Precise Object Detection Using Adversarially Augmented Local/Global Feature Fusion

Xiaobing Han[a], Tiantian He[a], Yew-Soon Ong[a] and Yanfei Zhong[b]

[a]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[b]LIESMARS, Wuhan University, Wuhan, China

ARTICLE INFO

Keywords:
High spatial resolution (HSR) remote sensing imagery
geospatial object detection
super resolution generative adversarial network
data augmentation
local/global feature fusion

ABSTRACT

Object detection, which aims at recognizing or locating the objects of interest in remote sensing imagery with high spatial resolutions (HSR), plays a significant role in many real-world scenarios, e.g., environment monitoring, urban planning, civil infrastructure construction, disaster rescuing, and geographic image retrieval. As a long-lasting challenging problem in both machine learning and geoinformatics communities, many approaches have been proposed to tackle it. However, previous methods always overlook the abundant information embedded in the HSR remote sensing images. The effectiveness of these methods, e.g., accuracy of detection, is therefore limited to some extent. To overcome the mentioned challenge, in this paper, we propose a novel two-phase deep framework, dubbed GLGOD-Net, to effectively detect meaningful objects in HSR images. GLGOD-Net firstly attempts to learn the enhanced deep representations from super-resolution image data. Fully utilizing the augmented image representations, GLGOD-Net then learns the fused representations into which both local and global latent features are implanted. Such fused representations learned by GLGOD-Net can be used to precisely detect different objects in remote sensing images. The proposed framework has been extensively tested on a real-world HSR image dataset for object detection and has been compared with several strong baselines. The remarkable experimental results validate the effectiveness of GLGOD-Net. The success of GLGOD-Net not only advances the cutting-edge of image data analytics, but also promotes the corresponding applicability of deep learning in remote sensing imagery.

1. Introduction

Automatically determining whether a given image comprises objects of interest and locating the coordinates of the predicted objects in an image, object detection is a significant civil related interpretation manner [4]. Due to the advances of aerospace and remote sensing, there are an increasing number of high spatial resolution (HSR) images captured by remote sensing satellites. Covering a wide range of resolutions, HSR images allow one to measure the terrestrial surface as precise as to sub-meter level and they are regarded as a significant data source for precise land use and land cover (LULC) investigation [1, 2, 3]. As highly geometrical structures and spatial patterns hidden in the HSR image are of great importance to many real-world applications [41], such as environmental monitoring, and emergency rescue, geospatial object detection has drawn much attention in the recent [2].

Object detection in HSR remote sensing images has been a challenging problem. Generally speaking, there are two kinds of objects, which are worthy detecting in HSR remote sensing images, i.e., the natural objects which are part of the background, and the artificial objects which are independent of the background (e.g., ships, buildings) [4]. To identify these objects, there have been several approaches proposed, which belong to one of the following four categories, including template matching-based, knowledge-based, object-based image analysis (OBIA)-based, and machine learning-based [4]. Though different categories of approaches exist, most of them to object detection in HSR images are machine-learning based. And those conventional learning-based methods always attempt to formulate the detection task as a four-step learning problem, including proposal generation, feature extraction, classification, and localization. The proposal generation stage mainly adopts the selective search (SS) algorithm [40], the feature extraction stage uses the handcrafted feature extractors (e.g., the histogram of oriented gradients (HOG) [39] and scale-invariant feature transform (SIFT) [21]), and the classification stage usually utilizes the shallow classifiers, such as support vector machine (SVM) [5, 15, 42], AdaBoost [39], conditional random fields [49], and sparse coding based classifiers [14, 45, 20, 19]. Although, methods for object detection which are based on conventional machine learning techniques can obtain relatively satisfactory performances, they are inefficient due to the high computational and parameter-tuning demand.

Besides conventional machine learning models, more modern learning paradigms are considered to be adopted to detect objects in HSR images. Amongst them, advanced deep learning (DL) frameworks [23, 25, 26, 24] are more frequently used as they may perform the task of object detection in HSR images in an automatic and unified manner. Different from traditional step-by-step approaches, deep learning frameworks, especially those R-CNN based, are capable of automatically extracting features hidden in the HSR image, so that more efficient and effective approaches to object detection are expected to be developed. Based on this capability, a series of DL based frameworks, e.g., anchor-free, one-stage, and two-stage object detection methods have been proposed [48, 53]. And two-stage methods for object detec-

*Corresponding author: Xiaobing Han; Tiantian He
ORCID(s): 0000-0002-2191-2864 (X. Han); 0000-0003-4839-681X (T. He)

tion become prevalent as they can learn meaningful latent features for different detection tasks via defining different pooling layers, e.g., region of interest (RoI), and position-sensitive region of interest (PSRoI) in deep learning frameworks. Though the two mentioned methods are effective to some extent, solely using either one of them may not detect those objects with low confidence values, leading to the degradation of detection rate.

Though previous methods, especially those two-stage ones, are sometimes capable of detecting objects effectively, they neither consider enlarging the detecting scale of detailed information in HSR images to improve their efficacy, nor attempt to suppress the adverse impact caused by inappropriate viewpoint, insufficient illumination, cluttered background, large shadow of the objects, and inadequate bounding boxes with labels. How to amplify the sample space, and capture more detailed information in HSR images as augmented training data, has been a open challenging problem for object detection in HSR images. To address the mentioned challenge, in this paper, we propose generative adversarial network augmentation and local/global feature fusion framework (GLGOD-Net), to precisely detect objects of interest in HSR remote sensing imagery. Different from previous approaches to object detection in HSR images, GLGOD-Net is a two-phase method, which is capable of learning pivotal local/global features from augmented HSR image data, and then utilizes them to detect object precisely. Adopting a super-resolution generative adversarial network (SRGAN) [27], GLGOD-Net is able to generate augmented HSR images followed by the multi-scale sampling strategy, each of which may capture the enhanced large-scale information hidden in the original image data. Given the augmented HSR image data, GLGOD-Net then attempts to learn the fusions of the local/global features, taking the advantage of R-FCN and Faster R-CNN, which enable both PSRoI and RoI pooling. And these feature fusions can be used to precisely detect objects in HSR remote sensing imagery, especially those hardly detected by previous approaches. To better illustrate the ideas supporting GLGOD-Net to perform the task, a schematic view of GLGOD-Net is provided in Fig. 1.

Specifically, the major contributions of this paper are summarized as follows.

- We propose GLGOD-Net, which is a novel deep framework for precisely detecting objects of interest in HSR remote sensing imagery. Different from previous approaches, GLGOD-Net is able to learn deep fusions of pivotal local/global features from augmented HSR images, which well capture the precise detailed information at a larger scale. These learned feature fusions may well preserve the information describing different objects, so that appropriately using these feature fusions can lead to a better performance of object detection.

- A novel strategy for the augmentation of HSR imagery is proposed and used in GLGOD-Net. To increase the sample number to facilitate the object detection processing, SRGAN is adopted by the proposed framework to amplify the HSR remote sensing images. Those images generated by SRGAN may well retain more detail structure information at a larger image scale, combining them with original HSR images may improve the possibility that one learns meaningful latent features for a higher detection rate.

- GLGOD-Net has been extensively tested with real-world data of HSR remote sensing images, and has been compared with both classical and advanced approaches to object detection. The experimental results show that GLGOD-Net outperforms all the compared baselines, which validates the effectiveness of the proposed framework.

The rest of this paper is organized as follows. The previous works which are related to the proposed framework are investigated in Section 2. The details of GLGOD-Net are elaborated in Section 3. The extensive experiments which are used to test the effectiveness of GLGOD-Net and other baselines, and corresponding analysis are presented in Section 4. Finally, we conclude the paper and discuss the future works.

