants, which are listed in the Table. II. The first GSD subnet architecture adopts a single branch structure. The second, third and fourth architecture are three-branch networks, which have same receptive field for each branch. The fifth architecture is a three-branch network but has different receptive field for each branch. The sixth architecture is the proposed architecture. It is still a three-branch network with different receptive field for each branch, in which the parameters are shared for each branch. We select 60,000 patches as a new training dataset and 10,000 patches as a new validation dataset from the original training patches. Results for GSD classification are shown in Table II. Compared with other architectures, the proposed method can achieve the best performance with a small number of parameters, which demonstrates the suitability of the proposed architecture for the GSD identification task. It is noticeable that architectures with same receptive field inferior to that with different receptive field on the test data. This further confirms that the extraction of GSD information need to be combined with the information of different receptive fields.

To verify the effectiveness of the proposed two subnets, we conduct ablative experiments by removing the two branches respectively. Specifically, we fix α to $1/C$ to avoid the influence of the GSD subnet, and term the variant of the proposed GSDet without using GSD subnet as Baseline+Scene and Baseline+Scene+Res. In the same way, we fix β to $1/C$ to avoid the influence of the scene subnet, and term the variant of the proposed GSDet without using scene subnet as Baseline+GSD and Baseline+GSD+Res. The result is shown in Table I. It can be seen that mAP degrades when each branch is removed, which demonstrates that the combination of the two subnet contributes to the performance improvement. In addition, though both branches improve the results, the GSD subnet focuses on categories with less physical size variation (*e.g.*, Ship, Swimming pool) while the scene subnet focuses on categories with scene constraints (*e.g.*, Plane). This result is consistent with the prior information we expected, *i.e.*, GSD prior knowledge and scene prior knowledge realize the reasoning process in a complementary way.

V. CONCLUSION

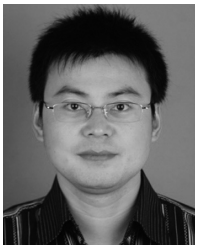
We present GSDet, which incorporates object-scale reasoning into the object detection modeling process. More specifically, we propose a novel GSD identification network to provide GSD distribution information, then combine it with the pixel range of proposals to take advantage of scale knowledge among categories. Scene information is also considered to promote the reasoning performance. Experiments reveal the distinct interpretability and demonstrate the stable performance improvement of our framework.

REFERENCES

[1] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Represent.*, vol. 34, pp. 187–203, Jan. 2016.
[2] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," in *Proc. Asian Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 150–165.
[3] Y. Long, Y. Gong, Z. Xiao, and Q. Liu, "Accurate object localization in remote sensing images based on convolutional neural networks," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 5, pp. 2486–2498, May 2017.
[4] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017.
[5] X. Yang *et al.*, "Automatic ship detection in remote sensing images from Google earth of complex scenes based on multiscale rotation dense feature pyramid networks," *Remote Sens.*, vol. 10, no. 1, p. 132, Jan. 2018.
[6] W. Ouyang *et al.*, "DeepID-Net: Object detection with deformable part based convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 7, pp. 1320–1334, Jul. 2017.
[7] Z. Xiao, Y. Gong, Y. Long, D. Li, X. Wang, and H. Liu, "Airport detection based on a multiscale fusion feature for optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 9, pp. 1469–1473, Sep. 2017.
[8] X. Li, M. Kan, S. Shan, and X. Chen, "Weakly supervised object detection with segmentation collaboration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9735–9744.
[9] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[10] J. Wang, J. Yao, Y. Zhang, and R. Zhang, "Collaborative learning for weakly supervised object detection," 2018, *arXiv:1802.03531*. [Online]. Available: http://arxiv.org/abs/1802.03531
[11] D. Zhang, J. Han, L. Zhao, and D. Meng, "Leveraging prior-knowledge for weakly supervised object detection under a collaborative self-paced curriculum learning framework," *Int. J. Comput. Vis.*, vol. 127, no. 4, pp. 363–380, Apr. 2019.
[12] D. Zhang, J. Han, G. Guo, and L. Zhao, "Learning object detectors with semi-annotated weak labels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3622–3635, Dec. 2019.
[13] X. Chen, L.-J. Li, L. Fei-Fei, and A. Gupta, "Iterative visual reasoning beyond convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7239–7248.
[14] X. Chen and A. Gupta, "Spatial memory for context reasoning in object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4086–4096.

[15] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3076–3086.
[16] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4145–4153.
[17] F. Yang, H. Fan, P. Chu, E. Blasch, and H. Ling, "Clustered object detection in aerial images," 2019, *arXiv:1904.08008*. [Online]. Available: http://arxiv.org/abs/1904.08008
[18] I. Misra, A. Gupta, and M. Hebert, "From red wine to red tomato: Composition with context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1792–1801.
[19] G.-S. Xia *et al.*, "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[20] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[21] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.
[24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[25] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[26] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[27] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7263–7271.
[28] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767
[29] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[30] X. Zhou, J. Zhuo, and P. Krahenbuhl, "Bottom-up object detection by grouping extreme and center points," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 850–859.
[31] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9627–9636.
[32] Y. Jiang *et al.*, "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*. [Online]. Available: http://arxiv.org/abs/1706.09579
[33] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.
[34] H. Xu, C. Jiang, X. Liang, L. Lin, and Z. Li, "Reasoning-RCNN: Unifying adaptive global reasoning into large-scale object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6419–6428.
[35] K.-H. Lee, X. He, L. Zhang, and L. Yang, "CleanNet: Transfer learning for scalable image classifier training with label noise," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5447–5456.
[36] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2015, *arXiv:1511.07122*. [Online]. Available: http://arxiv.org/abs/1511.07122
[37] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[38] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*. [Online]. Available: http://arxiv.org/abs/1706.05587
[39] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[40] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "DetNet: Design backbone for object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 334–350.
[41] J. Dai *et al.*, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[42] J. Peng, M. Sun, Z.-X. Zhang, T. Tan, and J. Yan, "POD: Practical object detection with scale-sensitive network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9607–9616.
[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[44] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50839–50849, 2018.
[45] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2849–2858.

Wei Li received the M.S. degree in computer science from Northwestern Polytechnical University, Xian, China, in 2021. His research interests include object detection and instance segmentation.



Wei Wei (Senior Member, IEEE) received the Ph.D. degree from Northwestern Polytechnical University, Xian, China, in 2012. He is currently an Associate Professor with the School of Computer Science, Northwestern Polytechnical University. He has been authored more than 40 articles, including IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, *Patten Recognition*, CVPR, ICCV, ECCV, AAAI, and IJCAI. His research interests include image processing, machine learning, and pattern recognition. He is a reviewer of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE REMOTE SENSING LETTERS, and IEEE TRANSACTIONS ON NEURAL NETWORK AND LEARNING SYSTEM. He has served as the PC Member for around ten major international conferences, including CVPR and ICME.

Lei Zhang (Member, IEEE) received the Ph.D. degree in computer science and technology from Northwestern Polytechnical University, Xian, in 2018. He was a Research Staff with the School of Computer Science, The University of Adelaide, Australia, from 2017 to 2019. He was a Research Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates, from 2019 to 2020. He is currently a Professor with the School of Computer Science, Northwestern Polytechnical University. His research interests include image processing, machine learning, and video analysis.