This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

SKNet: Detecting Rotated Ships as Keypoints in Optical Remote Sensing Images

Zhenyu Cui, *Member, IEEE*, Jiaxu Leng, Ying Liu, *Member, IEEE*, Tianlin Zhang, Pei Quan, and Wei Zhao

***Abstract*— Detecting rotated ships is difficult in optical remote sensing images due to the challenges of complex scenes. Existing advanced rotated ship detectors are typically anchor-based algorithms that require plenty of predefined anchors. However, the use of anchors brings three critical problems: 1) a large number of anchors bring a huge amount of calculation; 2) the attributes (e.g., size and aspect ratios) of anchors are designed via *ad hoc* heuristics; and 3) only a tiny fraction of anchors that overlap with ground-truth bounding boxes of ships tightly can be considered as positive samples, which causes an extreme imbalance between positive and negative samples. As a result, the detection accuracy will be influenced seriously when the design of anchors is not suitable. To address the above problems, this article proposes a novel anchor-free rotated ship detection framework, called SKNet, which detects rotated ships as keypoints in optical remote sensing images. In SKNet, a ship target is modeled as its center keypoint and morphological sizes, including the width, height, and rotation angle. Accordingly, we design two customized modules: orthogonal pooling and soft-rotate-nonmaximum suppression (NMS), where the former is to improve the prediction accuracy of the center keypoint and the morphological size, and the latter is to effectively remove redundant rotated ship detection results. Extensive experiments are conducted to demonstrate the effectiveness of SKNet on three optical remote sensing image data sets: HRSC2016, DOTA-ship, and HPDM-OSOD, which is collected by ourselves and published in this article. Empirical studies show that SKNet achieves state-of-the-art detection performance while being time-efficient. Overall, SKNet achieves the best speed–accuracy tradeoff.**

***Index Terms*— Keypoints, optical remote sensing, rotated ship, ship detection.**

Manuscript received October 22, 2020; revised December 14, 2020; accepted January 4, 2021. This work was supported in part by the Natural Science Foundation of China under Grant 71671178, in part by the Equipment Advance Research Fund under Grant 6142502180101, and in part by the Fundamental Research Funds for the Central Universities. *(Zhenyu Cui and Jiaxu Leng are co-first authors.) (Corresponding author: Ying Liu.)*

Zhenyu Cui and Ying Liu are with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101400, China, and also with the Lab of Data Mining and High Performance Computing, University of Chinese Academy of Sciences, Beijing 101400, China (e-mail: yingliu@ucas.ac.cn).

Jiaxu Leng is with the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing 40065, China.

Tianlin Zhang is with the National Centre for Text Mining, School of Computer Science, The University of Manchester, Manchester M1 7DN, U.K.

Pei Quan and Wei Zhao are with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101400, China.

HPDM-OSOD can be available at this URL: "http://hpdm.ucas.ac.cn/".

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TGRS.2021.3053311.

Digital Object Identifier 10.1109/TGRS.2021.3053311

I. Introduction

SHIP detection in optical remote sensing images is of vital significance for maritime security and other applications, e.g., traffic surveillance and protection against illegal fishers. In recent years, with the rapid development of remote sensing technology, plenty of available remote sensing images with high resolution [2] have greatly promoted ship detection [3]–[6]. Meanwhile, it also brings challenges due to the special shooting angles, the complex scenes, and the difference in the scales of the objects in a single image. In optical remote sensing images, ships are hardly deformed and the internal texture is usually stable. As a result, the orientation of the ship targets can be defined based on these consensuses, which is important for ship pose estimation, maritime traffic control, and so on. Therefore, it claims higher requirements for general object detectors in ship rotation angle prediction and boundary positioning.

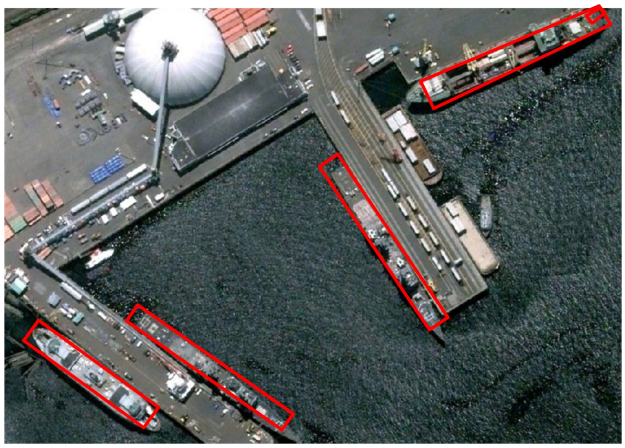
Recently, deep learning has become increasingly popular [7]–[9], which has achieved state-of-the-art results in public data sets (e.g., PASCAL VOC [10] and MS COCO [11]) for object detection. The convolutional neural network (CNN)-based object detection method can be roughly divided into one-stage algorithms and two-stage algorithms considering when the bounding boxes are generated. Two-stage algorithms (e.g., Faster R-CNN [7]) extract features by connecting CNNs and apply region proposal layers to generate dense anchors, which are used to fit the candidate detections. Two-stage object detectors have relatively high accuracy with low efficiency due to the massive computation for candidate generation and the repeated feature calculation for classification. In order to deal with the issue, one-stage object detectors (e.g., YOLO [8] and SSD [9]) are proposed and applied in real scenarios, which regress the detection boxes and categories simultaneously. One-stage object detectors greatly improve the inference speed. However, it makes a compromise in precision.

Despite the success of the above detectors for general purpose, the axis-aligned detection cannot outline the actual pose of the object accurately in some specific scenarios. Affected by the shooting angle or the geographical environment, general detection boxes usually contain plenty of background pixels and even regions supposed to belong to the neighboring objects. Rotated boxes can alleviate this uncertainty by fitting the boundary of ship targets tightly, which is also conducive to pose estimation and trajectory prediction. Nevertheless,

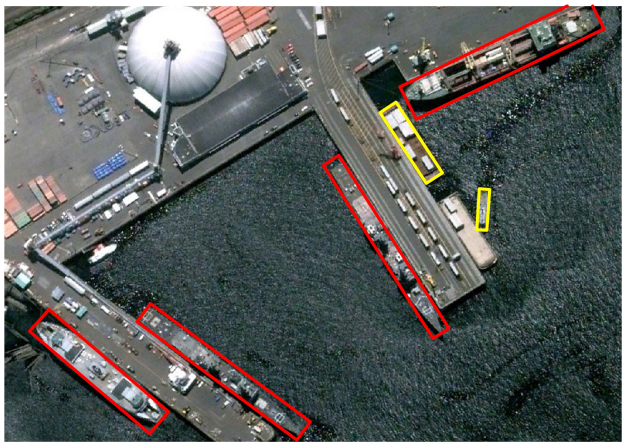
0196-2892 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a)



(b)

Fig. 1. Visual comparison between the proposed SKNet and a SOTA detector [1], where (a) shows the detection results of the SOTA detector and (b) shows the results of our SKNet in the same image. Ships detected by RoI-Transformer [1] are colored red, while those failed to be detected by RoI-Transformer but detected by SKNet are colored yellow. It can be seen that SKNet has advantages in ship locationing and boundary positioning.

rotated object detection is a relatively difficult task in object detection, which requires predicting not only the position and size of the target but also the rotation angle. A large number of rotated object detectors [12], [13] typically were derived from the general algorithms by enriching the prediction of the rotation angle. In order to predict the rotation angle, some existing methods [13], [14] extend the anchor, which has been applied to the region proposal network (RPN) [7] of two-stage detectors. Although the existing algorithms improved the rotated ship detection in optical remote sensing images in certain degree, the performance of such detectors is not able to meet the requirement of the real applications.

Fig. 1(a) shows the visual detection result of region of interest (RoI)-transformer [1], a state-of-the-art rotated object detector, while Fig. 1(b) shows the result of the proposed SKNet in the same image. As shown in Fig. 1(a), RoI-Transformer failed to accurately predict the boundary of ships. Specifically, for anchor-based algorithms represented by RoI-Transformer, the attributes of the anchors based on the heuristic design cannot accurately fit ship targets with large scale variance. In addition, although the sliding anchors are distributed densely, a tiny fraction of the positive samples are available that overlap with the ground-truth bounding boxes of ships tightly, resulting in the underfitting of the model. To get rid of the anchors, some anchor-free detectors [15], [16] are proposed to directly predict the attributes of the target rather than fitting it to a set of predefined anchors, which brings great difficulty to inference. In addition, most anchor-free detectors need to conduct corner grouping, which may lead to incorrect detections.

To overcome these problems, we propose a novel ship detection framework, called SKNet, to detect rotated ships. In our approach, a ship target is modeled as its center keypoint and morphological size, including the width, height, and rotation angle. Accordingly, we design two customized modules: orthogonal pooling and soft-rotate-nonmaximum suppression (NMS). Specifically, the contributions of our work are given as follows.

1) A novel one-stage anchor-free ship detection framework is proposed, which completely converts the detection of the rotated ships into the prediction of center keypoints and morphological sizes.
2) A novel module, called "orthogonal pooling," is proposed to extract the features of rotated ships, which can generalize ship features in any rotation angles and help to predict the keypoint and its corresponding attributes.
3) Combined with rotation, we adopt a novel module, called soft-rotate-NMS, which can remove the neighboring redundant detection boxes. We improve the soft-NMS by incorporating the features of rotation, which simultaneously suppresses the confidence of the nearby detections.
4) We verified the proposed SKNet in two public data sets, HRSC2016 [17] and DOTA [2], and an optical remote sensing image data set collected by ourselves, called HPDM-OSOD, which is also published in this article. Experimental results demonstrate that SKNet achieves state-of-the-art detection performance while being time-efficient.

The remaining of this article is organized as follows. Section II states the related work of object detection based on deep learning, especially the detection of rotated objects. In Section III, the proposed SKNet is presented in detail, and the experiment results are presented to demonstrate the effectiveness of our framework in Section IV. Finally, the conclusion of this article is provided in Section V.