2. Related works

Geospatial object detection is a fundamental problem in remote sensing community. Widely used approaches to object detection can be categorized into two classes, i.e., traditional methods for object detection and deep learning-based approaches to object detection. Most traditional methods for object detection are based on various machine learning techniques, which can formulate object detection as a multi-stage classification problem, consisting of proposal generation, feature extraction, classification, and localization stages.

2.1. Convolutional neural network

Besides the conventional approaches, modern approaches to object detection, especially those based on convolutional neural networks (CNN), have been experiencing a rapid development in recent years. And a number of CNN-based approaches have been proposed to detect different objects in HSR images. Having achieved a great success, CNN lays the foundation of DL series models for image processing. In contrast to traditional approaches, which always require a handcrafted feature design and extraction, CNN is a fully automatic framework enabling the learning of hierarchical features and representations. Although several improved versions of CNN frameworks, e.g., ZF [46], VGG16 [36], ResNet-50 [18], and ResNet-101 [18], have been proposed, like classical CNNs, they are always composed of several convolutional layers, pooling layers, nonlinear layers, and finally, the softmax classification layer. These

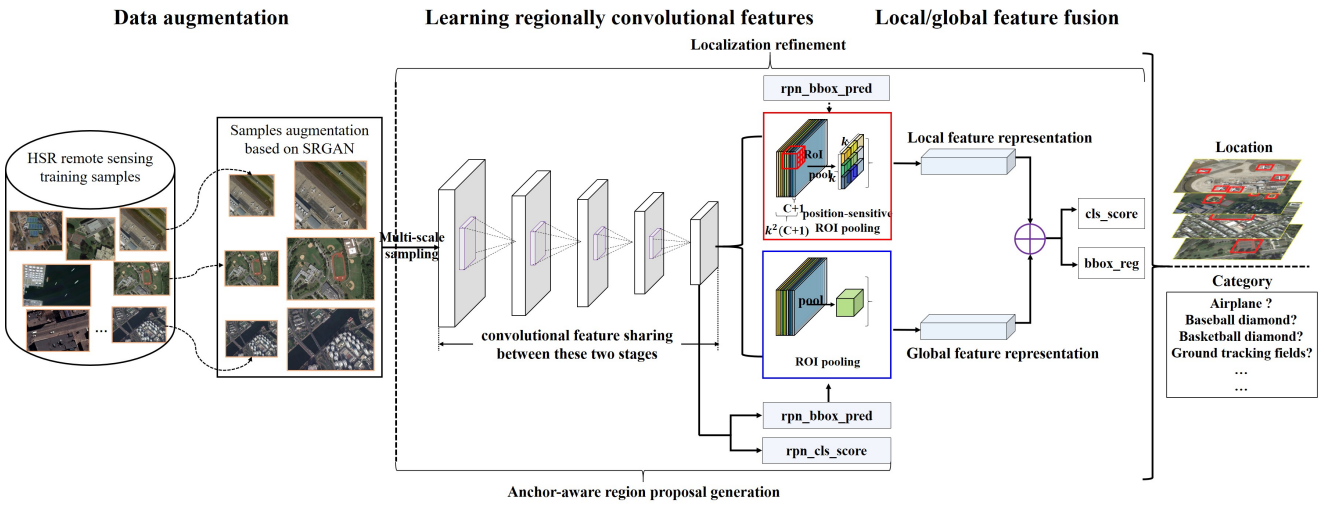


Figure 1: The schematic view of the proposed GLGOD-Net

layer-wise operations can be summarized as the follows.

$$
\begin{aligned}
&\text{Convolutional layer}\\
&z^s = \sum_i^q W_i^s * x^i + b_s\\
&\text{Nonlinear layer}\\
&a^s = f(z^s)\\
&\text{Pooling layer}\\
&P_G^s = \max_{i \in G} a_i^s\\
&\text{Softmax classification layer}\\
&J(W,b) = -\sum_i^{N_c} y_i \log(\hat{y}_i) + \lambda \sum_l^L \|W^{(l)}\|^2.
\end{aligned}
\tag{1}
$$

where x^i, W, and q denote the 3-D input array, convolutional filter bank, and the number of the kernels in each convolutional channel, respectively. $f(\cdot)$ can be any effective link function, e.g., sigmoid function, and rectified linear unit (ReLU). In the pooling operator, s and p represent the space between a grid of pooling units, and the spatial region size of pooling unit, respectively. In classification layer, K is the number of outputs, L is the total number of layers, $\hat{y}_i$ is the activations of the previous layer, λ is a regularization term, and l indexes the layer number, respectively.

2.2. CNN based object detection

There have been a number of CNN-based approaches to object detection. They can be categorized into three classes, including anchor-free object detection, one-stage object detection, and two-stage object detection methods. The anchor-free object detection methods, e.g., Foveabox [22], FCOS [38], and FSAF [51] adopt the specific network structure and prior information but ignore anchors or proposals which may directly output the location coordinates and the category of the objects. One-stage approaches to object detection, such as SSD [30], YOLO [31], YOLO9000 [32], YOLO-v3 [33], and DSOD [35], consider the detection task as fitting the regression probability of class-dependent bounding boxes. Compared with those two classes of approaches, more CNN-based approaches to object detection are two-stage ones. The two-stage object detection methods, utilize the region proposal generation stage to further locate the coordinates of the objects and classify the categories, containing R-CNN [11], SPPnet [17], Fast R-CNN [10], Faster R-CNN [34], R-FCN [8], Couplenet [52], Feature Pyramid based object detection method [29], etc.

Given the superb properties of CNN, a number of methods for object detection in HSR remote sensing imagery have been proposed. For example, a coupled weakly supervised CNN [47] is proposed to detect aircraft in HSR images, which enables the feature sharing between CRPNet and LOCNet so as to improve the detection performance. HSF-Net [28] proposed a hierarchical selective filtering layer to map features in different scales to the same scale space to improve the ship object detection performance. In [14], a deep framework for object detection is proposed. This framework combines deep Boltzmann machine and high-level feature learning. In [9], a multi-scale framework is proposed to detect objects in HSR remote sensing image. Different from those mentioned frameworks, in this paper, we propose GLGOD-Net, which considers the data augmentation strategy preserving the detail information, and the local/global pooling feature fusion when performing the detection tasks.

3. The proposed GLGOD-Net

In this section, we elaborate the proposed GLGOD-Net, which is a novel framework for object detection taking into the consideration the data augmentation and the fusion of local/global features. We first introduce how GLGOD-Net learns detail information from original image data, so as to

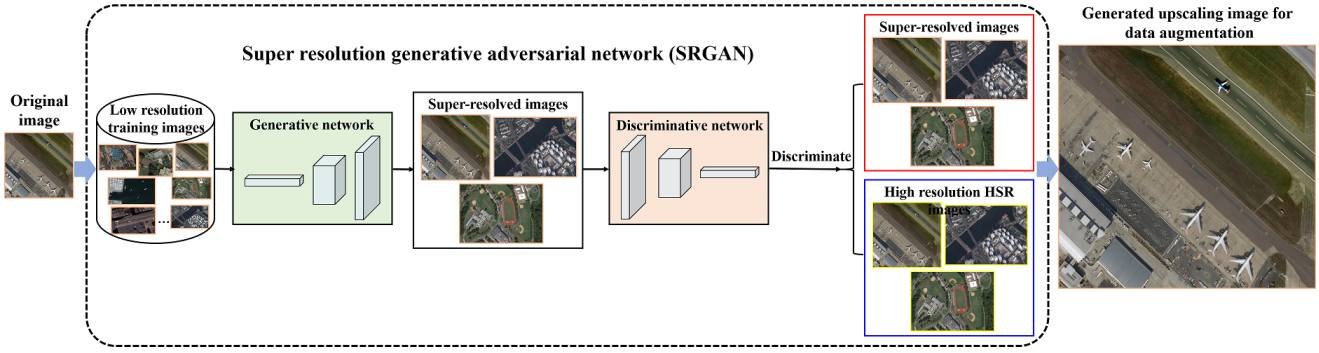


Figure 2: Super-resolution adversarial data augmentation.

adversarially generate images for the enhancement of network training. We then present how GLGOD-Net performs the task of precise object detection, making use of the fusions of local/global localization features, which are learned from enhanced HSR image data.

