SC-RPN: A Strong Correlation Learning Framework for Region Proposal

Wenbin Zou, Zhengyu Zhang, Yingqing Peng, Canqun Xiang, Shishun Tian, Lu Zhang

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

SC-RPN: A Strong Correlation Learning Framework for Region Proposal

Wenbin Zou, Zhengyu Zhang, Yingqing Peng, Canqun Xiang, Shishun Tian, and Lu Zhang

***Abstract*—Current state-of-the-art two-stage detectors heavily rely on region proposals to guide the accurate detection for objects. In previous region proposal approaches, the interaction between different functional modules is correlated weakly, which limits or decreases the performance of region proposal approaches. In this paper, we propose a novel two-stage strong correlation learning framework, abbreviated as SC-RPN, which aims to set up stronger relationship among different modules in the region proposal task. Firstly, we propose a Light-weight IoU-Mask branch to predict intersection-over-union (IoU) mask and refine region classification scores as well, it is used to prevent high-quality region proposals from being filtered. Furthermore, a sampling strategy named Size-Aware Dynamic Sampling (SADS) is proposed to ensure sampling consistency between different stages. In addition, point-based representation is exploited to generate region proposals with stronger fitting ability. Without bells and whistles, SC-RPN achieves AR_{1000} 14.5% higher than that of Region Proposal Network (RPN), surpassing all the existing region proposal approaches. We also integrate SC-RPN into Fast R-CNN and Faster R-CNN to test its effectiveness on object detection task, the experimental results achieve a gain of 3.2% and 3.8% in terms of mAP compared to the original ones.**

***Index Terms*—Region proposal, two-stage, strong correlation, SC-RPN.**

I. Introduction

OBJECT detection is one of the most fundamental and challenging tasks in computer vision, which is widely used in surveillance [1], biomedical analysis [2], digital map construction [3], and autonomous driving [4], [5]. As objects can exist at any positions with different scales in a given image, it is exhaustive to directly search everywhere. In order to reduce the searching area and improve the computation efficiency, region proposals are generated as the Region of Interest (ROI) where may contain objects. Modern two-stage object detectors usually begin with a region proposal approach and followed by a R-CNN head. For example, Faster R-CNN [6], in the first stage, generates region proposals as coarse location by using region proposal network (RPN), and then predicts classification and more accurate location based on the region proposals in the second stage. Since the subsequent detection is based on region proposals, the performance of two-stage object detector is largely determined by the quality of region proposals produced by region proposal approach. Thus it is essential to improve the performance of region proposal approaches.

Generally, a region proposal approach consists of several modules, each of which plays its own role. Different combinations of these modules will affect the performance of region proposal approach. In this paper, we focus on the region proposal approaches with two-stage framework, whose location results are predicted in the first stage and refined in the second stage. By revisiting the training process of the pervious two-stage region proposal approaches, we find several weak correlation issues in the existing approaches, which lead to performance degradation. Specifically, these weak correlation issues can be roughly summarized into three categories:

- **The correlation between location and classification of region proposals**. In previous region proposal approaches, the classification score [6] is the unique criterion to measure the location accuracy. Generally, the conventional binary cross entropy loss is applied to train the classification branch, which drives all the positive samples to learn their classification scores as high as possible without considering their location quality. Thus, the training of classification task and location task is independent of each other, which leads to the fact that the classification scores of region proposals cannot reflect their location accuracy correctly. As a result, some candidate region proposals with high location accuracy but low classification score are directly filtered out during Non-Maximum Suppression (NMS), which causes the performance degradation of region proposal.
- **The correlation of sampling strategy between the two stages**. In the training phase, sampling strategy is utilized to select positive samples and calculate losses. The previous two-stage region proposal approaches rarely consider the importance of the correlation of sampling strategy between the two stages. In this paper, this correlation is called "sampling consistency". Here, the sampling inconsistency problems are manifested in three phenomena: (1) The positive samples selected in the first stage will be all inside the ground-truth box of the object, but some positive samples selected in the second stage are outside. (2) The number of the positive samples is stable in the first stage, but it becomes extremely unstable in the second stage. (3) When combined with Feature Pyramid Networks (FPN) [7], the positive samples selected in the

Corresponding author: Shishun Tian.

W. Zou, Z. Zhang, Y. Peng, C. Xiang and S. Tian are with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Institute of Artificial Intelligence and Advanced Communication, College of Electronics and Information Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: wzou@szu.edu.cn; zyzhang23@163.com; 1800261055@email.szu.edu.cn; xiangcanqun2018@email.szu.edu.cn; stian@szu.edu.cn).

L. Zhang is with National Institute of Applied Sciences of Rennes (INSA de Rennes) and Institut d'Électronique et des Technologies du numéRique (IETR), Rennes, France. (lu.ge@insa-rennes.fr)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

first stage are associated with the object size, but in the second stage, the positive samples become unrelated to the object size.

- **The correlation between region proposals and anchor-based representation**. Bounding box representation has been shown effectiveness to describe region proposals. Actually, bounding box representation is a kind of anchor-based representation, which heavily relies on the setting of anchors. Since the preset anchors are the prototype of the predicted objects, higher performance can be obtained when the aspect ratio of the preset anchors are close to that of the predicted objects. Actually, even though several anchors with different aspect ratios are set up to meet as many kinds of object outlines as possible, it is still very difficult to fit all possible object outlines. Limited by the unavoidable shortcoming of anchors, anchor-based representation is inherently difficult to fit the outline of objects efficiently. However, region proposals need to fit the outline of objects well to facilitate the subsequent detection task.

To solve the weak correlation problems mentioned above, we propose a strong correlation learning framework for region proposal, abbreviated as SC-RPN. It mainly contains three novel components: 1) Light-weight IoU-Mask branch; 2) Size-Aware Dynamic Sampling; 3) point-based representation.

In summary, the contributions of this paper can be summarized as follows:

1. Facing the weak correlation issue of classification score and location accuracy in region proposal approach, we propose an extra Light-weight IoU-Mask branch, which predicts IoU mask to rebuild the correlation between classification scores and location accuracy of region proposals with a small amount of additional computation.
2. We delve into the correlation of sampling strategy between different stages in two-stage region proposal approach, and find that it is crucial to ensure the sampling consistency. Thus we propose a Size-Aware Dynamic Sampling (SADS) to establish a stronger correlation between different stages, in terms of location, number and size.
3. We investigate the inherent shortcoming of anchor-based representation in region proposals. To handle this problem, we integrate a point-based representation into region proposal approach to improve the outline fitting ability of region proposal approach.
4. We demonstrate that the proposed SC-RPN, substantially outperforms the state-of-the-art region proposal approaches. The SC-RPN is integrated into different kinds of mainstream object detectors and consistently achieves the best performance, which shows the effectiveness of our proposed framework.