II. Related Work

A. General Detectors

General detectors can be roughly divided into two categories: one-stage and two-stage detectors. Two-stage detectors generate candidates and classify each RoI using corresponding visual features respectively. As a representative of the two-stage detector, R-CNN [18] adopts selective search [19] for the first time to produce boundaries and classifies the combined bounding boxes by CNNs. However, the selective search is inefficient and unhelpful when it is performed in complex

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

scenes. Faster R-CNN [7] uses RPN to replace and optimize the bounding boxes generation and normalizes the features of RoIs for classification with RoI-Pooling. SPP-Net [20] proposes the spatial pyramid pooling layer, which directly deals with the generated fixed-length vectors extracted from the feature maps for the regression and classification of the candidate, which improves the inference speed. Similarly, R-FCN [21] improves feature selection by proposing a position-sensitive score map, which integrates the position of candidates into RoI-Pooling. Learning from cascading, Cascade R-CNN [22] connects three identical object regression networks with a series of increasing intersection-over-union (IOU) thresholds sequentially, which constantly predicts the bounding boxes correctly and achieves improvements. Later on, considering the huge cost of anchors, GA-RPN [23] adaptively generates sparse and varied anchors using the features of the image, which are placed in the region with a higher probability of interest. Libra R-CNN [24] proposes a more balanced loss function with a novel sampling method, which achieved further breakthroughs in public data sets.

Compared with two-stage detectors, one-stage detectors generate the boundary and the category of the objects simultaneously. YOLO [8] adopts a blocking strategy to label each block with a particular category. Although YOLO improved the efficiency, it does not work well for small targets because a single block cannot handle dense and small objects included inside. SSD [9] performs detection on the feature maps of different scales at different levels, which improves the detecting precision of small objects. Referring to the anchor mechanism in Faster R-CNN [7], RefineDet [25] further corrects the bounding boxes with the anchor refinement module. SNIP [26] proposes a new training method and fuses features of different scales in the feature pyramid. DSSD [27] enriches the features extracted by the neural networks and obtained great enhancement. FSAF [28] detects targets on the feature maps at different levels, which helps to select the feature pattern with high confidence.

B. Rotated Object Detectors

Rotated object detection not only predicts the location and size of the objects but also predicts the rotation angle, which is mainly used for optical character recognition (OCR) and remote sensing object detection.

RRPN [13] first introduces the idea of RPN to the rotated text detection. It enriches the rotation angle of the anchor in RPN, which enumerates the predefined bounding boxes in feature maps by different heights, widths, and rotation angles. Based on the generated detection boxes, RRoI-Pooling is proposed accordingly to extract features in rotated candidates. RRPN directly inherits the architecture of Faster R-CNN and achieves the analogous effectiveness based on the multiplicative inference time cost. Similarly, also based on Faster R-CNN, R2CNN [29] innovatively redefines an orientation object like a pair of points in clockwise and the height of a rotated rectangle. At the same time, considering the large aspect ratio of the text in orientation robust text scenes, R2CNN augments the size of RoI-Pooling with two extreme pooled sizes (3×11 and 11×3), which is beneficial to classification. Learning from the RRPN, R2CNN combines three pooling methods into the RRoI-Pooling layer to extract features in rotated regions, including RoI-Pooling, Free-form RoI-Pooling, and RRoI-Pooling. In order to detect rotated objects, the above methods adopt the rotated anchors and RoI-Pooling methods. However, the structure (CNN + RPN + classification) brings trivial precision but massive calculations. The research for the RoIs has never stopped. Ding *et al.* [1] raise a dedicated architecture, called RoI-Transformer, to learn rotated candidates in axis-aligned anchors and then locate and classify them with the output of rotated position-sensitive RoI-Align. Due to the compact anchors, the rotated detection results can be achieved faster. So far, RoI-Transformer is one of the state-of-the-art algorithms. Affected by the success of attention mechanism and context-aware learning, CAD-Net [6] raises a context-aware detection network, which consists of two subnetworks: GCNet and PLCNet, to consider more contextual information in the feature extraction stage and introduce spatial attention mechanism to the RoI-Align stage.

One-stage detectors also performed well in the rotated object detection. EAST [15] transforms the definition of rotated boxes and pioneers the anchor-free rotated object detector. It detects the feature points, the angle and the distance to the edges of the object. In the training stage, a degenerated region of feature points is helpful to obtain positive samples. Locality-aware NMS is proposed to merge and deduplicate the detection boxes generated by different layers and get great acceleration. TextBoxes++ [30] detects horizontal and rotated rectangles using the architecture of SSD [9], which tries an irregular rectangular convolution kernel to adapt the text with large aspect ratio. Meanwhile, a new online hard example mining (OHEM) [31] is helped to train the network in two phases with different proportion of positive and negative samples, which also improves the consumption of NMS by cascading a coarse bounding box filter to reduce the calculation of the Rotate-NMS. R3DET [5] claims that some feature points where the anchor is located are limited by the receptive field, which is not adequate for regression. Thus, it predicts the rotated objects based on the axis-aligned anchors and more precise rotated detection will be trained to approach the ground truth. The backbone follows RetinaNet [32], where feature refinement modules are cascaded for more refined feature maps.

C. Anchor-Free Detectors

Anchor-free detection algorithms directly cast away the anchor, for the purpose of preventing the selection of hyper-parameters and promote the inference speed. CornerNet [16] replaces anchors by predicting corner keypoint pairs. However, the embedding vector that is predicted to group the corner points is represented by a single scalar, and the resulting lack of similarity space carries mismatches for candidate objects. CenterNet-Triple [33] additionally joins the center keypoint of the object, which excludes wrong corner pairs effectively. Another CenterNet [34] predicts each object by a single keypoint and avoids the mismatch caused by grouping.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

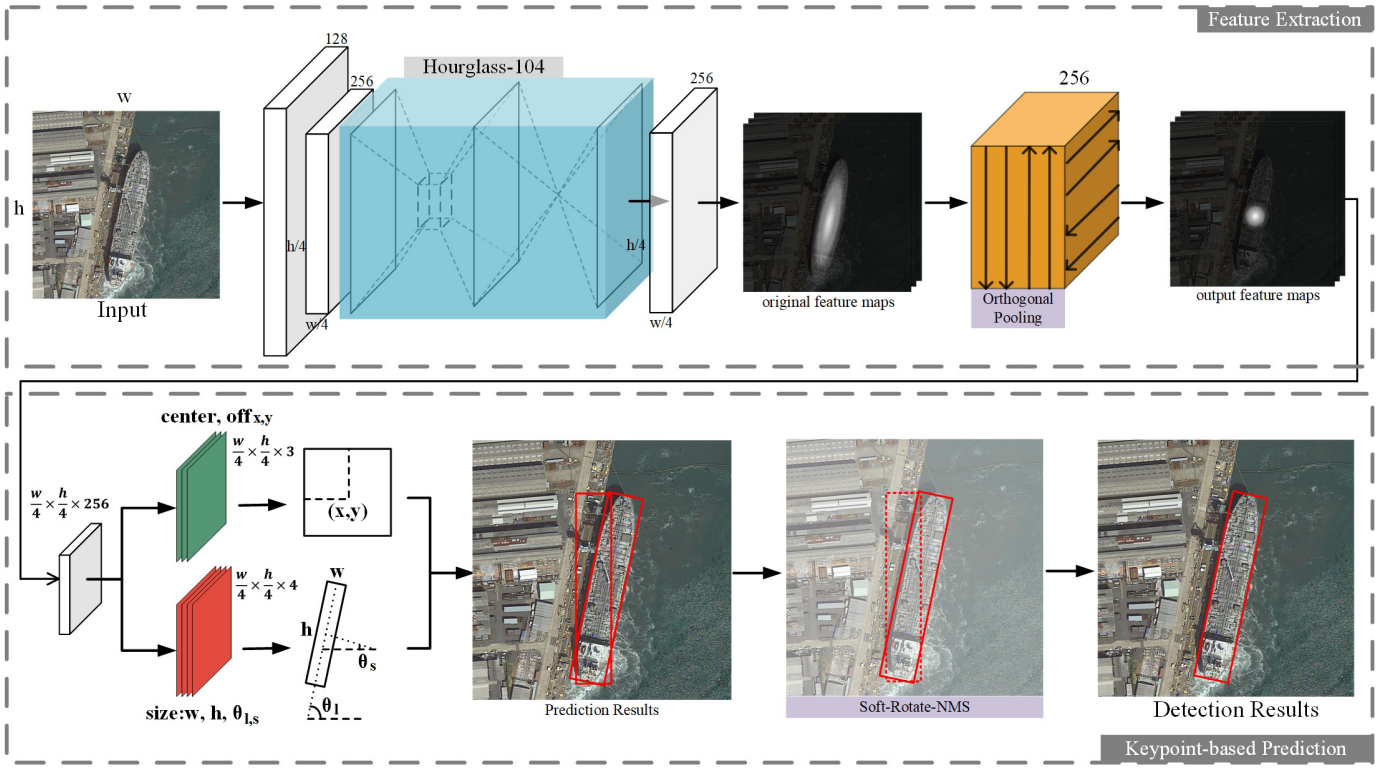


Fig. 2. Architecture of SKNet. SKNet consists of two phases: feature extraction and keypoint-based prediction. In the feature extraction phase, abundant visual features in multiple levels are extracted and integrated by orthogonal pooling. Then, the feature maps will be supervised for the prediction of the center keypoint and the morphological sizes. By combining the above predictions, soft-rotate-NMS produces the final detection results.

In this article, we propose a framework for detecting rotated ships based on keypoints. In order to predict the keypoints and attributes of ship targets effectively, we propose a novel module, called "orthogonal pooling," to enrich the center of the object with more global information of object. For the prediction of the rotation angle of the object, we directly performed the regression of the angle in multiple directions. In the inference stage, we adopt soft-rotate-NMS, which contributes to the mergence of redundant rotated boxes, compared with NMS. The experimental results on multiple data sets further demonstrate that our proposed SKNet is able to achieve the best speed–accuracy tradeoff.