3.1. Super-resolution adversarial data augmentation

How to recover the intimate information from low-resolution HSR images and preserve it for future analytical tasks, e.g., latent feature learning for object detection, has been a challenging problem in remote sensing community. Such high-quality intimate information may determine whether a detection framework can effectively identify various objects in HSR images. In the proposed framework, an effective adversarial strategy, dubbed Super-resolution GAN (SRGAN) is adopted to learn those intimate features from low-resolution images, so that the data can be augmented via generating new images according to the learned features. Constituted by a generator network and a discriminator network, SRGAN is able to estimate high-resolution images given low-resolution ones. In SRGAN, an L-layer feed-forward CNN with parameters $\theta_G = \{W_{1:L}; b_{1:L}\}$ is adopted to generate a high-resolution, super-resolved image i^{SR} from a low-resolution input image i^{LR}, where i^{LR} is obtained by using a Gaussian filter to i^{HR} with down-sampling factor r. The loss function of the generator network is defined in Eq. (2).

$$\hat{\theta}_G = \arg\min_{\theta_G} = \frac{1}{N}\sum_{1}^{N} l^{SR}(G_{\theta_G}, (I_n^{LR}), I_n^{HR}), \quad (2)$$

where, the number of the training images is defined as $n = 1, ..., N$.

The discriminator network in SRGAN is designed to maximize the discrimination between real HR images and generated SR samples. It has eight convolutional layers with an increasing number of 3×3 convolutional filter kernels by an incremental factor of 2 from 64 to 512 kernels followed by two dense layers and a sigmoid activation function. The adversarial min-max loss function of the discriminator network and the generator network is defined in Eq. (3).

$$\begin{aligned}\min_{\theta_G}\max_{\theta_D} & E_{I^{HR}\sim P_{train}(I^{HR})}[\log D_{\theta_D}(I^{HR})] \\ & + E_{I^{LR}\sim P_G(I^{LR})}[\log(1 - D_{\theta_D}G_{\theta_G}(I^{LR}))].\end{aligned} \quad (3)$$

The core idea of GAN [12] is that it allows one to train a generative model G and fool the differentiable discriminator D to distinguish super-resolved images from the real images. Hence, the GAN encourages to create images that are highly similar to real images and thus difficult to classify by D.

In SRGAN, the perceptual loss function l^{SR} is designed as the weighted sum of a content loss l_X^{SR} and an adversarial loss component (Eq. (4)).

$$l^{SR} = l_X^{SR} + 10^{-3} l_{gen}^{SR}, \quad (4)$$

where l_X^{SR} is the content loss and l_{gen}^{SR} is the adversarial loss.

For the content loss, the commonly utilized metric is the pixel-wise MSE loss shown in Eq. (5).

$$l_{MSE}^{SR} = \frac{1}{r^2 WH}\sum_{x=1}^{rW}\sum_{y=1}^{rH}(I_{x,y}^{HR} - G_{\theta_G}(I^{LR})_{x,y})^2 \quad (5)$$

where W and H represent the width and height of the low-resolution image, r is the upscaling factor.

The Euclidean distance between the feature representations of a reconstructed image $G_{\theta_G}(I^{LR})$ and the reference image I^{HR} is shown in Eq. (6).

$$l_{i,j}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\phi_{i,j}(H^{HR})_{x,y} - \phi_{i,j}(G_{\theta_G}(I^{LR})_{x,y})^2 \quad (6)$$

In addition to the content loss, the generative component of GAN to the perceptual loss (l_{gen}^{SR}) is also defined according to the probabilities of the discriminator ($D_{\theta_D}(G_{\theta_G}(I^{LR}))$) on all the training samples (Eq. (7)):

$$l_{gen}^{SR} = -\sum_{n=1}^{N}\log D_{\theta_D}(G_{\theta_G}(I^{LR})) \quad (7)$$

For HSR remote sensing images, in this paper, the SRGAN model is trained with the original scale HSR images, then the upscaling images of the HSR images can be generated using SRGAN which recovers the detailed information of the HSR remote sensing images. To ensure SRGAN to be convergent, the improved training strategy used by Wasserstein GAN [13] is adopted. How SRGAN super-resolves the HSR remote sensing images is shown in Fig. 2.

3.2. Learning regionally convolutional features

Given the enhanced HSR image data, GLGOD-Net is able to detect objects via fusing the convolutional spatial features learned from the alternative regions in HSR images.

3.2.1. Anchor-aware region proposal generation

Region proposal generation (RPN), aims at generating a number of rectangular object proposals in arbitrary-size images. Different from previous anchor-free approaches which always generate regions via a fully convolutional network (FCN), GLGOD-Net attempts to uncover anchor-based proposals after learning the HSR image data. Allowing GLGOD-net to be aware of such anchor clusters is very beneficial to precise object detection. First, having predefined shape and size, anchors lead to an easier refinement of location regression. Second, anchor clusters may contain multiple objects, which enables GLGOD-Net to simultaneously learn features of these objects. Last but not the least, the shape and size of the anchor can be determined by appropriate prior knowledge, potentially forcing GLGOD-Net to learn latent features which may better represent the objects in HSR imagery. Generically, the areas and length-to-width ratios of the anchor is set as $\{128^2, 256^2, 512^2\}$, and $\{1:1, 1:2, 2:1\}$, respectively.

Following the layer of Anchor-aware RPN, a box regression sibling fully connected layer and a box classification sibling fully connected layer are appended to capture the coordinate information of an object, and determine whether the generated object is predefined. Based on the above illustrations, the loss function of RPN is shown in Eq. (8).

$$\begin{aligned} L_{RPN} &= \sum_i L(p_i^r, \hat{p}_i^r, t_i^r, \hat{t}_i^r) \\ &= \frac{1}{N_{cls}} \sum_i L_{cls}(p_i^r, \hat{p}_i^r) + \lambda \phi_i^r L_{reg}(t_i^r, \hat{t}_i^r), \end{aligned} \tag{8}$$

where p_i^r represents the predicted probability of anchor being an object, the ground truth label $\hat{p}_i^r$ is 1 when the anchor is positive. t_i^r and $\hat{t}_i^r$ represent the 4 parameterized coordinates of the predicted anchor box and ground-truth anchor box, respectively. L_{cls} and L_{reg} represent the region proposal classification loss and regression loss, respectively.