II. Related Work

Over the past few years, object detection has attracted extensive attentions and different variations of object detectors have been proposed, including one-stage object detectors ($e.g.$, [8], [9], [10]) and two-stage object detectors ($e.g.$, [11], [12], [6]). Although the two-stage object detectors have achieved promising performance, recently researchers are still trying to exploit their potential from different aspects. Cao $et\ al.$ [13] reweight the training samples according to a novel ranking method named IoU-HLR. A novel AP-Loss [14] is formulated to replace the classification loss, which alleviates foreground-background class imbalance issue. Chen $et\ al.$ [15] propose a novel PIoU Loss to exploit both the angle and IoU for accurate oriented bounding box regression. Confluence [16] selects optimal bounding boxes and removes highly confluent neighboring bounding boxes according to "Manhattan Distance" instead of the conventional IoU. AugFPN [17] designs a more robust and powerful structure to further exploit the conventional FPN. In this paper, we focus on the first stage of two-stage object detectors, region proposal approaches, whose previous works can be roughly divided into three types: *grouping proposal approaches*, *window scoring proposal approaches* and *CNN-based proposal approaches*.

A. Grouping proposal approaches

In grouping proposal approaches, oversegmentation approaches are first adopted to generate superpixels for an image. Then the similar superpixels are grouped hierarchically with different merging strategies to obtain the proposals. Here, grouping proposal approaches are generally based on diverse low-level cues such as appearance color and superpixel shape. With manually similarity function, Selective Search [18], [19] generates proposals by greedily merging the most similar superpixels. Manen $et\ al.$ [20] innovatively propose a randomized superpixel merging strategy to address all the probabilities. Rantalankila $et\ al.$ [21] utilize novel features that differ from Selective Search, then the generated regions are regarded as seeds to generate more proposals. CPMC [22], [23] directly uses seeds and unaries to cut the graph on pixels, which avoids the initial oversegmentation. A hierarchical segmentation is established from occlusion boundaries in [24], [25], then different seeds and parameters are used to solve graph cuts. The resulting proposals are ranked by a wide range of cues. Rigor [26] uses multiple graph-cuts and fast edge detectors to speed up computing. In Geodesic [27], it starts with generating superpixels, then a set of precomputed geodesic distance transforms are selected as proposals. With multiple segmentation outputs, grouping proposal approaches can produce the proposals with high location accuracy, but they are also time-consuming and computationally expensive.

B. Window scoring proposal approaches

Window scoring proposal approaches firstly initialize a large number of candidate windows with different positions and scales in an image, then generate proposals by scoring and ranking each candidate window according to the probability that they contain objects. Objectness [28], [29] regards the salient locations in the image as candidate proposals, these proposals are then sorted by multiple low-level cues. Rahtu $et\ al.$ [30] initialize a large proposal pool which contains sampling regions and multiple randomly sampling boxes, and adopt a scoring strategy which is similar to Objectness.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

Blaschko *et al*. [31] add more low-level features on the basis of Rahtu to distinguish the quality of proposals in the subsequent scoring work. In Bing [32], a simple linear classifier is trained by edge information and runs as a sliding window approach to find high-scoring proposals. When Bing is applied in video sequences, an extra closed-loop proposal method [33] is proposed to exploit the sequential nature of videos which improves the quality of proposals. EdgeBoxes [34] begins with a coarse pattern of sliding window, then a subsequent refinement is applied to improve location accuracy. In RandomizedSeeds [35], each candidate window is scored by utilizing multiple randomised SEED superpixel maps, without any additional cues. Since this type of proposal approach does not return the segmentation results of proposals, they usually tend to be faster than grouping proposal approaches. However, in window scoring proposal approaches, all the sampling sliding windows are defined by hand-crafted scales and position, which leads to poor location accuracy.

C. CNN-based proposal approaches

As the flourish of convolutional neural network, CNN-based proposal approaches have been developed rapidly thanks to the powerful discrimination feature extracted by convolutional neural network. After obtaining the input feature, these approaches generally predict the coordinates of proposals for each local patch of feature. Multi-Box [36], [37] trains a neural network to produce a certain number of proposals which contains coordinates and scores, all the proposals are then sorted according to their scores. RPN [6] firstly generates the dense candidate proposals with Fully Convolutional Network [38], and then filters out the high overlapping proposals with Non-Maximum suppression (NMS). DeepProposal [39] uses sliding window to search proposals in CNN-based feature and trains a cascade linear classifiers to generate the high score proposals. Scale-aware prediction strategy is proposed in SPOP-net [40] that provides adaptive accurate prediction for objects of different sizes. Based on RPN, a two-stage manner is proposed in [41] (denoted as Iterative RPN in this paper), which refines the scores and location of proposals stage by stage. As a two-stage region proposal approach, GA-RPN [42] highlights the importance of feature alignment and adopts deformable convolution to align the feature before the second stage. Similar to GA-RPN, Cascade RPN [43] continues to focus on the feature alignment and proposes adaptive convolution to obtain better feature alignment. The recent CNN-based proposal approaches usually focus on the rule of feature alignment. Instead, in this paper, we pay closer attention to the weak correlation issues in two-stage region proposal approach, which distinguishes our work from others.

In summary, grouping proposal approaches can obtain high location accuracy, but they are more computationally expensive which results in lower speed, while window scoring proposal approaches are much more time-friendly but poor in location accuracy. Compared to these approaches based on hand-craft features, CNN-based proposal approaches achieve better speed/accuracy trade-offs, thanks to the CNN-based features with strong discrimination. To further improve the performance, a couple of previous CNN-based proposal approaches adopt two-stage framework to refine the result stage by stage. However, the performance of the two-stage region proposal approaches is still limited by the weak correlations between different modules. Therefore, in this paper, we delve into the existing weak correlation issues in two-stage region proposal approach, and some corresponding solutions are proposed to alleviate these issues.

III. The Proposed Framework

We propose a novel region proposal approach abbreviated as SC-RPN, which aims to alleviate the three weak correlation problems discussed above. Fig. 1 shows the overall pipeline of SC-RPN, the detailed process is shown as follows: 1) We firstly utilize FPN-based backbone to generate five feature maps, which are denoted as P_2-P_6, then each feature map is fed into each SC-RPN head. 2) In SC-RPN head, the initial location (denoted as init offset) is predicted from the feature map in the first stage. Then the feature alignment operation is adopted to generate classification feature map and regression feature map. 3) In the second stage of SC-RPN head, classification score is predicted from classification feature map, while the secondary location (denoted as secondary offset) and IoU mask are carried out from regression feature map. After that, the refined offset is obtained by combining init offset and secondary offset, and then the refined offset is converted into refined region proposals. 4) The mean of IoU mask and classification score is calculated as IoU-aware score, according to which the duplicate results in refined region proposals are filtered out in NMS. 5) Finally, an additional NMS is applied after the five SC-RPN heads, which filters out the duplicate region proposals again before the final results are output.