III. Our Approach

A. SKNet

Different from general detectors, the proposed SKNet is a one-stage, anchor-free framework. In SKNet, each rotated ship object is modeled as a center keypoint and morphological sizes, including the width, height, and rotation angle, which are combined into detection results. The entire detection process of SKNet can be described as follows:

$$\text{SKNet}(I) = f_{\text{SR}-\text{NMS}}(\eta(\mathcal{P}_{a,k}(\text{Op}(\xi(I_{i,j}))))) \quad (1)$$

where I is the input image for SKNet, ξ represents the backbone network that consists of stacked CNNs, Op represents the proposed novel module called "orthogonal pooling," $\mathcal{P}_k$ and $\mathcal{P}_a$ denote the prediction of the center keypoint and the morphological sizes, respectively, η denotes the combination of the predictions, and $f_{\text{SR}-\text{NMS}}$ denotes the proposed soft-rotate-NMS.

The architecture of SKNet is shown in Fig. 2. Proceeding from the input, SKNet mainly consists of a feature extraction phase and a keypoint-based prediction phase. SKNet adopts Hourglass [35] as the backbone network for feature extraction. To enrich the global visual pattern for better keypiont and morphological size prediction, we propose a novel module, called "orthogonal pooling." Then, the keypoint-based prediction simultaneously predicts the center keypoint heatmap of the object's center, the offset compensation of the keypoints caused by downsampling, and the size (width and height) of the object with its rotation angle. Among them, the heatmap is a single-channel feature map, where each value is normalized to $0 \sim 1$ and represents the probability of the center keypoint of the ship. We screen the potential targets out through the center keypoint heatmap and restore the original position by the predicted offset at the center's position. Features of the potential center position contain rich abstract patterns that are beneficial to the prediction of the semantic attribute. At the same time, the morphological size of the object at the corresponding position is also predicted and combined with the center position to generate the prediction results.

In addition, affected by the shooting angle and imaging methodology, the distortion of the ship occasionally hinders the angle prediction. We integrate the semantic features of the rotated ship in multiple directions to avoid the vagueness

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

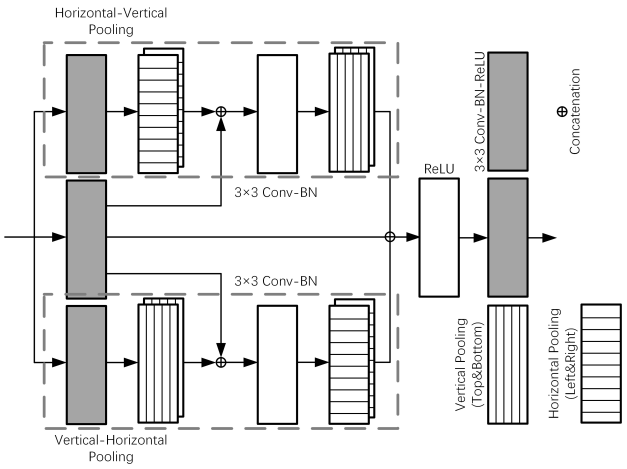


Fig. 3. Orthogonal pooling consists of two feature transfers in orthogonal directions, which arranges the horizontal and vertical pooling in sequence. The horizontal pooling consists of the maximum respond transfer in the left and right directions, while the vertical pooling consists of similar transfers toward the top and bottom.

caused by a single-angle prediction. Specifically, we make the same angle prediction for the long side and the short side of the ship and average the two predicted angles as (2) to integrate the rotation angles of the ship in different directions

$$\theta_{\text{pred}} = \frac{\theta_l + \left(\frac{\pi}{2} - \theta_s\right)}{2} \tag{2}$$

where θ_l represents the orientation of the long side, θ_s represents the other, and θ_{pred} represents the predicted rotation angle. Finally, we design a novel NMS method, called soft-rotate-NMS, to merge redundant rotated detection boxes.

B. Orthogonal Pooling

The prediction of the center keypoint is crucial in SKNet. In order to aggregate more useful information at the center, more global features need to be considered to estimate the keypoints and bounding boxes. SKNet learns to make each ship collapse into a single point at the center and abstract semantic features (height, width, and rotation angle) at the corresponding position. Therefore, the global features of the object need to be transferred to its center point to effectively augment the feature representation.

Rotate RoI-Pooling [13] is a general method for pooling in rotated regions, which divides the rotation candidates into quantitative blocks and takes the maximum response to construct feature vectors. However, Rotate RoI-Pooling is not suitable in our SKNet because the single-stage detector cannot perform exclusive pooling in a predictable specific direction. The geometric centers of objects do not necessarily convey very recognizable visual patterns [33]. Given that ship targets usually have extreme aspect ratios, local pooling methods with a fixed size (e.g., max-pooling with a size of 3×3) are likely to contain noticeable visual noises. To address this issue, we first look along an orthogonal direction to extend a maximum value on the boundary, transfer it to the centers in the other orthogonal direction, and concatenate all possible combinations of transferable paths. Therefore, we propose orthogonal pooling for ships with arbitrary rotation angles, which aggregates the global maximum response of an object to its center keypoint.

The maximum response distributed in any orientation can always be transmitted to the center through the decomposition in two orthogonal vectors: vertical and horizontal vectors. Therefore, we design the orthogonal pooling based on the above theory. Specifically, for a series of feature maps extracted by the backbone network, we first pass a potential maximum value on it in both vertical and horizontal directions and generate the transferred feature maps simultaneously. Then, we pass the value in another orthogonal direction for these two kinds of transferred feature maps, respectively. We denote $f_{(i,j)}$ as a stronger response value than the neighbor, which is waiting to be transferred in the input feature maps, g as the transferred feature map, and $o_{(x,y)}$ as the output value at the center of the object. We take the take vertical downward and horizontal leftward, respectively, as an example. The pooling process of maximum value transferred in the orthogonal pooling can be expressed as follows:

$$\begin{cases} g_{(i,y)} = \max(f_{(i,j)}, f_{(i,j-1)}, \ldots, f_{(i,y)}) \\ g'_{(i,y)} = \text{BN}(\text{Conv}_{3\times3}(g_{(i,y)})) \\ o_{(x,y)} = \max\left(g'_{(i,y)}, g'_{(i-1,y)}, \ldots, g'_{(x,y)}\right). \end{cases} \tag{3}$$

Due to the uncertainty of the rotation angle, it is necessary to combine any possible direction for the transfer of the maximum value. The complete structure of the orthogonal pooling is shown in Fig. 3, where some essential 3×3 Conv-BN [36] layers and ReLU layers are incorporated to enrich the orthogonal pooling to learning incrementally.

As shown in Fig. 3, we first generate feature maps with three branches by performing the 3×3 convolutional block with batch normalization [36] and ReLU [37]. For the feature maps of the first branch and the third one, we perform the horizontal pooling (leftward and rightward) and the vertical pooling (upward and downward), respectively. Then, we fuse these feature maps with the second branch to generate the transferred feature maps simultaneously. Besides, we perform the vertical pooling and the horizontal pooling for the transferred feature maps of the first and third branches. Finally, we combine the output of the three branches by the 3×3 convolutional block with batch normalization and ReLU to complete the output feature maps.

C. Soft-Rotate-NMS

NMS [38], as important postprocessing, is used to merge redundant detections of the given object. In remote sensing images, rotated and neighboring ships, sometimes, produce a large overlap at a certain angle with each other in Fig. 5(a). When the IOU of different objects exceeds a predefined parameter, the correct detections will be removed unfortunately by the original NMS. Motivated by Soft-NMS [39], we propose a novel strategy, called soft-rotate-NMS, which updates the remaining neighboring detections' confidence with its overlap when selecting detections with the highest confidence, instead of removing them directly by the overlap.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

Algorithm 1 Soft-Rotate-NMS

Input: $B = [b_1, \ldots, b_n], S = [s_1, \ldots, s_n]$
Output: Boxes with corresponding Scores: D, S

$D \leftarrow \{\}$
while $B \neq null$ **do**
  $i \leftarrow argmax(S)$
  $D \leftarrow D \cup \{b_i\}$
  $B \leftarrow B - \{b_i\}$
  for b_j *in* B **do**
    $I \leftarrow Rotate\_IOU(b_i \cap b_j)$
    $dec = I/(area(b_i) + area(b_j) - I)$
    $s_i \leftarrow s_i \times dec$
  end
end
return D, S

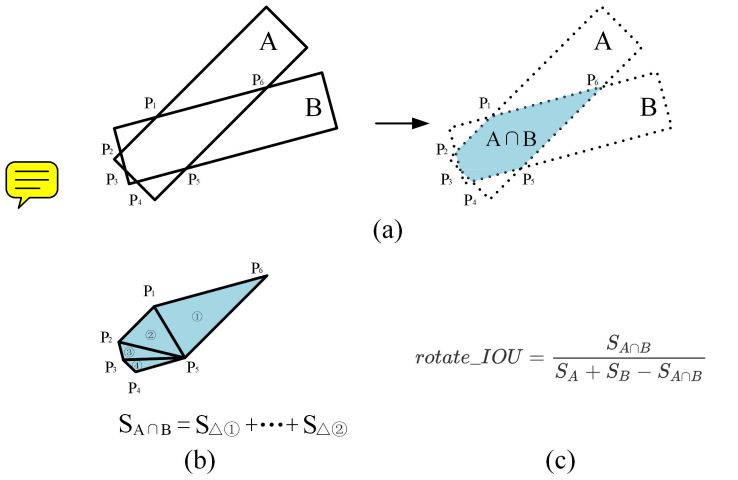


Fig. 4. Calculation of Rotate_IOU. (a) Calculation convex hull $(P_1, \ldots, P_n)$ of the intersection area of two rotated rectangles A and B. (b) Division of the convex hull and the calculation of its area.(c) Calculation of the rotate_IOU.