3.2.2. Localization refinement

The localization of objects generated in the RPN stage is always coarse. GLGOD-Net adopts a Fast R-CNN to refine the proposed regions. Fast R-CNN first processes the whole image with a CNN model to produce a convolutional feature map, and then utilizes a set of object proposals to score. In this procedure of the two-stage object detection network, a fixed-length feature vector is generated for each object proposal from the feature map with a region of interest (RoI) pooling layer or a position-sensitive region of interest (PSRoI) pooling layer, which converts the feature map into a small feature map with a fixed spatial extent $H \times W$, where H and W are independent layer hyper-parameters and the size of sub-windows is $h/H \times w/W$. Afterwards, the generated feature vector is imported into a sequence of fully connected layers followed by a layer to produce softmax probabilities over K object proposals plus a "background" class and a layer to produce four refined coordinate values for each of the K classes. The loss function of the Fast R-CNN (FR-CNN) for localization is shown in Eq. (9).

$$\begin{aligned} L_{FRCNN} &= \sum_n L(p_n^f, \hat{p}_n^f, t_n^f, \hat{t}_n^f) \\ &= \frac{1}{N_{cls}} \sum_n L_{cls}(p_n^f, \hat{p}_n^f) + \lambda \phi_n^f L_{reg}(t_n^f, \hat{t}_n^f). \end{aligned} \tag{9}$$

Besides, non-maximum suppression (NMS) is used to select the maximum classification confidence value and to eliminate redundant bounding boxes in the refinement stage.

3.3. Feature fusion for object detection

In the RPN stage, two kinds of features, i.e., local and global features can be learned by GLGOD-Net. When being used to detect object in HSR images, local/global features have advantages and deficiencies individually.

Object-specific features (local features) are learned by the PSRoI pooling layer in R-FCN. Generically, PSRoI pooling layer in the RPN stage attempts to learn a set of part-sensitive score maps with a 1×1 convolutional layer appended by $k^2(C+1)$ channels. Here k and C represent the number of parts into which each object is divided, and the number of object categories, respectively. Each k^2 channels are used to encode a specific part of the object and the probability that a part belongs to a category can be determined by voting the k^2 responses generated by the part-specific channels. At last, the $(C+1)-d$ vector is obtained to indicate the class-specific probability. It is seen that features learned by the PSRoI pooling concern more about the internal structure and components of the object belonging to a particular category.

Different from those local features generated by PSRoI pooling, features learned by RoI pooling in Faster R-CNN are capable of encoding the region-level information of each object. Disregarding the size of the object, RoI pooling in the RPN stage attempts to generate a fixed length convolutional layer ($1024-d$ 1×1) to capture the global structure of each object. Then, two convolutional layers with the kernel size $k \times k$ and 1×1 are adopted to learn the global representations of the object of interest. At last, the output of 1×1 convolution is fed into the classifier whose output is also a $(C+1)-d$ vector. Though global information is well preserved by RoI pooling, it is difficult to precisely detect objects only relying on such global features, as many HSR remote sensing images contain many objects with occlusions or truncations.

Either the locally sensitive features generated by the PSRoI pooling in R-FCN, or the global context features generated by RoI pooling in Faster R-CNN, is ex-parte and thereby insufficient for precisely detecting objects in HSR images. Aiming at providing a comprehensive feature space to precisely detect objects in HSR remote sensing imagery, we propose to fuse the local and global features. To match the same order of magnitude, a normalization operation among each kind of features is needed before combined. The element-wise summation is then performed to model the scale. Compared with object detection using either local or global features, feature fusion for object detection may fully explore structural features learned by GLGOD-Net, and thereby provide more information when detecting different objects, especially for hard-to-detect ones in HSR remote sensing images.

3.4. Computational complexity

In this subsection, the computational complexity of two essential components of the proposed model, which are learning regionally convolutional features and feature fusion for object detection, is analyzed. For learning regionally convolutional features, the time complexity follows the order of $O(\sum_{l=1}^{D} h_m^l \cdot w_m^l \cdot h_k^l \cdot w_k^l \cdot C^{l-1} \cdot C^l)$, where h_m and w_m represent the size of convolutional feature map, h_k and w_k stand for the size of the convolutional kernel, C^{l-1} and C^l are the numbers of the input and output channel of the convolutional kernel, and D represents the number of ResNet-101.

The feature fusion stage is composed of the global RoI pooling and local PSRoI pooling. The time complexity of the global RoI pooling follows the order of $O(c \cdot ((h - h_p + p_h + s_h)/s_h \cdot (w - w_p + p_w + s_w)/s_w)$, where $c \times h \times w$ is the size of pooling layer, $h_p \times w_p$ represents the size of the pooling window, (p_h, p_w) and (s_h, s_w) stand for the padding and stride of pooling. The time complexity of local PSRoI pooling follows the order of $O(k \cdot k \cdot c \cdot (((h/k) - h_{p_l} + p_{h_l} + s_{h_l})/s_{h_l} \cdot ((w/k) - w_{p_l} + p_{w_l} + s_{w_l})/s_{w_l})$, where $c \times (h/k) \times (w/k)$ is the size of $k \times k$ local regions on the input convolutional feature map, $(h_{p_l} \times w_{p_l})$ represents the size of the pooling window, (p_{h_l}, p_{w_l}) and (s_{h_l}, s_{w_l}) stand for the padding and stride of pooling.

As the computational complexity of learning regionally convolutional features is much higher, the overall complexity of the proposed model follows the order of $O(\sum_{l=1}^{D} h_m^l \cdot w_m^l \cdot h_k^l \cdot w_k^l \cdot C^{l-1} \cdot C^l)$.

4. Experiments and analysis

In this section, a series of experiments are conducted to validate the effectiveness of GLGOD-Net against both traditional and state-of-the-art deep learning based object detection methods.

4.1. Experimental setup

4.1.1. Dataset description

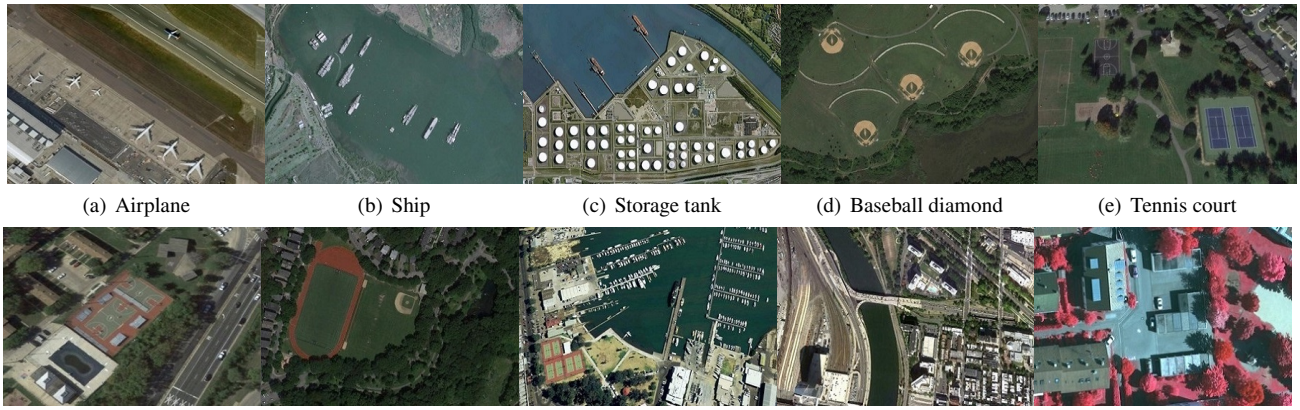
In our experiments, we used NWPU VHR-10 object detection dataset to test the effectiveness of different approaches. As a multi-source and multi-resolution object detection dataset, NWPU VHR-10 contains 800 images in total, 715 images of which are collected from Google Earth with 0.5-2.0m spatial resolutions, and the rest of which are pan-sharped color infrared images collected from Vaihingen dataset with 0.08m spatial resolution. What makes the detection task in this dataset challenging is that not all images are labeled with one or more objects. 150 images out of 800 are without any labeled object. In our experiment, the split ratios of the positive dataset used are 20% for the training dataset (130 images), 20% for the validation dataset (130 images), and 60% for the test dataset (390 images). Fig. 3 demonstrates the category examples of the dataset. In the experiment, for the proposed GLGOD-Net framework, the initial learning rate (0.001) with a "step" strategy of gamma (0.1) of the ResNet-101-ohem is set. The momentum (0.9), weight decay (0.0005), and the total iteration number (10000) are set. The experiment has been run for five times to compute the average performance of all the approaches.