There are three novel components in SC-RPN: Light-weight IoU-Mask branch, Size-Aware Dynamic Sampling (SADS) and point-based representation. Specifically, IoU mask is predicted by the Light-weight IoU-Mask branch and region proposals are described by the point-based representation. In addition, the Size-Aware Dynamic Sampling (SADS) is used for generating the training samples and calculating the losses in the training phase, as shown in Fig. 2. Based on the forward propagation structure in Fig.1, the training phase of SC-RPN is detailed below: 1) Firstly, the ground-truth (abbreviated as GT) and init region proposals are input into SADS. 2) After that, positive samples are output and all of them are regarded as regression samples in the first stage. 3) Then in the second stage, positive and negative samples are output, among which all the positive samples are regarded as regression samples in the second stage, while all the positive samples and the negative samples after random sampling are regarded as classification samples. 4) Finally, the difference between GT and the four predictions, including init region proposals, refined region proposals, IoU mask and classification score are computed as shown in the formula (4)-(7). All the proposed components of SC-RPN are detailed in the following sections.

A. Light-weight IoU-Mask Branch

As mentioned above, regarding classification score as the unique criterion to measure the location accuracy is subop-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

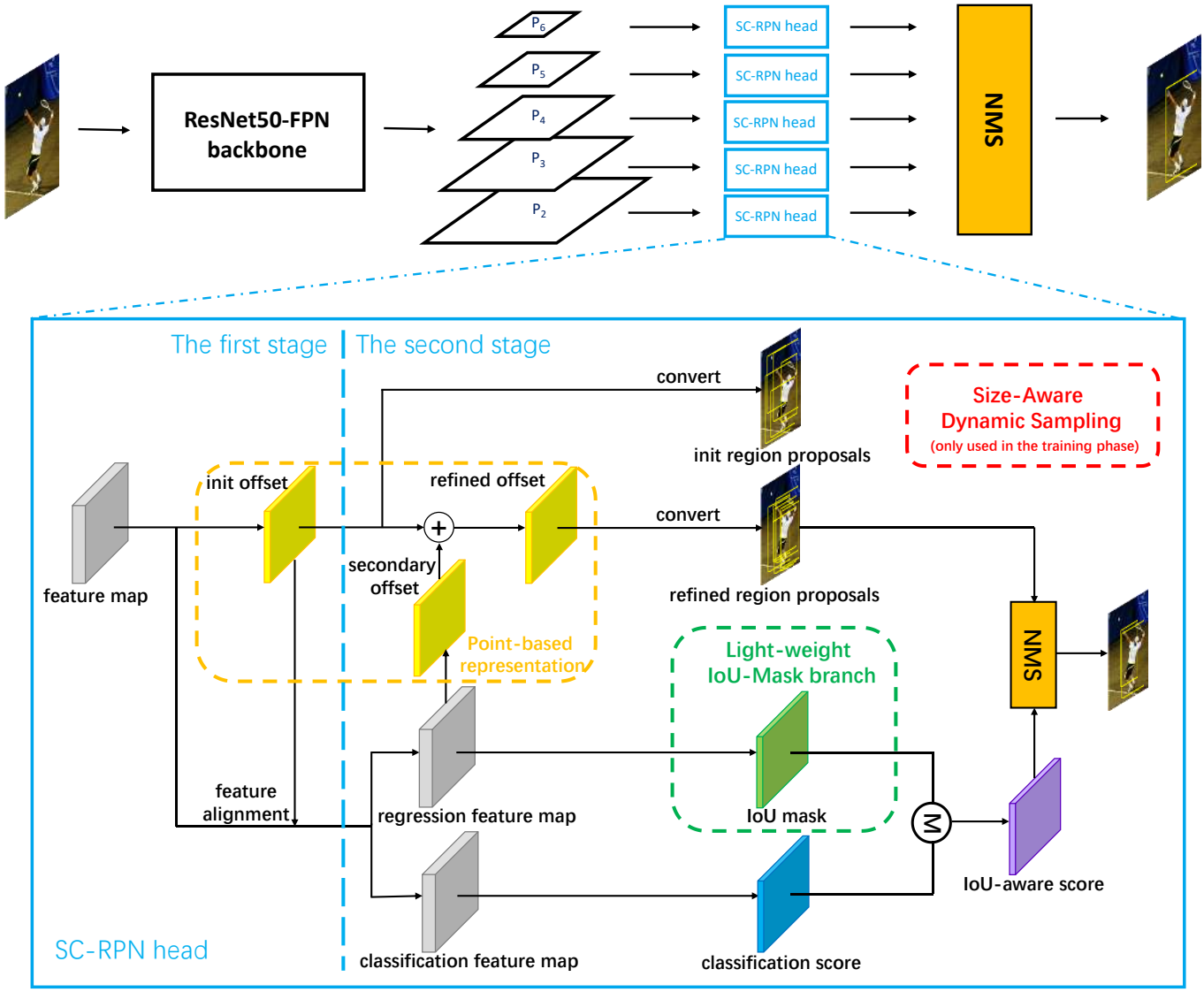


Fig. 1. The overall pipeline of SC-RPN. IoU mask is predicted by the Light-weight IoU-Mask branch and region proposals are described by the point-based representation. In addition, Size-Aware Dynamic Sampling (SADS) is used for generating the training samples and calculating the losses, which is only used in the training phase.

timal. Thus we need a novel criterion to establish a strong correlation between location accuracy and classification score. In previous work, IoU-Net [44] directly predicts an exact IoU between ground-truth box and bounding box as location score, and then replaces the conventional NMS with IoU-guided NMS. However, predicting an exact IoU requires several fully-connected layers in IoU-Net, which is too computationally expensive in region proposal approach. In addition, aiming to prevent the high location accuracy but low classification score region proposals from being removed in NMS, IoU-Net focuses on applying a new location score, but ignores the importance of classification score itself. In this work, instead of predicting an exact IoU for each region proposal, we directly refine the suboptimal distribution of classification score map to correct the wrong classification score of high-quality region proposals. Therefore, we propose a Light-weight IoU-Mask branch to predict IoU mask, which is the distribution map of IoU between ground-truth boxes and region proposals. In the testing phase, we calculate the mean of IoU mask and classification score as IoU-aware score, which is regarded as a novel criterion to measure the location accuracy.