We do not adopt the strategy that fuses the factor of rotation angle to the confidence updating directly. The exploratory experimental results also showed that this method has negligible impact on the further improvement of the effect. We believe that there are two reasons behind it. On the one hand, two objects with great variance in direction rarely produce a large overlap, which can hardly result in removal. On the other hand, Rotate_IOU between ships with little variance in direction has indicated the difference well. We denote N detections and their corresponding confidences as $(b_1, \ldots, b_n)$ and $(s_1, \ldots, s_n)$, respectively. The process of soft-rotate-NMS is described in Algorithm 1, where Rotate_IOU(A ∪ B) denotes the IOU between the rotated box A and box B.

The computation of the Rotate_IOU is shown in Fig. 4, which first calculates the convex hull $(P_1, \ldots, P_n)$ of the intersection area of two rotated rectangles and then calculates the area of this convex hull to produce the overlap and the IOU. Fig. 5 shows a sample of the comparison between soft-rotate-NMS and Rotate-NMS algorithms.

Fig. 5. Comparison of original NMS and soft-rotate-NMS. (a) Effectiveness of the original rotate-NMS. (b) Effectiveness of the soft-rotate-NMS.

D. Details in Training and Inference

1) Loss Function: SKNet can be trained end-to-end. We predict the center keypoints (C) of the ship with its offset (O), width (W), height (H), and rotation angle (θ) simultaneously, which are supervised during the training. The multiloss of SKNet is shown in (4), which is optimized by the Adam [40] optimizer for gradient descent

$$\text{Loss} = L_c + \alpha(L_w + L_h) + \beta L_\theta + \gamma L_O. \tag{4}$$

In (4), L_c denotes the loss of the center keypoint. An enhanced focal loss [32] is adopted to compromise extremely few and unbalanced positive pixel-level samples, as shown in (5), where p_{ij} denotes the predicted probability of the central point at (i, j), and g_{ij} denotes the ground-truth heatmap, where 1 for the center location of objects and 0 for the others. However, as the number of the positive samples is much smaller than that of the negative ones, it brings difficulties to network learning. Specifically, even if the predicted center is close to the ground-truth value, the loss is the same as the one far away from it since they are nonoverlapping, which does not reward the former situation. Therefore, we use the Gaussian blur to increase the probability around the center keypoints to reward the nearby prediction. Assuming the reward value is g_{ij}, the polynomial $(1 - g_{ij})$ in (5) effectively reduces the prediction error of the confidence of the position around the positive samples. Parameters, α and β, are set at 2 and 4, respectively. In the remaining of (4), L_w, L_h, L_θ, and L_O are all the same as the smooth L1 loss [41], as shown in (6)–(9), where gt and pred mean the ground truth and the predicted value. Especially, P_k in L_O means the position of the positive sample. Due to the downsampling in the backbone network, polynomial $(P_k/m) - \lfloor P_k/m \rfloor$ is the deviation to its original picture, where m denotes the sampling multiple

$$L_c = -\frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\begin{cases}(1-p_{ij})^\alpha \log(p_{ij}), & g_{ij}=1\\ (1-g_{ij})^\beta (p_{ij})^\alpha \log(1-p_{ij}), & \text{others}\end{cases} \tag{5}$$

$$L_w = \frac{1}{N}\sum_{k=1}^{N}\text{SmoothL1}(\text{gt\_}W_k, \text{pred\_}W_k) \tag{6}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

$$L_h = \frac{1}{N}\sum_{k=1}^{N} \text{SmoothL1}(\text{gt\_}H_k, \text{pred\_}H_k) \tag{7}$$

$$L_\theta = \frac{1}{N}\sum_{k=1}^{N} \text{SmoothL1}(\text{gt\_}\theta_k, \text{pred\_}\theta_k) \tag{8}$$

$$L_O = \frac{1}{N}\sum_{k=1}^{N} \text{SmoothL1}\left(\frac{P_{k_{x,y}}}{m} - \left\lfloor \frac{P_{k_{x,y}}}{m} \right\rfloor, \text{pred\_}O_{k_{x,y}}\right). \tag{9}$$

The loss of the rotation angle is slightly different from the others, which is worth a unique preprocessing. If the difference between the predicted angle and the ground truth is more than 90°, the direct use of smooth L1 loss will bring a greater loss than the case of 90°. However, in actual detection, there is no two detections whose intersection angle exceeds 90°. Two detections that are greater than 90° can always be reexpressed with a complementing angle less than 90°. Thus, we adjust the ground truth to guide the rotation angle to regress to its complement angle for such cases. The adjustment policy is expressed as follows:

$$\theta_{\text{gt}} = \begin{cases} \theta_{\text{gt}} + \pi, & \theta_{\text{pred}} - \theta_{\text{gt}} > \dfrac{\pi}{2} \\ \theta_{\text{gt}} - \pi, & \theta_{\text{pred}} - \theta_{\text{gt}} < -\dfrac{\pi}{2}. \end{cases} \tag{10}$$

For the case of intersection angles greater than 90°, we set the actual angle (θ) to $\theta - \pi$. For the ones less than $-90°$, we set the actual angle (θ) to $\theta + \pi$. Then, the loss calculated by the adjusted angles could effectively overcome the problem.

2) Details in Training: Learning from the previous practice, the weights in (4): α, β, and γ are set to 0.1, 0.1, and 1, respectively. We use a pretrained model [33] for transfer learning in SKNet to speed up the training. We train the network on Titan X (Pascal). When training, we set the batch size to 3, the initial learning rate to $2.5*10^{-4}$, and the learning rate decaying to 0.1 with a step of 10K. Model is trained until validation errors stop decreasing. We apply data augmentation, including random mirroring, random proportional scaling, and random cutting to process the images to a preset size (511 × 511) and feed them to SKNet due to the limit of the memory size.

3) Details in Inference: During inferring, we use a kernel size of 3 × 3 to suppress the nonmaximum around the keypoints on the heatmap for removing redundant center keypoints. Then, we obtain the points whose scores are greater than the threshold (0.5) and their corresponding parameters ($\text{offset}_{x,y}$, W, H, and θ) to form the detecting boxes. Finally, by synthesizing the flipped image with the original one, we use soft-rotate-NMS to remove the redundant detecting boxes.

IV. Experimental Results and Analysis

A. Data Set

We verified the proposed SKNet in three optical remote sensing image data sets: HRSC2016, DOTA-ship, and HPDM-OSOD. HRSC2016 and DOTA are two popular public data sets, which have been widely used to measure the general performance of different models. However, they still insufficient in the diversity of scenes and ship targets, which failed to reflect different harbors and marine environments in reality. Particularly, as shown in Fig. 10(a), most images in HRSC2016 only contain a few large-scale salient objects (usually 1-2). Similarly, small ship targets (usually occupying less than 1000 pixels) account for 76.5% of all ship targets in DOTA. In order to make up for the extremely unbalanced scale distribution of the public data sets, we collect and will publish an optical remote sensing image data set in this article, called HPDM-OSOD (oriented ship object data set), specifically for ship detection to make more convincing ablation studies and fill the gaps in the industry.

1) HRSC2016: HRSC2016 [17] is a popular public optical remote sensing image data set, which contains 1070 images with an average width of 1000 pixels. HRSC2016 contains 2976 instances with rotated bounding boxes in the annotation. As one of the most popular optical ship detection data sets, HRSC2016 has advantages in abundant object features, distinctive shooting angle, multiple directions, and complicated backgrounds and has become the benchmark in the industry. Many previous studies are based on HRSC2016 and have milestone significance. Therefore, HRSC2016 is used as one of the benchmarks in our work for longitudinal comparison with the state-of-the-art algorithms. We use 626 of the 1070 pictures for training and the rest for validation.

2) DOTA-Ship: DOTA [2] is one of the biggest public optical remote sensing image data sets for object detection, which contains ∼3k images with an average width of 4k pixels. To verify the proposed ship detection method, we extracted all the pictures that contain ship targets and performed validation based on this subdata set of DOTA (DOTA-ship). DOTA-ship contains 573 images with 43 738 rotated ship targets: 435 of them are for training and the rest are for testing. Particularly, due to the excessively large-scale pictures in DOTA, we crop the pictures into the size of 1024 × 1024 and keep the target that contains at least 70% internally.

3) HPDM-OSOD: In order to complement the lack of diversity in the public data set, we collected and will publish an optical remote sensing image data set for ship detection, called HPDM-OSOD, to verify the performance of SKNet in various scenes. HPDM-OSOD contains 1127 images and 5564 rotated ship targets, as shown in Fig. 10(b). The source image of HPDM-OSOD comes from Google Earth with distinct scenes around the world. In addition, the average resolution of these pictures is about 1000 × 2000. In comparison, the quantity and quality of the ship in HPDM-OSOD are greatly higher than the analogous data sets [17], [42], [43].

We label each target as four corner points, that is (x_1, y_1, . . . , x_4, y_4), where 1∼4 denote four corners in clockwise and (x_i, y_i) denotes the coordinate, as shown in Fig. 6. All these annotations are format as both PASCAL [10] and COCO [11]. We randomly select 3/5 of original images for training and 2/5 for validation.

4) Comparison With HRSC2016 and HPDM-OSOD:

a) Complex scenes: In HPDM-OSOD, we averagely collect pictures in lands, ports, bays, and sea on six continents, including Asian, Africa, North/South America, Europe, and Oceania, which increases the diversity of HPDM-OSOD. Moreover, compared with HRSC2016, HPDM-OSOD contains more small objects. As shown in Fig. 7, we make statistics

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

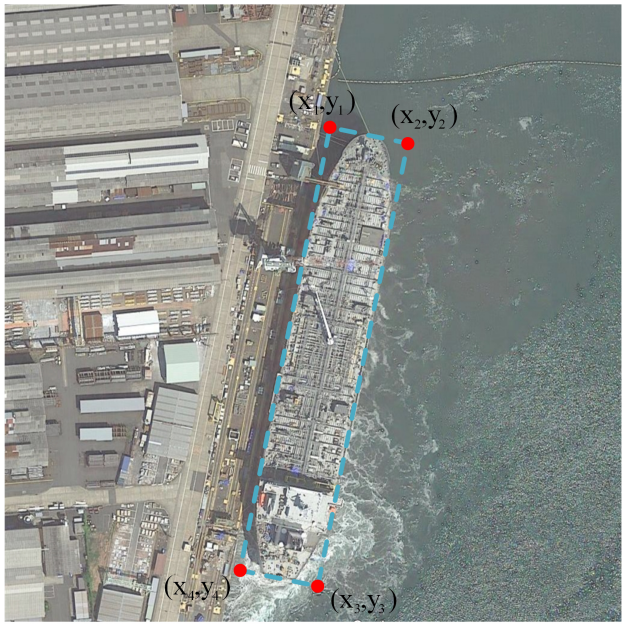


Fig. 6. Sample of the annotation in HPDM-OSOD.