4.1.2. Baselines for comparison

To validate the effectiveness of GLGOD-Net, we selected nine approaches to object detection, which are either conventional-based, or deep learning-based, as baselines, including BoW [42], spatial sparse coding BoW (SSCBoW) [37], Fisher discrimination dictionary learning (FDDL) [5], collection of part detectors (COPD) [6], transferred CNN (TCNN) [7], newly trained CNN (NTCNN) [7], the RICNN without fine-tuning (RICNN-w) [7], the RICNN with fine-tuning (RICNN-f) [7], the R-P-Faster R-CNN (RPFR-CNN) [16], and PSB [50].

BoW, SSCBoW, FDDL, and COPD are four representative methods for object detection which are based on classical machine learning techniques. BoW describes the statistics of the occurrence of visual words with the combination of spectral and texture features on the basis of patch detection and description to divide and represent various subregions of objects comprising multiple homogeneous components. SSCBoW uses a new spatial mapping strategy to encode the geometric information representing the relative position of the parts of a target with rotation variations, and then utilizes the sparse coding to achieve a much lower reconstruction error on the basis of BOW model. FDDL uses saliency prediction model to generate a small set of target candidate areas, and then utilizes discriminative sparse coding to learn a dictionary to have small within-class scatter and big between-class scatter to obtain the final object categorization performance. COPD is composed of a set of representative and discriminative part detectors, where each part detector is a linear support vector machine classifier used for the detection of objects or recurring spatial patterns within a certain range of orientation.

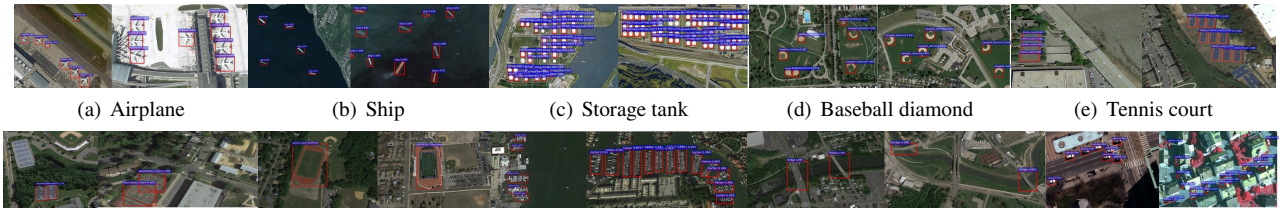
Transferred CNN, newly trained CNN, RICNN-w, RICNN-f, R-P-Faster R-CNN, and PSB are six deep learning frameworks for object detection. Transferred CNN object detection method uses the pre-trained model on a large-scale natural image dataset to train the object images to ob-



(a) Airplane (b) Ship (c) Storage tank (d) Baseball diamond (e) Tennis court

(f) Baseball court (g) Ground track field (h) Harbor (i) Bridge (j) Vehicle

Figure 3: Exemplified images from NWPU VHR-10 dataset. All the figures need redrawing to keep their height same.



(a) Airplane (b) Ship (c) Storage tank (d) Baseball diamond (e) Tennis court

(f) Baseball court (g) Ground track field (h) Harbor (i) Bridge (j) Vehicle

Figure 4: The qualitative results of the proposed GLGOD-Net with the NWPU VHR-10 dataset.

tain the category information. Newly trained CNN learns the object classifier from scratch on the object images to obtain the object category information. RICNN without fine-tuning proposes to train an object detection classifier with rotation-invariance properties from scratch. RICNN with fine-tuning trains an object detection classifier with rotation-invariance properties with fine-tuning the CNN with pre-trained network weights from a large-scale natural image dataset. R-P-Faster R-CNN method fine-tunes the Faster R-CNN from a pre-trained network weight from a large-scale natural image dataset, which can greatly promote the object detection performance. PSB deals with the object detection problem with considering both the translation-invariance in the classification stage and the translation-variance in the object stage with introducing the position-sensitive balancing framework on the basis of ResNet-101 for geospatial object detection.

4.1.3. Evaluation metrics

Four evaluation metrics, including average precision (AP), mean average precision (mAP), precision-recall curves (PRCs), and recall rate, are used in our experiments to evaluate the performance of different approaches [4, 45, 44, 43]. To compute these metrics, Precision and Recall have to be obtained according to Eq. (10), where TP, FP, and FN represent the number of true positives, false positives, and false negatives, respectively. AP (average precision) calculates the average precision on the PRCs, where $p(r)$ adopts the largest value of precision among the right side of the point in PRCs. mAP (mean average precision) measures the mean value of the average precision values w.r.t. all the n object categories.

$$\begin{aligned} &\text{Precision} \\ &P = \frac{TP}{TP + FP} \\ &\text{Recall} \\ &R = \frac{TP}{TP + FN} \\ &\text{AP} \\ &AP = \int_0^1 p(r)\mathrm{d}r \\ &\text{mAP} \\ &mAP = \frac{\sum_{l=1}^{n} AP_l}{n} \end{aligned} \tag{10}$$

These four selected metrics have been widely used to evaluate the performance of object detection in previous works [4]. AP reflects the proportion of the correctly predicted objects among all the predicted targets. Recall reflects the proportion of the correctly predicted objects that are also in the true targets, the correct detection rate, and the false detection rate. Compared with AP and Recall, mAP evaluates the performance of object detection in a complementary way, which includes precision and recall of all object categories into the evaluation. It is well known that one per-

forms well when precision and recall are high at the same time. Thus, we use PRC curve to reveal the inherent relation trend between precision and recall. PRC curve at the top-right means it covers more area and indicates a better performance of object detection. Given how these four metrics evaluate the performance of object detection, we know that AP and Recall take more emphasis on the category-wise performance, while mAP and PRC take into the consideration precision and recall of all categories, so that they can provide an entire evaluation on object detection across all categories. These metrics can complementarily measure the performance of object detection, and deeply reveal robustness of different approaches. In addition to evaluate the accuracy of different frameworks, we also consider to assess the efficiency of different approaches. Thus, we recorded the computational time costed by different methods.

4.2. Performance analysis

4.2.1. Qualitative results

In order to verify the effectiveness of the proposed framework, GLGOD-Net is tested on NWPU VHR-10 dataset and is compared with other baselines from both qualitative and quantitative perspectives. Ten object categories, including airplane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, and vehicle are considered as the qualitative benchmarks used to test the effectiveness of the proposed framework. The corresponding results are demonstrated in Fig. 4. As the figure shows, GLGOD-Net is able to successfully detect various objects in HSR remote sensing imagery, despite their sizes, orientations, and contexts are different. In addition, it is also seen that, GLGOD-Net can better detect the classes of airplane, ship, tennis court, and ground track field.

4.2.2. Quantitative results

Although qualitative results may directly reveal whether GLGOD-Net can prominently perform the task of object detection, quantitative evaluations, e.g., the comparisons between GLGOD-Net and prevalent methods can demonstrate the proposed framework is more advanced. The objects detected by GLGOD-Net have been evaluated by different metrics, including AP, mAP, PRCs, and recall rate, and have been compared with different baselines. The corresponding results are summarized in Tables 1, 2, Fig. 5, and Fig. 6, respectively.