Light-weight Design. Light-weight IoU-Mask branch is designed to be parallel to the secondary regression branch. Compared to the complex structure of IoU-Net, Light-weight IoU-Mask branch consists of a 1×1 convolution layer and two sigmoid function layers, which is extremely computational friendly for region proposal approach. With a 1×1 convolution layer, Light-weight IoU-Mask branch predicts IoU mask, the distribution map of location score. Then the following sigmoid function layer can keep the IoU mask to the range of (0,1). Here, IoU mask is used to refine classification score and generate IoU-aware score. Since we do not want to change the distribution of classification score dramatically, we use the sigmoid function layer twice to get a smaller distribution

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

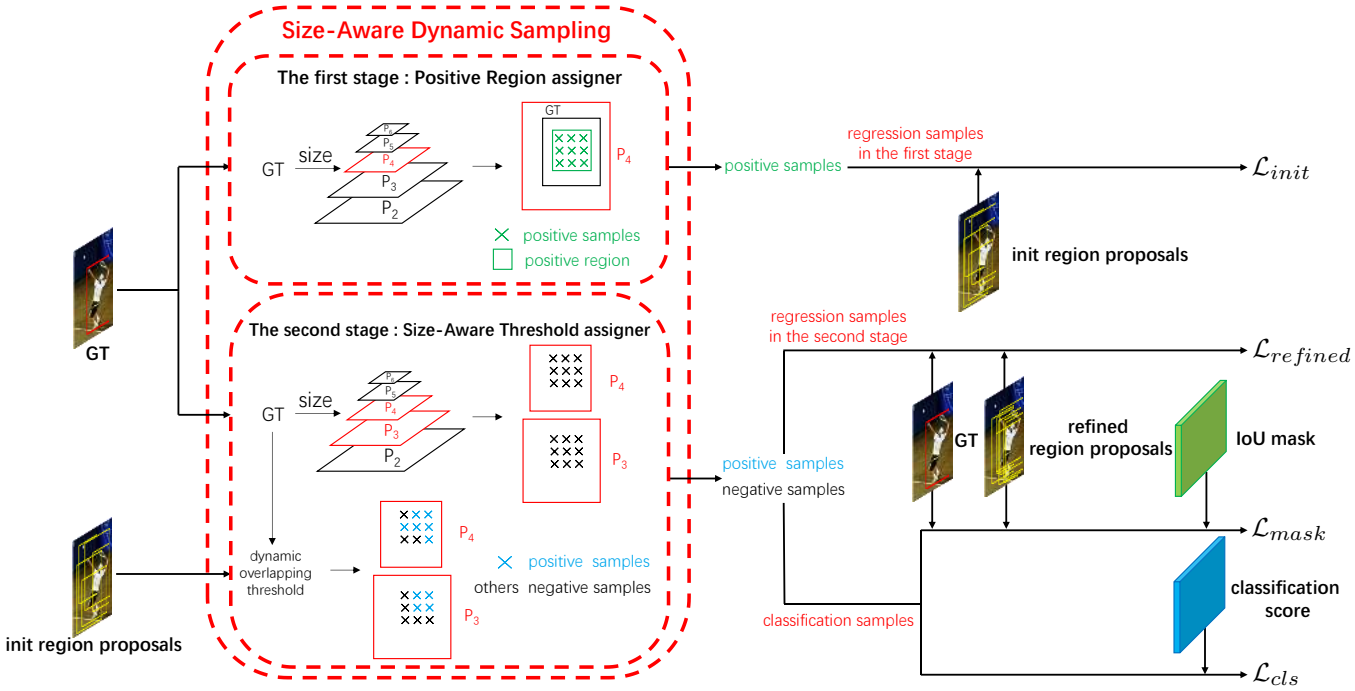


Fig. 2. The training phase of SC-RPN with Size-Aware Dynamic Sampling (SADS). In the first stage, positive samples are output and all of them are regarded as regression samples. Then in the second stage, positive and negative samples are output, among which all the positive samples are regarded as regression samples, while all the positive samples and the negative samples after random sampling are regarded as classification samples.

variances map of IoU mask.

Weak Supervision Information. In order to predict the distribution of IoU between ground-truth boxes and region proposals, we need the ground-truth of IoU distribution which is not provided in the original supervision information. Without introducing additional supervision information, we utilize the IoU between GT and the predicted refined region proposals as supervision information for training Light-weight IoU-Mask branch. In back-propagation, classification samples selected in the second stage are regarded as the training samples of the Light-weight IoU-Mask branch. We firstly calculate the IoU between GT and refined region proposals, and then the IoU of classification samples are regarded as GT IoU. Finally, the conventional binary cross entropy loss is adopted to calculate the point-wise distance which denoted as $\mathcal{L}_{mask}$. The generation of supervision information and the training phase of Light-weight IoU-Mask branch are shown in Fig. 3.

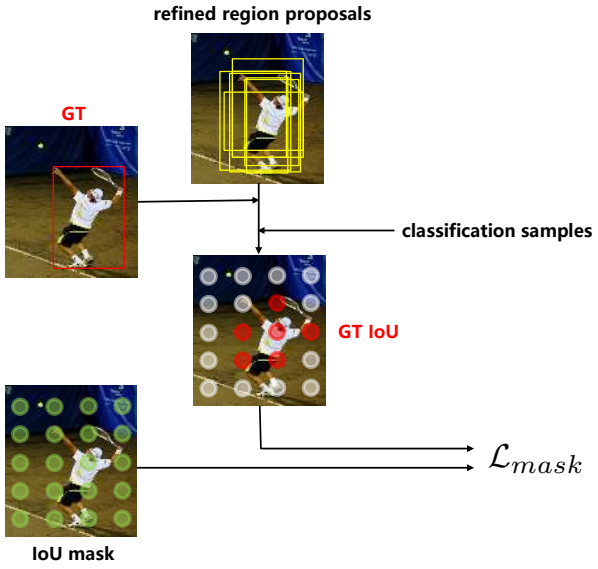


Fig. 3. The generation of supervision information and the training phase of Light-weight IoU-Mask branch. As a result, red points in ground-truth IoU map denote the ground-truth of training samples, while white points denote the irrelevant samples.

Why IoU mask Works. As mentioned above, classification score should not be the unique criterion to measure the location accuracy because classification score cannot correctly reflect the location accuracy of region proposals. Now, Light-weight IoU-Mask branch predicts the distribution map of IoU between ground-truth boxes and region proposals to refine the classification score. In the training phase, the proposed branch is supervised by location information. In the testing phase, the IoU-aware score is generated from the mean of classification score and IoU mask. In this way, the IoU-aware score can better reflect the location accuracy of region proposals. When multiple region proposals are put into NMS, the region proposal with the highest classification score will be retained first, and then it filters out other region proposals with high overlap. As shown in Fig. 4, without IoU mask, region proposal A with high classification score but low location accuracy will be produced which suppresses those with low classification score but high location accuracy, thus the suboptimal region proposal A will be retained. With IoU

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

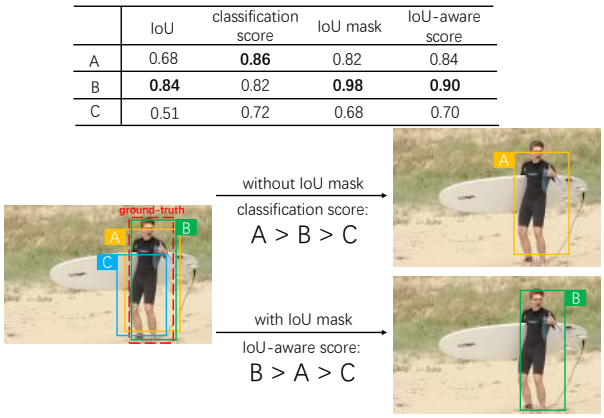


Fig. 4. With IoU mask, the IoU-aware score can better reflect the location accuracy of region proposals. It prevents the high-quality region proposals from being removed, thus the best region proposal B will be retained instead of region proposal A. (Best viewed in color.)

mask, the classification score will be refined according to IoU mask, and the IoU-aware score can better reflect the location accuracy. In this situation, the best region proposal B will be retained.