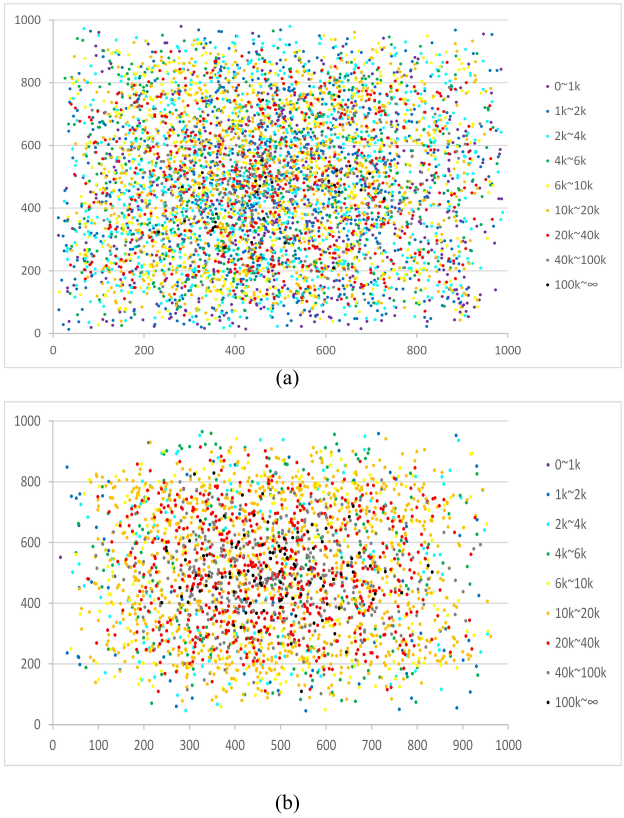
TABLE I
STATISTIC OF THE SCALE DISTRIBUTION IN HRSC2016 AND HPDM-OSOD

Object Scale /pixels	HRSC2016 /pcs	HPDM-OSOD /pcs
$0\sim 1k$	1	**624**
$1k\sim 2k$	95	**844**
$2k\sim 4k$	136	**886**
$4k\sim 6k$	204	**591**
$6k\sim 10k$	319	**798**
$10k\sim 20k$	**1008**	876
$20k\sim 40k$	**632**	599
$40k\sim 100k$	**366**	315
$100k\sim \infty$	**107**	31

on the distribution of the object position at different scales in HRSC2016 and HPDM-OSOD, which illustrates the challenge of our HPDM-OSOD on detecting small and dense ship targets.

b) Multiscales: Considering the large aspect ratio of the ship, horizontal bounding boxes failed to measure rotated ships with such extreme aspect ratios. We use the number of pixels of the rotated bounding boxes as a measurement for the size of the instance. According to the consensus of the industry on the object scale [11], we properly divide and statistic the scale distribution in HRSC2016 and HPDM-OSOD, respectively, as shown in Table I and Fig. 8. It is clear that HPDM-OSOD keeps a balance between instances with different scales and makes the scenes more practical, which is of great significance to the multiscale object detection in reality.

c) Multipostures: In aerial images, the aspect ratio is vital for an anchor-based detection algorithm. Meanwhile, a large range of aspect ratios is also a challenge for the improvement of model adaptability. Besides, the rotation angle is another criterion for rotated object detection. Extensive rotation angles further increase the complexity and diversity of the data set. We statistic the postures, especially the rotation angle of the



(a)

(b)

Fig. 7. Distribution of the object position at different scales in HRSC2016 and HPDM-OSOD. The X- and Y-axes denote the relative positions of the objects, which is normalized to 1000 $\times$ 1000 pixels. Different colors indicate objects in different scales. (a) Statistics of HPDM-OSOD. (b) Statistics of HRSC2016.

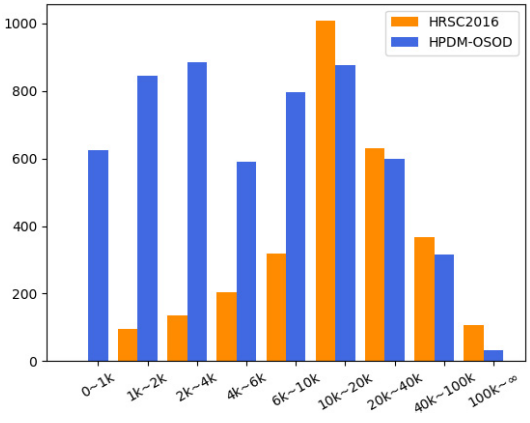


Fig. 8. Scale distribution in HRSC2016 and HPDM-OSOD.

ship in HRSC2016 and HPDM-OSOD, as shown in Table II and Fig. 9. Table II indicates that the angle distribution in HRSC2016 is more uniform due to the artificial rotation angle of the image. However, HPDM-OSOD is collected in the direction of longitude and latitude, which reveals the distribution of ships in the ocean all over the world. In order to make up for the imbalance in the direction, we can easily make it more generalized by randomly rotating the image.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE II

STATISTIC OF THE ROTATION ANGLE DISTRIBUTION IN HRSC2016 AND HPDM-OSOD

Rotation Angle /rad	HRSC2016 /pcs	HPDM-OSOD /pcs
$0 \sim \pi/8$	380	**1301**
$\pi/8 \sim \pi/4$	377	**1028**
$\pi/4 \sim 3\pi/8$	353	**394**
$3\pi/8 \sim \pi/2$	**405**	230
$\pi/2 \sim 5\pi/8$	**430**	175
$5\pi/8 \sim 3\pi/4$	**379**	365
$3\pi/4 \sim 7\pi/8$	302	**1030**
$7\pi/8 \sim \pi$	242	**1041**

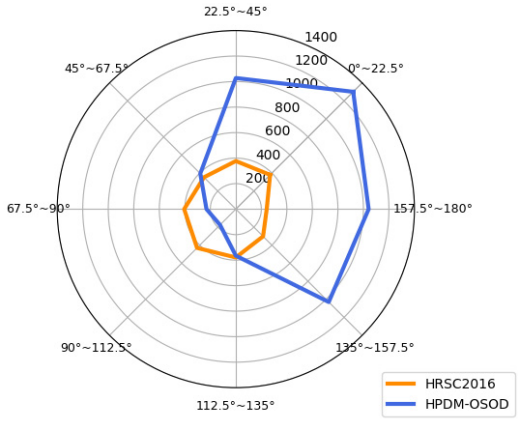


Fig. 9. Rotation angle distribution in HRSC2016 and HPDM-OSOD.

TABLE III

CONFUSION MATRIX

Confusion Matrix		Ground Truth	
		Positive	Negative
Prediction	Positive	TP	FP
	Negative	FN	TN

B. Evaluations and Criteria

Two measurements, average precision (AP) and recall, are adopted for universal performance comparison. Table III shows the confusion matrix for the evaluation of the detection algorithms. In the experiment, three evaluation indicators are used for performance evaluation in various scales and different overlap with the ground-truth boxes. The calculation of precision and recall are shown in (11) and (12). Especially, as the most popular criterion in detection study, AP is used to measure the algorithm comprehensively, which is calculated as in (13), where $P(k)$ and $r(k)$ denote the detection precision and its recall in the output

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{12}$$

$$\text{AP} = \sum_{k=1}^{N} P(k)\Delta r(k). \tag{13}$$

C. Results on HRSC2016

We apply the proposed SkNet on HRSC2016 and compare it with several popular algorithms, including the state-of-the-art detectors. The experimental results are shown in Table IV, which shows that based on the single scale (511×511) SKNet reached 88.3% in AP and 10 frames/s in inference speed. Compared with the state-of-the-art algorithms, SKNet doubled the speed ($6 \rightarrow 10$ frames/s) while further improved the AP.

Multiscale verification is one of the important techniques to improve the detection effect and is widely used in state-of-the-art models [16], [20], [44], which improves the robustness for detecting objects of various scales by inputting larger or smaller images. We conducted the validation at two scales (0.6, 1), which is denoted as SKNet* in Table IV, to further verify the effectiveness of the proposed SKNet. The results show that we traded less efficiency ($10 \rightarrow 8$ fps) for the state-of-the-art performance (reached 90.4% on AP) in anchor-free methods. Within the anchor-free detectors, SKNet not only outperforms the other algorithms in AP but also reached the average level in speed. Besides, the floating-point operations (FLOPs) of the backbone networks are measured in different networks. Despite the large number of FLOPs required by SKNet ($\sim 1.3 \times 10^{11}$ FLOPs), the Hourglass network has a higher degree of parallelism in computation than other deeper backbone networks [45]. Moreover, the concise postprocessing of SKNet improves the general detection effect significantly.

Compared with anchor-based methods, SKNet still has relatively high precision. Several anchor-based algorithms [46]–[48] are slightly ahead of SKNet in AP by setting dense multipose anchors, which are set based on the prior knowledge of a specific data set in order to achieve the best detection effect on a particular data set for verification while increasing the computational complexity during prediction. However, the proposed SKNet avoids the setting of anchors, strengthening the generalization ability of the model, and is more efficient (8–10 fps), which promotes real-time performance.

Overall, it is evident that SKNet outperforms the state-of-the-art detection algorithm in terms of accuracy and time-efficiency.