Objects detected by GLGOD-Net and other baselines have been evaluated by AP and mAP. And the comparative results have been summarized in Table 1. As the table shows, GLGOD-Net performs robustly when detecting different categories of objects. When compared with the handcrafted feature based and deep learning based object detection methods, GLGOD-Net outperforms all the baselines when detecting 8 categories of objects out of 10, except for the cases of storage tank and baseball diamond. These remarkable results indicate GLGOD-Net is very effective when performing different tasks of object detection in HSR remote sensing images.

When it comes to the comparisons between GLGOD-Net and itself using different learning strategies, the experimental results still show GLGOD-Net performs better. As it can be seen in Table 2, GLGOD-Net adopting the learning strategies proposed in this paper, outperforms 84% of the mean AP values, obtained by other GLGOD-Nets utilizing alternative learning strategies, including V1, V1-L, V1-G, V2-L, and V2-G. Compared with either GLGOD-Net without adversarial data augmentation (e.g., V1), or that concerning local or global features (V2-L and V2-G), GLGOD-Net is able to detect objects in HRS images with a higher accuracy. Taking into the consideration both data augmentation and local/global feature fusion, GLGOD-Net is able to obtain an improved performance when detecting objects belonging to the categories of ship, storage tank, tennis court, ground track field, harbor, and bridge. Many objects of those mentioned categories possess structural features scaling in a wide range, which lead the object detection in the tanglesome HSR images to be a challenging task. It is the adoption of adversarial data augmentation, which allows detailed multi-scale information on the object structure to be well preserved, and the local/global feature fusion, which provides more meaningful latent features for object identification, that makes GLGOD-Net performs robustly in various detection tasks in HSR remote sensing imagery.

When the detected objects are evaluated using PRCs, GLGOD-Net still outperforms other baselines in most testing cases. Fig. 5 illustrates the PRCs of different object categories, which are obtained by different methods. As shown, GLGOD-Net is better than any other baseline in most cases, particularly in the categories of airplane, storage tank, tennis court, basketball court, ground track field, and harbor. Based on the remarkable results obtained by GLGOD-Net, it is said that the data augmentation and feature fusion adopted by the proposed framework, is very effective in improving the performance of object detection in HSR remote sensing imagery.

We also recorded the recall values (See Fig. 6) of different object categories predicted by GLGOD-Net, to investigate its effectiveness. As demonstrated, although the ground-truth objects may come from different categories, the recall values of them are relatively high, indicating GLGOD-Net is able to correctly detect a higher proportion of objects in the ground-truth database. Given the remarkable results obtained from the extensive experiments, it is said that GLGOD-Net framework is very effective in detecting different categories of objects in HSR remote sensing imagery.

4.3. Sensitivity analysis

In GLGOD-Net, the number of bounding boxes is an important parameter influencing the performance of object detection in HSR image data. To investigate how different settings of bounding boxes may affect the detection performance, we run GLGOD-Net to detect objects in NWPU VHR-10 dataset, setting the number of boxes as $\{200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000\}$.

Table 1
The AP values of the comparison as well as the proposed object detection methods.

Methods	BoW	SSCBoW	FDDL	COPD	TCNN	NTCNN	RICNN-w	RICNN-f	RPFR-CNN	PSB	GLGOD-Net
Airplane	0.025	0.506	0.292	0.623	0.661	0.701	0.860	0.884	0.904	0.907	0.909
Ship	0.585	0.508	0.376	0.689	0.569	0.637	0.760	0.773	0.750	0.811	0.817
Storage tank	0.632	0.334	0.770	0.637	0.843	0.843	0.850	0.853	0.444	0.805	0.807
Baseball diamond	0.090	0.435	0.258	0.833	0.816	0.836	0.873	0.881	0.899	0.894	0.896
Tennis court	0.047	0.003	0.028	0.321	0.350	0.355	0.396	0.408	0.797	0.773	0.813
Basketball court	0.032	0.150	0.036	0.363	0.459	0.468	0.579	0.585	0.776	0.898	0.852
Ground track field	0.078	0.101	0.201	0.853	0.800	0.812	0.855	0.867	0.877	0.871	0.965
Harbor	0.530	0.583	0.254	0.553	0.620	0.623	0.665	0.686	0.791	0.769	0.797
Bridge	0.122	0.125	0.215	0.148	0.423	0.454	0.585	0.615	0.682	0.784	0.784
Vehicle	0.091	0.336	0.045	0.440	0.429	0.448	0.680	0.711	0.732	0.690	0.768
Mean AP	0.246	0.308	0.245	0.546	0.597	0.618	0.710	0.726	0.765	0.820	0.841

Table 2
Mean AP values/AP values of the proposed GLGOD-Net with different strategies. Different strategy abbreviation: V1: no data augmentation but with local/global feature fusion, V1-L: no data augmentation with only local feature (namely, R-FCN), V1-G: no data augmentation with only global feature (namely, Faster R-CNN+ResNet-101-ohem); V2: with data augmentation but no local/global feature fusion, V2-L: with data augmentation and only local feature, V2-G: with data augmentation and only global feature.

Methods	V1-L	V1-G	V1	V2-L	V2-G	GLGOD-Net
Airplane	0.907	0.904	0.906	0.909	0.906	0.909
Ship	0.786	0.799	0.812	0.862	0.791	0.817
Storage tank	0.816	0.673	0.813	0.774	0.625	0.807
Baseball diamond	0.897	0.895	0.904	0.893	0.900	0.896
Tennis court	0.812	0.716	0.807	0.814	0.800	0.813
Basketball court	0.824	0.772	0.897	0.827	0.730	0.852
Ground track field	0.963	0.976	0.978	0.934	0.958	0.965
Harbor	0.736	0.707	0.753	0.715	0.759	0.797
Bridge	0.681	0.707	0765	0.717	0.727	0.784
Vehicle	0.766	0.652	0.698	0.779	0.743	0.768
Mean AP	0.819	0.780	0.833	0.826	0.794	0.841

Those detected objects are then evaluated by Recall rate and AP, and the corresponding results have been shown in Fig. 7. As the figure shows, GLGOD-Net may perform robustly when the number of bounding boxes is set between 800 and 1800. One is recommended to tune the performance of GLGOD-Net by changing the setting of bounding boxes within this recommended range.

4.4. Time cost on object detection

Besides testing the effectiveness of the proposed framework, we also investigate whether GLGOD-Net can perform the task of object detection efficiently. We recorded the time costed by different object detection approaches. When detecting different objects in HSR images, the time costed by GLGOD-Net is 0.12s. As for other compared baselines, including BoW, SSCBoW, FDDL, COPD, Transferred CNN, Newly trained CNN, RICNN without fine-tuning, RICNN with fine-tuning, R-P-Faster R-CNN, PSB, the detection time is 5.32s, 40.32s, 7.17s, 1.07s, 5.24s, 8.77s, 8.77s, 8.77s, 0.15s, and 0.10s, respectively. It is seen that the computational time of GLGOD-Net is slightly higher than PSB. According to the computational complexity analyzed in section 3.4, we know that the feature fusion in the object detection stage is more computationally demanding. This is the reason why the proposed framework performs the task of object deteciton slightly slower than PSB. Still, the proposed framework is more efficient than other baselines except for PSB. Based on the comparative results related to time consumption, it is said that the proposed framework is computationally efficient when performing different tasks of object detection.