B. Size-Aware Dynamic Sampling

Generally, in two-stage region proposal approach, the problems of sampling inconsistency between the two stages are manifested in location, number and size level, as shown in the following three phenomena:

1. In the first stage, since all the candidate samples are uniformly distributed, the selected positive samples will be all inside the ground-truth box of the object. After adjusting the location in the first stage, all the candidate samples are no longer uniformly distributed. As a result, some positive samples selected in the second stage are outside the ground-truth box of the object. This phenomenon is more common in objects with slender outlines, such as toothbrush and knife.

2. The number of positive samples is stable in the first stage, but it becomes extremely unstable in the second stage. The instability in the second stage is mainly caused by two factors: (1) Some easily identifiable objects may have a large number of positive samples, while objects that are difficult to identify may have very few positive samples. (2) High threshold leads to the shortage of positive samples in the early training, while low threshold results in an excessive number of positive samples in the late training.

3. When combined with Feature Pyramid Networks (FPN), positive samples are only selected from the appropriate feature map corresponding to the object size in the first stage, which reduces the learning pressure of each feature map. However, the positive samples selected in the second stage are completely unrelated to the object size and their corresponding feature map. These size-independent positive samples will conflict with the size-aware samples selected in the first stage.

Here, the above unreasonable sampling results mainly lie on the commonly used sampling strategy, Max IoU assigner, which is used in both the two stages. In Max IoU assigner, given a hand-craft overlapping threshold, the sample whose IoU with ground truth box higher than the threshold is regarded as positive sample, otherwise it is regarded as negative sample. Therefore, in order to address the suboptimal sampling strategy and ensure the sampling consistency, we propose Size-Aware Dynamic Sampling (SADS), a simple and efficient sampling strategy to select training samples. Specifically, it can be divided into two sampling strategies: Positive Region assigner in the first stage and Size-Aware Threshold assigner in the second stage, which are shown in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Positive Region assigner

Input:
$\mathcal{G}$: all ground-truth boxes on a certain image
$\mathcal{F}^i$: the i_{th} feature map
σ: the center ratio of positive region
Output:
$\mathcal{P}_1$: a set of positive samples in the first stage

for *each ground-truth box* $g \in \mathcal{G}$ **do**
  $\mathcal{S}_g \leftarrow$ *calculate the area of* g
  $i_g \leftarrow$ *assign positive feature map number based on* $\mathcal{S}_g$
  $\mathcal{R}_g \leftarrow$ *assign positive region on* $\mathcal{F}^{i_g}$ *based on* σ
  $\mathcal{P}_g \leftarrow$ *assign samples inside* $\mathcal{R}_g$ *as positive samples*
  $\mathcal{P}_1 = \mathcal{P}_1 \cup \mathcal{P}_g$
end for
return $\mathcal{P}_1$

Algorithm 2 Size-Aware Threshold assigner

Input:
$\mathcal{G}$: all ground-truth boxes on a certain image
$\mathcal{F}^i$: the i_{th} feature map
$\mathcal{T}$: all training samples
$\mathcal{T}^i$: all training samples on $\mathcal{F}^i$
k: the closest samples number selected per feature map
Output:
$\mathcal{P}_2$: a set of positive samples in the second stage
$\mathcal{N}_2$: a set of negative samples in the second stage

for *each ground-truth box* $g \in \mathcal{G}$ **do**
  $\mathcal{S}_g \leftarrow$ *calculate the area of* g
  $i_g \leftarrow$ *assign positive feature map number based on* $\mathcal{S}_g$
  $j_g \leftarrow$ *assign adjacent positive feature map number based on* $\mathcal{S}_g$ *and* i_g
  $\mathcal{D}_g \leftarrow$ *select* k *samples closest to* g *from both* $\mathcal{T}^{i_g}$ *and* $\mathcal{T}^{j_g}$ *based on L2 distance*
  $m_g \leftarrow$ *calculate the mean IoU of* $\mathcal{D}_g$
  for *each sample* $d \in \mathcal{D}_g$ **do**
    if *IoU of* $d > m_g$ *and the center of* d *in* g **then**
      $\mathcal{P}_2 = \mathcal{P}_2 \cup d$
    end if
  end for
end for
$\mathcal{N}_2 = \mathcal{T} - \mathcal{P}_2$
return $\mathcal{P}_2$,$\mathcal{N}_2$

How SADS works. In the first stage, we apply Positive Region assigner, an anchor-free sampling strategy, to select

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

positive samples. Negative samples are not necessary because we only predict init offset in the first stage. Since there may be multiple ground-truth boxes on an image, we assign positive samples to each ground-truth box of each image. Concretely, for each ground-truth box g on an image, we firstly calculate area $\mathcal{S}_g$, and then get the corresponding positive feature map number i_g according to the following formula:

$$i_g = \left\lfloor \frac{\log_2 \mathcal{S}_g - 9}{2} \right\rfloor \tag{1}$$

as described in Line 4 to 5 of Algorithm 1, given a preset center ratio σ in [0,1], we regard the center region of ground-truth box as the positive region of $\mathcal{F}^{i_g}$, and the size of the center region is controlled by center ratio σ. The training samples inside $\mathcal{F}^{i_g}$ are recognized as positive samples in the first stage.

In the second stage, Size-Aware Threshold assigner is adopted to select positive and negative samples. Firstly, we calculate the area $\mathcal{S}_g$ and the corresponding positive feature map number i_g for each ground-truth box g as we did in the first stage. If g is in moderate size, an extra positive feature map number j_g is assigned for g. Here, feature map $\mathcal{F}^{j_g}$ is adjacent to feature map $\mathcal{F}^{i_g}$:

$$j_g = i_g - 1 + \left\lfloor \frac{\mathcal{S}_g}{2^{2\times i_g+9}} \right\rfloor \qquad if\ \mathcal{S}_g \in [\ 2^{10}\ ,\ 2^{18}\ ] \tag{2}$$

For each assigned positive feature maps, we firstly select k samples closest to the center of ground-truth box g based on L2 distance, and then calculate the IoU between their corresponding init region proposal and ground-truth box g. These IoU of selected samples are denoted as $\mathcal{D}_g$. After that, the mean of all the IoUs in $\mathcal{D}_g$ is calculated and denoted as m_g. Here, m_g is regarded as a dynamic overlapping threshold for ground-truth box g. Finally, the sample whose IoU is greater than m_g and located inside the center of ground-truth box g is selected as positive samples. On the contrary, the other samples are considered as negative samples. In addition, the sample with the highest IoU will be selected if a sample is assigned to multiple ground-truth boxes. Combined with Positive Region assigner and Size-Aware Threshold assigner in SC-RPN, positive samples with different sizes can be assigned to different feature maps for better learning. Besides, different dynamic overlapping threshold is assigned to different ground-truth box in different iteration.