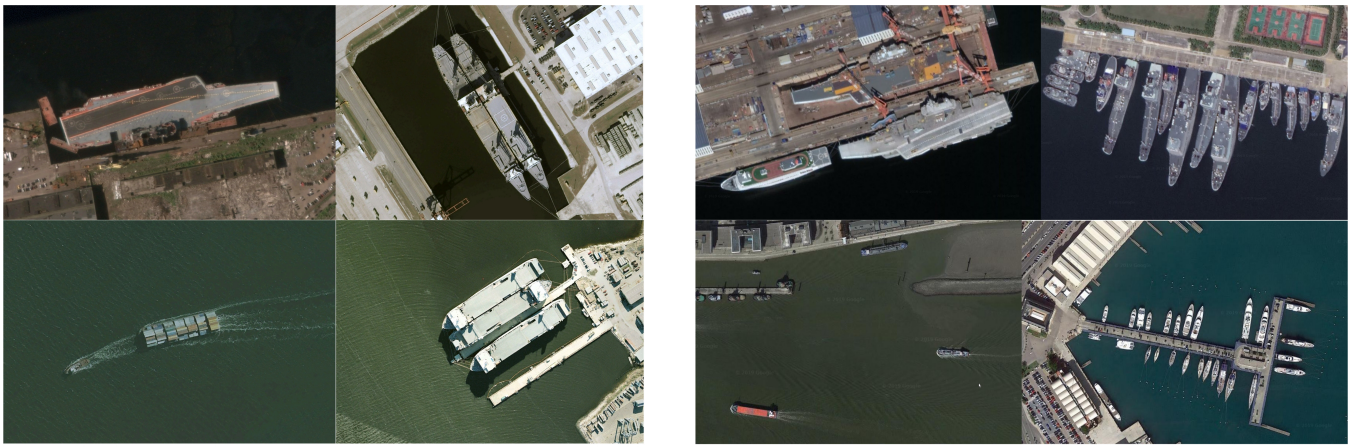
D. Results on DOTA-Ship

We performed the same comparison with more methods on DOTA-ship with larger images and more ship targets. The experimental result in Table V shows that the proposed SKNet also achieves the best performance when performing single-scale validation (reached 82.5% on AP). When performing the multiscale validation, SKNet achieves 83.9% on AP, a competitive result among existing algorithms. We observed that the anchor-based detectors perform better than the anchor-free detectors. Theoretically, the concentrated ship scale distribution (about 75% of the instances are small targets in DOTA-ship) helps to set up overfitting anchors based on statistics. We argue that the resulting improvement cannot fully support the robustness of the algorithm. In comparison, the proposed SKNet has achieved better performance while avoiding the settings of hyperparameters.

In all, the experimental results on public data sets verify the robustness and effectiveness of the proposed SKNet.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.



(a) (b)

Fig. 10. Sample images in (a) HRSC2016 and (b) HPDM-OSOD.

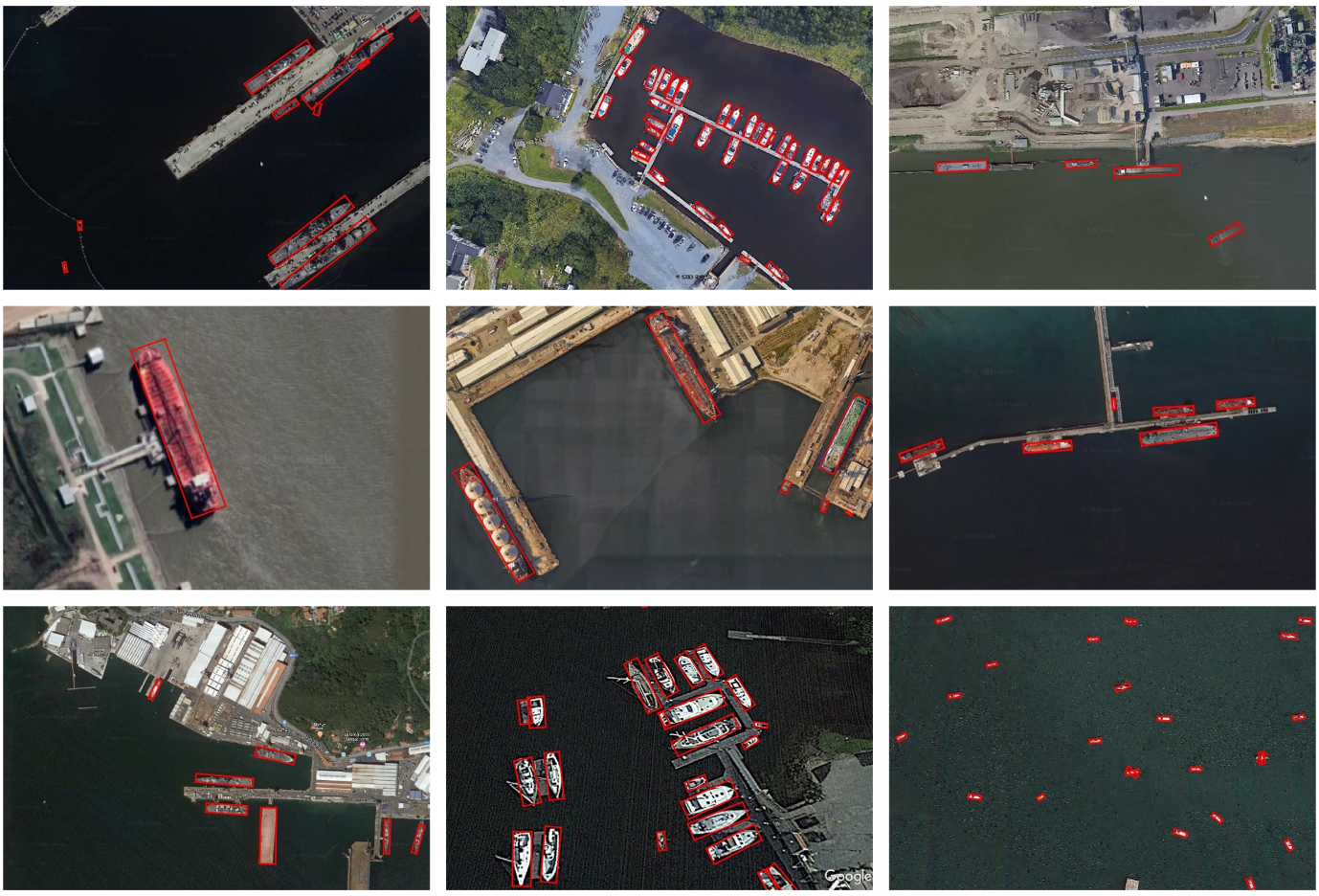


Fig. 11. Visualization of the detection results of SKNet on the HPDM-OSOD data set.

E. Results on HPDM-OSOD

To further evaluate the proposed method, we apply SKNet in a more challenging data set HPDM-OSOD. Fig. 11 shows a visualization of the detection results of SKNet, which illustrates the effectiveness of the proposed structure in detecting objects with multiscales and multipostures in complex scenes. We compared SKNet with three popular anchor-free detectors. For Cornernet [16] and CenterNet-Triple [33], we enrich two corners into four corners to describe the rotated bounding box with top-left, top-right, bottom-left, and bottom-right keypoints. When grouping, the four embeddings of the corners are predicted and grouped in the same manner with the original threshold (0.5). For CenterNet [34], a branch for rotation angle

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE IV
AP OF DIFFERENT METHODS ON THE HRSC2016 DATA SET

Method	Data Aug	Backbone	Image height	Image width	FLOPs of Backbone	AP	Speed
Anchor-based Methods							
R^2CNN[29]	×	ResNet101 [45]	800	800	$\sim 9.6 \times 10^{10}$	73.07	2fps
RC1&RC2[17]	–	VGG16 [49]	–	–	–	75.7	<1fps
RRPN[13]	×	ResNet101	800	800	$\sim 9.6 \times 10^{10}$	79.08	3.5fps
R^2PN[50]	√	VGG16	–	–	–	79.6	<1fps
RRD[51]	–	VGG16	384	384	$\sim 4.5 \times 10^{10}$	84.3	slow
RoI-Transformer [1]	×	ResNet101	512	800	$\sim 6.1 \times 10^{10}$	86.2	6fps
Xu2020 [52]	√	ResNet101	512	800	$\sim 6.1 \times 10^{10}$	88.2	-
R^3Det[5]	√	ResNet101	800	800	$\sim 9.6 \times 10^{10}$	89.26	12fps
Liu2020 [53]	√	ResNet101	512	512	$\sim 4 \times 10^{10}$	90.3	-
RBox-CNN [46]	√	ResNet101	600	600	$\sim 5.4 \times 10^{10}$	91.9	-
Wang2020 [47]	√	ResNet101(FPN)	1024	1024	$> 1.6 \times 10^{11}$	92.8	**6**fps
Liu2019 [48]	√	ResNet50	1200	1600	$\sim 1.5 \times 10^{11}$	**93.7**	slow
Anchor-free Methods							
RetinaNet – H	√	ResNet101	800	800	$\sim 9.6 \times 10^{10}$	82.89	14fps
RetinaNet – R	√	ResNet101	800	800	$\sim 9.6 \times 10^{10}$	89.18	10fps
IENet[54]	√	ResNet101	1024	1024	$\sim 1.6 \times 10^{11}$	75.01	9fps
Yang2020 [55]	√	ResNet152	–	–	–	89.6	-
SKNet(proposed)	√	Hourglass-104	511	511	$\sim 1.3 \times 10^{11}$	88.3	10fps
SKNet*(proposed)	√	Hourglass-104	511	511	$\sim 1.3 \times 10^{11}$	**90.4**	**8**fps

TABLE V
AP OF DIFFERENT METHODS ON THE DOTA-SHIP DATA SET

Method	Backbone	Advances	AugData	AP
RetinaNet – H	ResNet50	×	×	63.6
RetinaNet – R	ResNet50	×	×	75.2
R^2CNN	VGG16	×	√	55.8
RRPN	VGG16	×	√	57.3
ICN	ResNet101	DIN	√	70.0
SCRDet	ResNet101	SF+MDA	√	72.4
R^3Det	ResNet101	FRM	√	77.5
RoI – Transformer	ResNet101	RoI-Trans	√	**83.6**
FFA[56]	ResNet101	RPN-O	√	83.5
SKNet	Hg101	OP+SR-NMS	√	82.5
SKNet*	Hg101	OP+SR-NMS	√	**83.9**

prediction is added, which is consistent with other object properties of the center points. The other settings of the networks are consistent with the original [34].