4.5. Case study-objects detected by different approaches

To further investigate whether the proposed learning scheme, i.e., adversarial data augmentation and local/global feature fusion may truly improve object detection in HSR images, we elaborately analyze the objects detected by the proposed framework, and compared them with those detected by other baselines. In Fig. 8, some objects identified

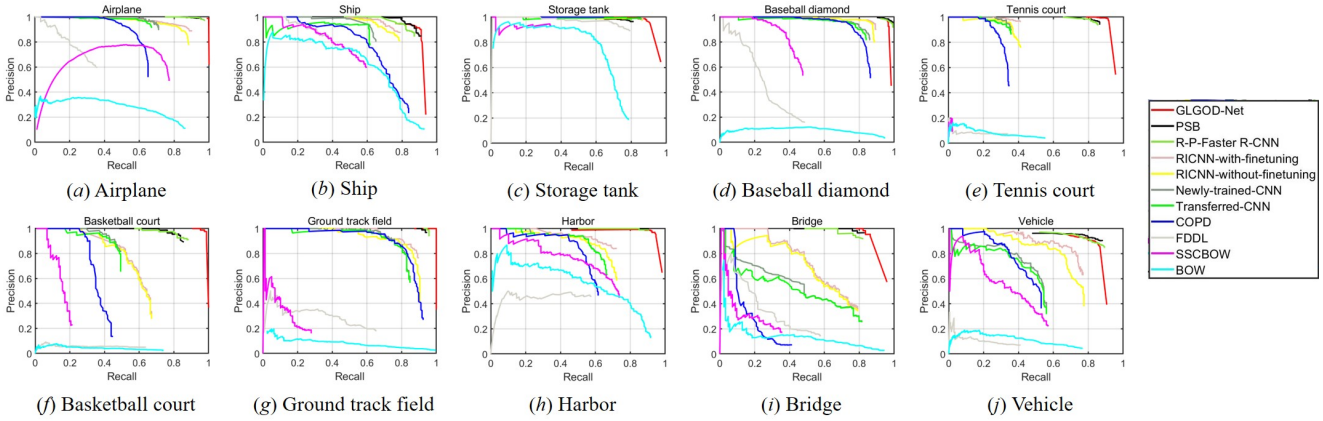


Figure 5: PRCs of GLGOD-Net and other baselines.

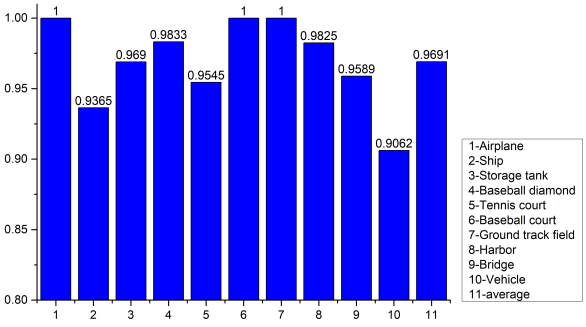


Figure 6: Recall values for the 10 object categories.

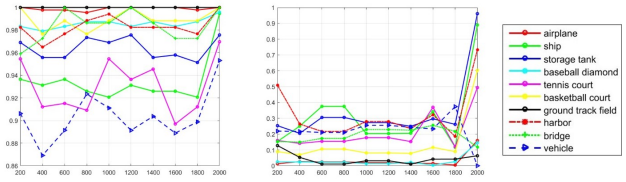


Figure 7: Sensitivity test on the number of object proposals.

by GLGOD-Net and other baselines are exemplified. When compared with R-FCN ignoring data augmentation and feature fusion (Fig. 8. (a)), GLGOD-Net is able to detect objects with different scales, so that the missing rate of detection is reduced. When compared the objects detected by GLGOD-Net with those done by GLGOD-Net only adopting adversarial data augmentation, the proposed framework still performs better (Fig. 8. (b)). Taking the advantage of the fusion of features preserving local/global structural information, GLGOD-Net is capable of distinguishing objects from the complex background. As Fig. 8. (b) shows, challenging objects like basketball court, tennis court, harbor, and car, can be detected by GLGOD-Net with a higher accuracy, while they can not be done by GLGOD-Net (V2). Similar results can also be observed when GLGOD-Net is compared with the baseline only considering feature fusion. As depicted in Fig. 8. (c), the missing rate of airplane, basketball court, bridge, harbor, and ship, detected by GLGOD-Net is much lower than that of GLGOD-Net (V1). It is observed that GLGOD-Net may obtain a better performance when considering utilizing the fused features which are learned from the enhanced HSR image data generated by SRGAN. Under such a learning paradigm, the multi-scale structural information generated by SRGAN can be well utilized to learn the local/global features, comprehensively representing different objects. Consequently, the fusion of these features learned by GLGOD-Net are capable of producing accurate results of object detection in HSR remote sensing imagery.

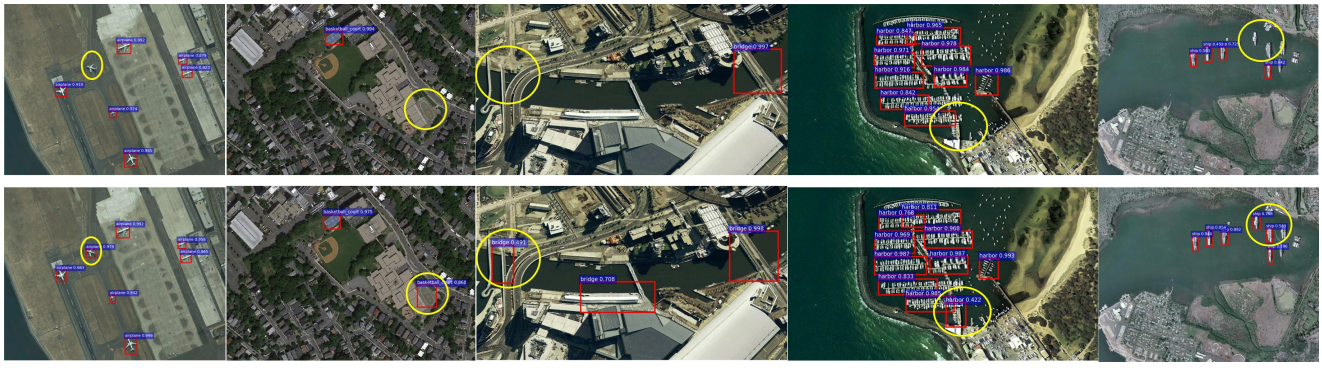
Conclusion

In this paper, a novel deep framework, dubbed GLGOD-Net is proposed to precisely detect objects of interest in HSR remote sensing imagery. Different from previous methods, GLGOD-Net takes into the consideration both data augmentation and local/global feature fusion to improve the accuracy of detecting objects in HSR images. GLGOD-Net firstly adopts SRGAN to recover detail information hidden in the image and amplify the number of samples used for classification-oriented training. Then the fusions of local/global features, which are learned by a fine-designed CNN, are generated according to normalization and element-wise operation strategy. These fusions can be used for precisely detecting objects in HSR image data. The proposed framework has been extensively tested with real world HSR image data and has compared with both classical and state-of-the-art approaches to object detection. The experimental results show that GLGOD-Net outperforms other comparison baselines when detecting different types of objects in HSR images.

In future, we will further improve the performance of

(a) Comparisons on the objects detected by R-FCN and GLGOD-Net. Top: R-FCN; Bottom: GLGOD-Net

(b) Comparisons on the objects detected by GLGOD-Net (V1) and GLGOD-Net. Top: GLGOD-Net (V1); Bottom: GLGOD-Net.



(c) Comparisons on the objects detected by GLGOD-Net (V2) and GLGOD-Net. Top: GLGOD-Net (V2); Bottom: GLGOD-Net.