Three corresponding solutions for the three sampling inconsistency phenomena. The proposed SADS provides some solutions for the three sampling inconsistency phenomena mentioned above. They are detailed as follows:

1. With Positive Region assigner and uniformly distributed candidate samples in the first stage, all the selected positive samples are guaranteed to be inside the ground-truth box. In the second stage, since all the candidate samples are no longer uniformly distributed, Size-Aware Threshold assigner directly considers samples outside the ground-truth box as negative samples. Therefore, all positive samples are strictly limited inside the ground-truth box, which ensures the sampling consistency throughout the training phase.

2. The instability of the positive sample number in the second stage is mainly caused by two factors: (1) Some easily identifiable objects may have a large number of positive samples, while objects that are difficult to identify may have very few positive samples. (2) High threshold leads to the shortage of positive samples in the early training, while low threshold results in an excessive number of positive samples in the late training. As for the former, Size-Aware Threshold assigner defines a set of candidate samples for each object and then selects positive samples for it, guaranteeing a balance of positive sample number between different objects. For the latter, since the dynamic overlapping threshold can reflect the quality of current candidate samples, the model can maintain a stable number of positive samples throughout the whole training process. Thus SADS can keep the quantity and quality of positive samples more reasonable.

3. When combined with FPN, Positive Region assigner in the first stage maintains the characteristic of size perception. Specifically, each object is assigned a positive feature map number based on their size, and then all positive samples are selected from their corresponding positive feature map. However, this characteristic is often ignored in the second stage. Here, Size-Aware Threshold assigner assigns two corresponding positive feature maps for most objects based on their size, and generates dynamic overlapping threshold to select positive samples. With Size-Aware Dynamic Sampling, size-aware positive samples are produced to ensure the sampling consistency, which improves the training efficiency.

Selecting $\mathcal{D}_g$ with L2 distance instead of positive region in the second stage. To select the sample closest to the center of the ground-truth box, positive region is adopted in the first stage while L2 distance is used in the second stage. Although they are both used to select the closest samples, there is a certain difference between these two methods. Given a fixed center ratio, even if two objects are assigned to the same positive feature map, the difference of area will lead to a large difference in the number of positive samples, which aggravates the phenomenon 2 mentioned above. On the contrary, given a fixed hyper-parameter k to define $\mathcal{D}_g$ for ground-truth box g based on L2 distance, the number of positive samples will be more reasonable and stable since all the positive samples are selected from $\mathcal{D}_g$.

Calculating the mean of size-aware samples from two positive feature maps as dynamic overlapping threshold. In order to ensure the sampling consistency, Size-Aware Threshold assigner selects positive samples which are related to the object size. For each ground-truth box g, since only positive samples of i_g feature map are trained in the first stage, we can easily find that the quality of region proposals predicted from i_g feature map is higher than any other feature maps. In the second stage, if all the feature maps are used to calculate the dynamic overlapping threshold like ATSS [45], the positive samples selected in this situation will no longer be associated with the size-aware positive samples selected in the first stage. On the contrary, if only one feature map is used for calculating, the threshold will be too high, which leads to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

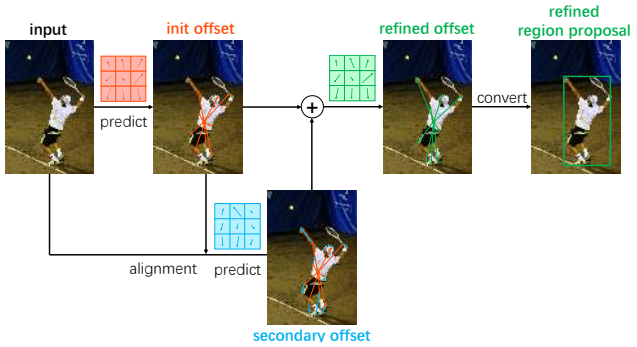


Fig. 5. The entire process of point-based representation. For better visualization, we use the original image instead of the feature map.

the shortage of positive samples. Thus we apply two positive feature maps to calculate the dynamic overlapping threshold, ensuring a reasonable dynamic overlapping threshold and a moderate number of positive samples. Besides, different from calculating the overlapping threshold with both variance and mean, we only calculate the mean of $\mathcal{D}_g$ as the dynamic overlapping threshold in Size-Aware threshold assigner, which is more concise to produce a reasonable threshold for g.

C. Point-based Representation

As discussed above, even if the number of anchors is increased, it is still very difficult to meet all possible object outlines. Furthermore, with the increasing number of anchors, higher accuracy will be obtained, as well as more parameters and computations. However, region proposals need to fit the outline of objects well, which conflicts with the inherent shortcoming of anchors. Hence we draw the conclusion that such a weak correlation between region proposals and anchor-based representation is suboptimal.

Inspired by Reppoint representation [46], we integrate a point-based representation into SC-RPN to describe region proposal instead of anchor-based representation. Specifically, for each feature point on the feature map, we firstly predict init offset that utilizes 9 groups of offset to adaptively find the object boundary. After that, to get a more accurate location, we adopt the point alignment operation and then predict secondary offset. Finally, refined offset is generated by combining init offset and secondary offset, then the maximum and the minimum value of the refined offset are converted into the boundary of the refined region proposal. Different from anchor-based representation, point-based representation automatically finds the boundary point of the object and forms a semantic set of points, which gets rid of the cumbersome anchor setting. Thus, point-based representation can capture the object information in a more detailed way and enhance the outline fitting ability in SC-RPN. The entire process of point-based representation is shown in Fig. 5.

D. Learning

As an end-to-end approach, SC-RPN is trained under the guidance of a standard multi-task objective function. The total loss $\mathcal{L}_{total}$ is generated by adding up the losses of all the branches, which is defined as follow:

$$\begin{aligned}\mathcal{L}_{total} &= \lambda_{cls}\mathcal{L}_{cls} + \lambda_{mask}\mathcal{L}_{mask} \\ &+ \lambda_{init}\mathcal{L}_{init} + \lambda_{refined}\mathcal{L}_{refined},\end{aligned} \tag{3}$$

here, $\mathcal{L}_{cls}$ is the classification loss and $\mathcal{L}_{mask}$ is the loss of Light-weight IoU-Mask branch. $\mathcal{L}_{init}$ and $\mathcal{L}_{refined}$ are the regression loss of the first stage and the second stage, respectively. In addition, λ_{cls}, λ_{mask}, λ_{init} and $\lambda_{refined}$ denote the weight of $\mathcal{L}_{cls}$, $\mathcal{L}_{mask}$, $\mathcal{L}_{init}$ and $\mathcal{L}_{refined}$, respectively. Finally, classification branch and Light-weight IoU-Mask branch are both driven by the conventional binary cross entropy loss (denoted as L_{BCE}), the regression branch of the two stages are both supervised by IoU loss [47] (denoted as L_{IoU}). The loss functions of all branches are defined below:

$$\mathcal{L}_{cls} = \frac{1}{M_{cls}}\Big(\sum_{i\in pos,neg} L_{BCE}(c_i, \hat{c}_i)\Big), \tag{4}$$

$$\mathcal{L}_{mask} = \frac{1}{M_{cls}}\Big(\sum_{i\in pos,neg} L_{BCE}(m_i, \hat{m}_i)\Big), \tag{5}$$

$$\mathcal{L}_{init} = \frac{1}{N_{reg}}\Big(\sum_{i\in pos} L_{IoU}(\mathcal{B}_i, \hat{\mathcal{B}}_i)\Big), \tag{6}$$

$$\mathcal{L}_{refined} = \frac{1}{M_{reg}}\Big(\sum_{i\in pos} L_{IoU}(\mathcal{B}_i, \hat{\mathcal{B}}'_i)\Big), \tag{7}$$

here, N_{reg} is the number of regression samples in the first stage, M_{cls} and M_{reg} are the number of classification and regression samples in the second stage. Furthermore, $\hat{c}$, $\hat{m}$, $\hat{\mathcal{B}}$ and $\hat{\mathcal{B}}'$ denote the predictions of the classification branch, Light-weight IoU-Mask branch, initial regression branch and secondary regression branch, respectively. The predictions with no hat represent their corresponding ground-truth.

IV. Experiment Results

A. Experiment Settings

1) Dataset. Most of the experiments are based on MS COCO 2017 detection dataset [48]. Specifically, both region proposal approaches and object detectors are trained on train split which contains 115k images. The performance of region proposal approaches and ablation experiments are tested on val split which contains 5k images. The performance of object detectors is tested on test-dev split which contains 20k images.

2) Implementation Details. All the region proposal approaches consist of two stages except for RPN. ResNet50-FPN is used as the backbone network. Without changing the aspect ratio, the input images are resized to the scale of 1333×800 for both training and testing. No data augmentation is used except for standard flipping. The center ratio σ is set to 0.2 for selecting positive samples in the first stage. In the multi-task loss function, we assign different weights based on the importance of each loss. Specifically, we use λ_{cls} = 2.0, λ_{init} = 0.5, $\lambda_{refined}$ = 10.0 and λ_{mask} = 1.0 to balance each loss. The NMS post-processing is applied for each head, whose overlapping threshold is set to 0.8. With SGD optimizer, all the detectors are trained with 2 GPUs and a total batch size of 4 for 12 epochs. We use an initial learning rate of 0.005

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

and divide the learning rate by 10 after 8 and 11 epochs. The runtime is measured on GTX 1080Ti GPU. All the code are implemented with mmdetection [49].

3) Evaluation Metrics. The performance of region proposal approach is measured with Average Recall (AR), which is the average of recalls across IoU thresholds (from 0.5 to 0.95 with a step of 0.05). AR_{100}, AR_{300}, and AR_{1000} mean AR for 100, 300, and 1000 region proposals per image. Computed for 100 region proposals, AR for small, medium, and large objects are denoted as AR_S, AR_M, AR_L. Detection results are reported with the averages mAP of IoUs from 0.5 to 0.95 of standard COCO metric.

B. Results

1) Region Proposal Performance. As Table I shows, we compare SC-RPN with the state-of-the-art region proposal approaches, including SharpMask [50], GCN-NS [51], AttractioNet [52], ZIP [53], RPN [6], Iterative RPN [41], GA-RPN [42] and Cascade RPN [43]. The weak correlations limit the performance of the previous region proposal approaches. By using the proposed Light-weight IoU-Mask branch, point-based representation and Size-Aware Dynamic Sampling, the correlations between these modules are greatly enhanced. Without bells and whistles, SC-RPN achieves an improvement of 14.5% in terms of AR_{1000} compared to the RPN. Even under different region proposal numbers and object sizes, SC-RPN consistently outperforms all the existing region proposal approaches.

2) Detection Performance. To further investigate the ability of generating high-quality region proposals and its potential to improve the detection performance, we integrate SC-RPN into two common object detectors, including Fast R-CNN and Faster R-CNN. In Fast R-CNN, the pre-computed region proposals are produced by SC-RPN. While in Faster R-CNN, we utilize SC-RPN to generate region proposals instead of RPN in the first stage and train the whole model end to end. The previous works have proved that aiming to train a detector successfully, several adjustments should be made when replacing RPN with other high-quality region proposal approaches. Therefore, following [42], the overlapping threshold in R-CNN is set to 0.65 and the region proposal number is set to 300. The detection performance are reported in Table II. Besides, the detection performance of Iterative RPN, GA-RPN, Cascade RPN are cited from the previous paper [43]. With RPN, Fast R-CNN yields 37.0 mAP while Faster R-CNN yields 37.1 mAP. However, integrating SC-RPN into Fast R-CNN and Faster R-CNN can boost the performance to 40.2 mAP and 40.9 mAP, respectively. Here, it can be found that the improvement of SC-RPN for Faster R-CNN is greater than that of Fast R-CNN, which shows that an end-to-end manner can better exploit the potential of SC-RPN and achieve higher detection performance.

C. Ablation Studies

Component-wise Performance. In order to demonstrate the effectiveness of each component of SC-RPN, we show an overall component-wise performance in Table III, which omits different components progressively. Firstly, RPN with three preset anchors is regarded as baseline, yielding AR_{1000} of 58.3. Then the performance falls to 55.8 when we use only one anchor, implying that the model cannot fit the object outline well. Even after predicting location twice, the AR_{1000} is still similar to the baseline. After that, we apply the conventional alignment operation and IoU loss, the performance improves to 65.7 and 66.3, respectively. The incorporation of point-based representation increases the AR_{1000} to 66.7, implying that abandoning the use of anchors can produce more high-quality region proposals. When the Size-Aware Dynamic Sampling is added, the AR_{1000} incrementally surges to 72.4, showing the effectiveness of ensuring the sampling consistency. Finally, applying IOU mask gets the AR_{1000} of 72.8, indicating that IoU-aware score can reflect the location accuracy correctly. Overall, SC-RPN achieves 17.1%, 15.5%, and 14.5% higher than that of RPN in terms of AR_{100}, AR_{300}, and AR_{1000}.

1) Ablation Studies on Size-Aware Dynamic Sampling (SADS). Table IV shows the experiment results of different sampling strategy combinations in two stages. The proposed Size-Aware Dynamic Sampling is made up of Positive Region assigner in the first stage and Size-Aware Threshold assigner in the second stage. Here, PosR and SAT denote Positive Region assigner and Size-Aware Threshold assigner, respectively. When Max IoU assigner is applied to both the first and the second stage, the model yields the AR_{1000} of 66.7. The incorporation of SAT in the second stage can surge the performance to 69.6, indicating the importance of ensuring sampling consistency. Furthermore, when PosR is added in the first stage, the AR_{1000} increases to 72.4, implying that selecting positive samples with the preset anchors will limit the model performance.