For a more detailed comparison, COCO evaluation is used to measure the above methods, which compares the AP and recall of the objects at different scales under different IOU threshold. The comparative results are shown in Table VI.

Table VI shows that SKNet achieved the highest AP at 84.6%, 43.9%, and 44.5% when IOU = 0.5, 0.75, and 0.5 ∼ 0.95. In the case of IOU = 0.5, we found that CornerNet and CenterNet-Triple did not achieve ideal results, which was attributed to a general problem as follows. Both the above algorithms predict multiple keypoints on the boundary of each target and then group them into integral detections. As known to all, three key points are necessary to be able to reconstruct a rotated detection box. The lack of any keypoint will reduce the detection performance. In addition, both the algorithms use a value of 0 ∼ 1 to indicate the matching degree of the keypoint, which is prone to mismatch. Fortunately, since both algorithms predict the keypoints of the boundary, the overlap with the ground-truth bounding boxes is extremely tight once they are combined into an integral, which smooths the performance under stricter IOUs (0.75 and 0.5 ∼ 0.95).

Overall, the proposed SKNet has higher effectiveness than the other popular anchor-free detectors.

F. Ablation Study

We perform ablation studies to further identify the proposed contributions in detail over the HPDM-OSOD data set for a more exhaustive evaluation.

1) Backbone Network: Different backbone networks usually bring diverse results in distinct tasks or structures. Therefore, we choose several classical CNN structures as the backbone to explore the impact of different backbone networks on SKNet. By predicting the offset of the position of keypoints, SKNet can be adapted to the change in the size of the feature map output by any backbone, which supports the replacement of the backbone. Meanwhile, we use the authoritative pretrained models trained in public data sets to initialize these networks. The data augmentation and training settings are consistent, which includes random flipping, random cropping, and random zooming. The effect of different backbones (VGG16, ResNet101, DLA-34, Hourglass-52, and Hourglass-104) is shown in Table VII. It is clear that the network based on Hourglass-104 achieved the highest AP with 84.6% when the IOU = 0.5.

In the crosswise comparison, the result shows that, compared to the deeper backbone networks (VGG16 and ResNet101), the wider network structure (DLA-34 and Hourglass) is more effective for ship detection based on keypoints in SKNet, which achieves the AP for more than 80%. On the comparison under stricter IOUs (0.5, 0.75, 0.5 ∼ 0.95), it also demonstrates the comprehensive leadership of the Hourglass network, which indicates that its structure is more suitable for the keypoint-based detection. In addition, the experimental results also show that Hourglass-104 with two five-order nesting structures is more effective than Hourglass-52 with

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE VI
AP OF DIFFERENT ANCHOR-FREE METHODS ON THE HPDM-OSOD DATA SET UNDER DIFFERENT IOUS

Method	Data Aug	Backbone	Average Precision(AP)		
			IOU=0.50	IOU=0.75	IOU=0.5:0.95
CornerNet [16]	√	Hourglass-104	49.7	39.7	32.5
CenterNet-Triple [33]	√	Hourglass-104	49.9	41.9	33.3
CenterNet [34]	√	Hourglass-104	75.3	42.3	43.6
SKNet(proposed)	√	Hourglass-104	**84.6**	**43.9**	**44.9**

TABLE VII
AP OF DIFFERENT BACKBONES UNDER DIFFERENT IOUS ON SKNET

Backbone	Average Precision(AP)		
	IOU=0.50	IOU=0.75	IOU=0.5:0.95
VGG16 [49]	64.5	38.0	36.4
ResNet101 [45]	78.1	42.2	42.6
DLA-34 [57]	81.1	42.6	43.7
Hourglass-52 [58]	82.4	43.1	44.2
Hourglass-104 [58]	**84.6**	**43.9**	**44.9**

TABLE VIII
EVALUATION OF DIFFERENT REGRESSION METHODS FOR BOUNDING BOX ON SKNET

Evaluation	Object Scale	SKNet-Center	SKNet-Corner
AP (IOU=0.5:0.95)	Small(<1024)	50.9	41.6
	Medium(1024:9216)	45.5	39.6
	Large(>9216)	28.2	19.2
AR (IOU=0.5:0.95)	Small(<1024)	54.0	41.8
	Medium(1024:9216)	56.9	48.9
	Large(>9216)	46.1	52.1
Evaluation	IOU	SKNet-Center	SKNet-Corner
AP	(IOU=0.5:0.95)	**44.9**	34.2
	(IOU=0.5)	**84.6**	52.7
	(IOU=0.75)	**43.9**	41.5

one of them, which indicates the superiority of deeper neural networks.

2) Combination of the Bounding Box: The keypoint-based algorithm can be roughly divided by the number of keypoints to detect for each object. Some algorithms detect a single keypoint for each object and use additional information to restore the detection boxes, while others detect multiple keypoints and then group them to form the bounding boxes. In order to compare the performance of the proposed keypoint-based strategy with other algorithms, we compared the impact of a different number of keypoints on the detection results.

As shown in Table VIII, SKNet with the central keypoint performs better than that with the corner keypoint in objects at different scales in general. However, the AP of SKNet-Center started a precipitous decline when the IOU is stricter. The changes of IOU have little effect on the detectors that detect four corner keypoints before grouping them into integral objects and achieved about 52.7% when IOU $= 0.5$. We believe that the SKNet-Corner forms a more precise bounding by directly grouping the maximum response on the boundary of the object. However, two factors result in an unsatisfactory AP. First, it tends to miss keypoint in multikeypoint detection, which leads to ineffectiveness. For the other, the lack of the embedding of keypoints makes it impossible to group the points in such a similarity space. In addition, we also discovered a phenomenon. Although SKNet-center has clear leadership in AP in objects with different scales, for large objects (scale $>$ 9216), the recall of SKNet-Corner is significantly higher (46.1% $\rightarrow$ 52.1%). It can be inferred that the sufficient visual features of large objects can enrich the abstract visual patterns at the corner. However, small targets are more suitable to collapse into center keypoints and abstract semantic features (height, width, and rotation angle). Although SKNet-Corner failed to achieve a better general detection effect, it still provides inspiration for the follow-up work of the proposed SKNet.

TABLE IX
EVALUATION OF THE EFFECTIVENESS OF DIFFERENT POOLING STRATEGIES IN SKNET ON HRSC2016 AND HPDM-OSOD

Method	AP(HRSC2016)	AP(HPDM-OSOD)
SKNet-MP	86.5	82.7
SKNet-CP	87.8	84.2
SKNet-OP(Add)	87.6	83.9
SKNet-OP(Concat)	**88.3**	**84.6**

3) Orthogonal Pooling Module: The orthogonal pooling is a novel module for the feature extraction of the rotated ships. To further identify and explain its effectiveness, we compared the performance of the network with and without the module. As shown in Table X, the orthogonal pooling module enhances the precision and recall in objects at different scales and increases the AP and recall of small, medium, and large objects significantly compared with the case without the module. The results show that the proposed orthogonal pooling can significantly improve the performance of target detection, especially the recall of large objects (about 4%) and the accuracy of small targets (about 3.7%), which reveals that the proposed orthogonal pooling can not only promote the prediction of keypoints but also improve the accuracy of the estimation of morphological sizes.

Furthermore, in order to explore the rationality of the customized pooling structure for the prediction of the center point and corresponding morphological attributes, we introduced several baseline algorithms to evaluate the orthogonal pooling comprehensively. The three baseline pooling methods in Table IX are set as follows: 1) SKNet-MP represents the SKNet that adopts the max-pooling as a customized pooling method; 2) SKNet-CP represents the framework that adopts center pooling [33]; and 3) for SKNet-OP(Add) and SKNet-OP(Concat), we applied two feature map aggregation strategies, including elementwise addition and concatenation, to compare their influences on SKNet. The experimental

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE X
EVALUATION OF THE EFFECTIVENESS OF THE ORTHOGONAL POOLING MODULE AND THE SOFT-ROTATE-NMS IN SKNET

Orthogonal Pooling	NMS	SR-NMS	Average Precision(AP)(IOU=0.5:0.95)			AR(IOU=0.5:0.95)		
			Small	Medium	Large	Small	Medium	Large
	√		47.2	43.9	27.4	49.4	53.4	40.2
		√	48.6	44.3	27.6	51.3	54.1	41.9
√	√		50.2	45.2	29.1	53.0	55.4	44.9
√		√	**50.9**	**45.5**	**29.3**	**54.0**	**56.9**	**46.1**

results are shown in Table IX. We performed single-scale verification on HRSC2016 and HPDM-OSOD. It shows that SKNet-OP (Concat) achieved the highest AP in the above two data sets (88.3% and 84.6% on HRSC2016 and HPDM-OSOD). The proposed OP improves 1.8%∼1.9% on AP over the max-pooling. Center pooling (CP) [33] is a feature extraction module based on the anchor-free framework, which superimposes target responses in the axis-aligned direction to keypoints. The results in Table IX prove that OP cascaded in multiple directions is more helpful to detect ship targets with various scales and extreme aspect ratios than CP. Furthermore, we compared the impact of different feature fusion strategies (Add and Concat). Experiments show that although the elementwise addition strategy saves the consumption of the feature layer (the channels are reduced from 256 to 128), it destroys the prime features and produces a suboptimal result. In comparison, the concatenation strategy is more conducive to transferring prime features and achieves the highest AP.