Figure 8: Objects detected by GLGOD-Net, R-FCN, GLGOD-Net (V1), and GLGOD-Net (V2). Rectangles in the circles represent the objects which are correctly detected.

object detection in the following promising directions. First, we will propose the improved version of GLGOD-Net to effectively detect complex, and small-scale objects in HSR remote sensing images. Second, we will try to refine the oriented box for object detection, and integrate it into different frameworks for object detection in HSR images. Third, we will try to develop novel techniques for deep metric learning, so as to reduce interclass disparity and increase intraclass variability. Last but not the least, we intend to propose novel learning frameworks that can effectively learn the object-object relation for precise object detection.

Acknowledgment

This paper is supported in part by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-RP-2018-004), the National Natural Science Foundation of China under Grant 61802317, and the Data Science and Artificial Intelligence Research Center at Nanyang Technological University. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of National Research Foundation, Singapore.

References

[1] Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS journal of photogrammetry and remote sensing 65, 2–16.

[2] Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., Van der Meer, F., Van der Werff, H., Van Coillie, F., et al., 2014. Geographic object-based image analysis–towards a new paradigm. ISPRS journal of photogrammetry and remote sensing 87, 180–191.

[3] Bontemps, S., Bogaert, P., Titeux, N., Defourny, P., 2008. An object-based change detection method accounting for temporal dependences in time series with medium to coarse spatial resolution. Remote Sensing of Environment 112, 3181–3191.

[4] Cheng, G., Han, J., 2016. A survey on object detection in optical remote sensing images. ISPRS Journal of Photogrammetry and Remote Sensing 117, 11–28.

[5] Cheng, G., Han, J., Guo, L., Qian, X., Zhou, P., Yao, X., Hu, X., 2013. Object detection in remote sensing imagery using a discriminatively trained mixture model. ISPRS Journal of Photogrammetry and Remote Sensing 85, 32–43.

[6] Cheng, G., Han, J., Zhou, P., Guo, L., 2014. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. ISPRS Journal of Photogrammetry and Remote Sensing 98, 119–132.

[7] Cheng, G., Zhou, P., Han, J., 2016. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 54, 7405–7415.

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

[9] Deng, Z., Sun, H., Zhou, S., Zhao, J., Lei, L., Zou, H., 2018. Multi-scale object detection in remote sensing imagery with convolutional neural networks. ISPRS journal of photogrammetry and remote sensing 145, 3–22.

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

[11] Girshick, R., Donahue, J., Darrell, T., Malik, J., 2015. Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence 38, 142–158.

[12] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680.

[13] Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C., 2017. Improved training of wasserstein gans, in: Advances in neural information processing systems, pp. 5767–5777.

[14] Han, J., Zhang, D., Cheng, G., Guo, L., Ren, J., 2014a. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. IEEE Transactions on Geoscience and Remote Sensing 53, 3325–3337.

[15] Han, J., Zhou, P., Zhang, D., Cheng, G., Guo, L., Liu, Z., Bu, S., Wu, J., 2014b. Efficient, simultaneous detection of multi-class geospatial targets based on visual saliency modeling and discriminative learning of sparse coding. ISPRS Journal of Photogrammetry and Remote Sensing 89, 37–48.

[16] Han, X., Zhong, Y., Zhang, L., 2017. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. Remote Sensing 9, 666.

[17] He, K., Zhang, X., Ren, S., Sun, J., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence 37, 1904–1916.

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

[19] He, T., Chan, K.C., 2018. Discovering fuzzy structural patterns for graph analytics. IEEE Transactions on Fuzzy Systems 26, 2785–2796.

[20] He, T., Liu, Y., Ko, T.H., Chan, K.C., Ong, Y.S., 2019. Contextual correlation preserving multiview featured graph clustering. IEEE transactions on cybernetics .

[21] Hu, F., Xia, G.S., Wang, Z., Huang, X., Zhang, L., Sun, H., 2015. Unsupervised feature learning via spectral clustering of multidimensional patches for remotely sensed scene classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8.

[22] Kong, T., Sun, F., Liu, H., Jiang, Y., Shi, J., 2019. Foveabox: Beyond anchor-based object detector. arXiv preprint arXiv:1904.03797 .

[23] Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

[24] LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature 521, 436.

[25] LeCun, Y., Boser, B.E., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W.E., Jackel, L.D., 1990. Handwritten digit recognition with a back-propagation network, in: Advances in neural information processing systems, pp. 396–404.

[26] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

[27] Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4681–4690.

[28] Li, Q., Mou, L., Liu, Q., Wang, Y., Zhu, X.X., 2018. Hsf-net: Multiscale deep feature embedding for ship detection in optical remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing 56, 7147–7161.

[29] Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2117–2125.

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

[31] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

[32] Redmon, J., Farhadi, A., 2017. Yolo9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271.

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

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

[35] Shen, Z., Liu, Z., Li, J., Jiang, Y.G., Chen, Y., Xue, X., 2017. Dsod: Learning deeply supervised object detectors from scratch, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1919–1927.

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

[37] Sun, H., Sun, X., Wang, H., Li, Y., Li, X., 2011. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. IEEE Geoscience and Remote Sensing Letters 9, 109–113.

[38] Tian, Z., Shen, C., Chen, H., He, T., 2019. Fcos: Fully convolutional one-stage object detection. arXiv preprint arXiv:1904.01355 .

[39] Tuermer, S., Kurz, F., Reinartz, P., Stilla, U., 2013. Airborne vehicle detection in dense urban areas using hog features and disparity maps. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6, 2327–2337.

[40] Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W., 2013. Selective search for object recognition. International journal of

computer vision 104, 154–171.

[41] Xia, G.S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. Aid: A benchmark data set for performance evaluation of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing 55, 3965–3981.

[42] Xu, S., Fang, T., Li, D., Wang, S., 2009. Object classification of aerial images with bag-of-visual words. IEEE Geoscience and Remote Sensing Letters 7, 366–370.

[43] Yao, X., Han, J., Cheng, G., Qian, X., Guo, L., 2016. Semantic annotation of high-resolution satellite images via weakly supervised learning. IEEE Transactions on Geoscience and Remote Sensing 54, 3660–3671.

[44] Yao, X., Han, J., Guo, L., Bu, S., Liu, Z., 2015. A coarse-to-fine model for airport detection from remote sensing images using target-oriented visual saliency and crf. Neurocomputing 164, 162–172.

[45] Yokoya, N., Iwasaki, A., 2015. Object detection based on sparse representation and hough voting for optical remote sensing imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 8, 2053–2062.

[46] Zeiler, M.D., Fergus, R., 2014. Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer. pp. 818–833.

[47] Zhang, F., Du, B., Zhang, L., Xu, M., 2016. Weakly supervised learning based on coupled convolutional neural networks for aircraft detection. IEEE Transactions on Geoscience and Remote Sensing 54, 5553–5563.

[48] Zhao, Z.Q., Zheng, P., Xu, S.t., Wu, X., 2019. Object detection with deep learning: A review. IEEE transactions on neural networks and learning systems .

[49] Zhong, P., Wang, R., 2007. A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Transactions on Geoscience and Remote Sensing 45, 3978–3988.

[50] Zhong, Y., Han, X., Zhang, L., 2018. Multi-class geospatial object detection based on a position-sensitive balancing framework for high spatial resolution remote sensing imagery. ISPRS journal of photogrammetry and remote sensing 138, 281–294.

[51] Zhu, C., He, Y., Savvides, M., 2019. Feature selective anchor-free module for single-shot object detection. arXiv preprint arXiv:1903.00621 .

[52] Zhu, Y., Zhao, C., Wang, J., Zhao, X., Wu, Y., Lu, H., 2017. Couplenet: Coupling global structure with local parts for object detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4126–4134.

[53] Zou, Z., Shi, Z., Guo, Y., Ye, J., 2019. Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055 .