Orthogonal pooling is equally important and effective for single or multiple ships presented in the image. When dealing with multiple targets, the uncertainty of the ship candidate will cause a few interferences of features between targets. However, the essential 3×3 Conv-BN layers and ReLU layers (3×3 Conv-BN blocks in Fig. 3) alleviate the interference caused by the feature transfer in the orthogonal direction. The results of Table IX show that the orthogonal pooling finally improves the features that are meaningful for predicting single or multiple targets.

Overall, the proposed orthogonal pooling module strengthens the response of feature maps at specific locations and improves the objects' responses at the center by transmitting the maximum global value of the rotated object into the orthogonal direction.

4) Soft-Rotate-NMS: Eliminating redundant detection boxes has always been an effective means in the inference stage. The proposed soft-rotate-NMS calculates the IOU with the rotation angle to decrease the confidence for each candidate. In order to measure the effectiveness of the proposed method, we perform ablation studies for different NMS methods in SKNet. The diversity between the Rotate-NMS and the soft-rotate-NMS is shown in Table X. Experimental results show that the proposed SR-NMS contributes more to the accuracy than the original NMS, which improves the AP of objects at different scales by about 0.2%–1.4% and the corresponding recall by about 0.7%–1.9%. Considering the rotation angle, the algorithm is more sensitive to ship targets with the fixed and larger ratios at different scales. Overall, soft-rotate-NMS achieved better results than the original.

V. CONCLUSION

In this article, we propose a ship detection framework in optical remote sensing images, called SKNet, which models each ship as a center keypoint and corresponding morphological sizes. In order to further aggregate the global information to the center of the object, we propose a novel module named "orthogonal pooling" to transmit the maximum response information within the candidates. In addition, based on the NMS algorithm, we innovatively adopt soft-rotate-NMS to merge neighboring redundant detections.

We verified the proposed SKNet in three data sets (HRSC2016, DOTA-ship, and HPDM-OSOD). In the experiments, we compared and analyzed the performance of the proposed algorithm with popular object detectors. In HRSC2016, the proposed SKNet achieves 90.4% in AP with 10 frames/s in inference speed. Meanwhile, in the more challenging data set, HPDM-OSOD, SKNet has also achieved SOTA performance, which reaches 84.6% in AP. The ablation experiment further verified the effectiveness of the above modules proposed in SKNet. Empirical studies show that SKNet achieves state-of-the-art detection performance while being time-efficient. In addition, inspired by instance segmentation advances [59]–[61], we believe that the instance segmentation of individual ships will be critical to improve the detection effect. Furthermore, it is more powerful to integrate unique features for multiscale ships. In the future, we will carry out more relevant research on these fields.

REFERENCES

[1] 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.

[2] 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.

[3] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," 2018, *arXiv:1805.09512*. [Online]. Available: https://arxiv.org/abs/1805.09512

[4] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "$\mathcal{R}^2$-CNN: Fast tiny object detection in large-scale remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5512–5524, Aug. 2019, doi: 10.1109/TGRS.2019.2899955.

[5] X. Yang, J. Yan, Z. Feng, and T. He, "R3Det: Refined single-stage detector with feature refinement for rotating object," 2019, *arXiv:1908.05612*. [Online]. Available: http://arxiv.org/abs/1908.05612

[6] G. Zhang, S. Lu, and W. Zhang, "CAD-Net: A context-aware detection network for objects in remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10015–10024, Dec. 2019, doi: 10.1109/TGRS.2019.2930982.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[8] 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.

[9] W. Liu *et al.*, "SSD: Single shot MultiBox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[11] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.

[12] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv:1711.09405*. [Online]. Available: https://arxiv.org/abs/1711.09405

[13] J. Ma *et al.*, "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3111–3122, Nov. 2018.

[14] X. Yang *et al.*, "SCRDet: Towards more robust detection for small, cluttered and rotated objects," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2018, pp. 8232–8241.

[15] X. Zhou *et al.*, "EAST: An efficient and accurate scene text detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5551–5560.

[16] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[17] Z. Liu, L. Yuan, L. Weng, and Y. Yang, "A high resolution optical satellite image dataset for ship recognition and some new baselines," in *Proc. 6th Int. Conf. Pattern Recognit. Appl. Methods*, 2017, pp. 324–331.

[18] 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.

[19] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[21] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[22] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6154–6162.

[23] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2965–2974.

[24] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra R-CNN: Towards balanced learning for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 821–830.

[25] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4203–4212.

[26] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection–SNIP," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3578–3587.

[27] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*. [Online]. Available: http://arxiv.org/abs/1701.06659

[28] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," 2019, *arXiv:1903.00621*. [Online]. Available: http://arxiv.org/abs/1903.00621

[29] 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

[30] M. Liao, B. Shi, and X. Bai, "TextBoxes++: A single-shot oriented scene text detector," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3676–3690, Aug. 2018.

[31] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 761–769.

[32] 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.

[33] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6569–6578.

[34] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019, *arXiv:1904.07850*. [Online]. Available: http://arxiv.org/abs/1904.07850

[35] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," 2016, *arXiv:1603.06937*. [Online]. Available: https://arxiv.org/abs/1603.06937

[36] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[37] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," *J. Mach. Learn. Res.*, vol. 15, no. 4, pp. 315–323, 2011.

[38] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, vol. 3, 2006, pp. 850–855.

[39] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS—Improving object detection with one line of code," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5561–5569.

[40] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980

[41] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[42] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.

[43] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0924271614002524

[44] 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.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 630–645.

[46] J. Koo, J. Seo, S. Jeon, J. Choe, and T. Jeon, "RBox-CNN: Rotated bounding box based CNN for ship detection in remote sensing image," in *Proc. 26th ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, Nov. 2018, pp. 420–423.

[47] J. Wang, W. Yang, H.-C. Li, H. Zhang, and G.-S. Xia, "Learning center probability map for detecting objects in aerial images," *IEEE Trans. Geosci. Remote Sens.*, early access, Jul. 28, 2020, doi: 10.1109/TGRS.2020.3010051.

[48] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," 2019, *arXiv:1906.02371*. [Online]. Available: http://arxiv.org/abs/1906.02371

[49] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[50] Z. Zhang, W. Guo, S. Zhu, and W. Yu, "Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1745–1749, Nov. 2018.

[51] M. Liao, Z. Zhu, B. Shi, G. Xia, and X. Bai, "Rotation-sensitive regression for oriented scene text detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, Jun. 2018, pp. 5909–5918.

[52] Y. Xu *et al.*, "Gliding vertex on the horizontal bounding box for multi-oriented object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Feb. 18, 2020, doi: 10.1109/TPAMI.2020.2974745.

[53] L. Liu, Y. Bai, and Y. Li, "Locality-aware rotated ship detection in high-resolution remote sensing imagery based on multi-scale convolutional network," 2020, *arXiv:2007.12326*. [Online]. Available: http://arxiv.org/abs/2007.12326

[54] Y. Lin, P. Feng, and J. Guan, "IENet: Interacting embranchment one stage anchor free detector for orientation aerial object detection," 2019, *arXiv:1912.00969*. [Online]. Available: http://arxiv.org/abs/1912.00969

[55] X. Yang and J. Yan, "Arbitrary-oriented object detection with circular smooth label," 2020, *arXiv:2003.05597*. [Online]. Available: http://arxiv.org/abs/2003.05597

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

[56] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 294–308, Mar. 2020.

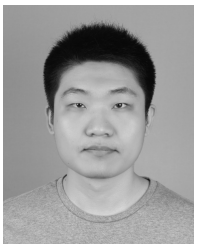
[57] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2403–2412.

[58] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.

[59] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8759–8768.

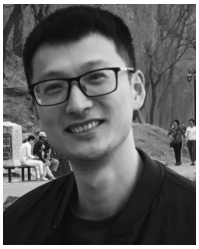
[60] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 1–12.

[61] S. W. Zamir *et al.*, "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2019, pp. 28–37.



Zhenyu Cui (Member, IEEE) received the B.S. degree in computer science and technology from the China University of Petroleum (East China), Qingdao, China, in 2018. He is pursuing the master's degree with the University of Chinese Academy of Sciences, Beijing, China.

His research interests include data mining and computer vision.



Jiaxu Leng received the B.S. degree in electronic information science and technology from Tianshui Normal University, Gansu, China, in 2012, the M.S. degree in electrical engineering from the University of Electronic Science and Technology, Sichuan, China, in 2015, and the Ph.D. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2020.

After obtaining his master's degree, he worked at Hisense, Qingdao, China, for two years as an Algorithm Engineer. He is working at the College of Computer Science and Technology, Chongqing University of Posts and Telecommunications, Chongqing, China. His research interests include computer vision, deep learning, object detection, and stereovision.

Ying Liu (Member, IEEE) received the B.S. degree in computer science from Peking University, Beijing, China, in 1999, and the M.S. and Ph.D. degrees in computer engineering from Northwestern University, Evanston, IL, USA, in 2001 and 2005, respectively.

She is a Professor with the University of Chinese Academy of Sciences, Beijing, the Head of the Lab of Data Mining and High Performance Computing, and an Adjunct Professor with the Key Lab of Big Data Mining and Knowledge Management. Her research interests include data mining, parallel computing, and big data.

Tianlin Zhang received the B.S. degree from the School of Software, Nankai University, Tianjin, China, in 2017, and the M.S. degree in computer science from the University of Chinese Academy of Sciences, Beijing, China, in 2020. He is pursuing the Ph.D. degree in computer science with The University of Manchester, Manchester, U.K.

His research interests include natural language processing and deep learning.

Pei Quan received the B.S. degree in software engineering from Chongqing University, Chongqing, China, in 2017. He is pursuing the Ph.D. degree with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China.

His research interests include graph representation learning and machine learning.

Wei Zhao received the B.S. degree in electrical engineering from the Beijing University of Technology, Beijing, China, in 2016, and the M.S. degree in computer application technology from the University of Chinese Academy of Sciences, Beijing, in 2019.

Her research interests include computer vision, face recognition, and object detection.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on March 08,2021 at 08:52:55 UTC from IEEE Xplore. Restrictions apply.