The performance dependency of different hyper-parameters in SADS is shown in Table V. At the top of the Table V, firstly, we evaluate the effectiveness of Positive Region assigner with different hyper-parameters σ, which denotes the center ratio of positive region in the first stage. The experiment results show that the performance is similar to each other when σ is set to 0.1, 0.2 or 0.3. After that, the performance decreases with the increase of σ, implying that training the model with a small positive region can achieve a more reliable result.

Then the performance dependency results of different hyper-parameters k in Size-Aware Threshold assigner are reported at the middle of the Table V. Here, k denotes the number of the closest samples selected per feature map in Size-Aware Threshold assigner. As shown in the Table, a small k value results in the shortage of positive samples and performance degradation. On the contrary, a big k value leads to the excessive number of positive samples, which has a negative impact on model performance. Thus, adopting a moderate value, $k = 9$, can maintain a reasonable number of positive samples and obtain the best model performance.

Finally, we explore the effectiveness of different number of feature map N selected in Size-Aware Threshold assigner, and the experiment results are shown at the bottom of the Table V. The results show that when the number of the selected feature maps is insufficient, too few positive samples are assigned for proper training due to the high dynamic overlapping threshold.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

TABLE I
PERFORMANCE OF REGION PROPOSAL APPROACHES ON MS COCO 2017 VAL SPLIT.

Approach	Backbone	AR_{100}	AR_{300}	AR_{1000}	AR_S	AR_M	AR_L	Time(s)
SharpMask [50]	ResNet50	36.4	-	48.2	-	-	-	0.76
GCN-NS [51]	VGG-16(Sync BN)	31.6	-	60.7	-	-	-	0.10
AttractioNet [52]	VGG-16	53.3	-	66.2	31.5	62.2	77.7	4.00
ZIP [53]	BN-inception	53.9	-	67.0	31.9	63.0	78.5	1.13
RPN [6]	ResNet50-FPN	44.6	52.9	58.3	29.5	51.7	61.4	**0.04**
Itertive RPN [41]		48.5	55.4	58.8	32.1	56.9	65.4	0.05
GA-RPN [42]		59.1	65.1	68.5	40.7	68.2	78.4	0.06
Cascade RPN [43]		61.1	67.6	71.7	42.1	69.3	**82.8**	0.06
SC-RPN(ours)		**61.7**	**68.4**	**72.8**	**42.9**	**69.9**	**82.8**	0.07

TABLE II
PERFORMANCE OF OBJECT DETECTORS ON MS COCO 2017 TEST-DEV SPLIT.

Approach	Proposal Approach	# proposals	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Fast R-CNN [12]	RPN	1000	37.0	59.5	39.9	21.1	39.4	47.0
	Cascade RPN		40.1	59.5	43.7	**22.8**	42.4	50.9
	SC-RPN(ours)		**40.2**	**59.7**	**44.0**	**22.8**	**42.6**	**51.3**
	RPN	300	36.6	58.6	39.5	20.3	39.1	47.0
	Iterative RPN		38.6	58.8	42.2	21.1	41.5	50.0
	GA-RPN		39.5	59.3	43.2	21.8	42.0	50.7
	Cascade RPN		40.1	59.4	43.8	22.1	42.4	51.6
	SC-RPN(ours)		**40.2**	**59.5**	**44.2**	**22.2**	**42.5**	**51.8**
Faster R-CNN [6]	RPN	1000	37.1	59.3	40.1	21.4	39.8	46.5
	Cascade RPN		40.5	59.3	44.2	22.6	42.9	51.5
	SC-RPN(ours)		**40.8**	**59.6**	**44.6**	**23.1**	**43.2**	**51.7**
	RPN	300	36.9	58.9	39.9	21.1	39.6	46.5
	Iterative RPN		39.2	58.2	43.0	21.5	42.0	50.4
	GA-RPN		39.9	59.4	43.6	22.0	42.6	50.9
	Cascade RPN		40.6	58.9	43.6	22.0	42.8	**52.6**
	SC-RPN(ours)		**40.9**	**59.5**	**45.0**	**22.5**	**43.3**	52.4

TABLE III
COMPONENT-WISE PERFORMANCE OF SC-RPN.

Baseline	One anchor	Two-stage	Alignment	IOU loss	Point-based rep	SADS	IOU mask	AR_{100}	AR_{300}	AR_{1000}
✓								44.6	52.9	58.3
✓	✓							44.7	51.2	55.8
✓	✓	✓						42.9	51.7	58.9
✓	✓	✓	✓					55.1	61.7	65.7
✓	✓	✓	✓	✓				55.7	62.3	66.3
✓	✓	✓	✓	✓	✓			56.1	62.7	66.7
✓	✓	✓	✓	✓	✓	✓		61.3	67.9	72.4
✓	✓	✓	✓	✓	✓	✓	✓	**61.7**	**68.4**	**72.8**
Over Improvement								**+17.1**	**+15.5**	**+14.5**

Meanwhile, the excessive number of selected feature maps introduces too much noise and limits the model performance. When two feature maps are selected in SADS, the model gets the best performance, achieving 61.3, 67.9, and 72.4 in terms of AR_{100}, AR_{300}, and AR_{1000}, respectively.

To further demonstrate the effectiveness of SADS, the IoU distribution of region proposals with and without SADS are shown in the Fig. 6. Here, x coordinate denotes IoU between region proposals and ground-truth boxes, while y coordinate denotes the statistics number of region proposals tested on MS COCO 2017 val split. The figure shows that the incorporation of SADS significantly reduces the number of region proposals with low IoU. Besides, SADS consistently increases the number of high-quality region proposals under different IoUs.

TABLE IV
ABLATION ANALYSIS OF DIFFERENT SAMPLING STRATEGY COMBINATIONS IN SADS. POSR AND SAT DENOTE POSITIVE REGION ASSIGNER AND SIZE-AWARE THRESHOLD ASSIGNER, RESPECTIVELY.

PosR	SAT	AR_{100}	AR_{300}	AR_{1000}
		56.1	62.7	66.7
	✓	58.2	65.0	69.6
✓	✓	**61.3**	**67.9**	**72.4**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

TABLE V
PERFORMANCE DEPENDENCY ON DIFFERENT HYPER-PARAMETERS IN SADS. HERE, σ DENOTES THE CENTER RATIO OF POSITIVE REGION IN POSITIVE REGION ASSIGNER. k DENOTES THE NUMBER OF THE CLOSEST SAMPLES SELECTED PER FEATURE MAP AND N DENOTES THE NUMBER OF THE SELECTED FEATURE MAPS IN SIZE-AWARE THRESHOLD ASSIGNER. THE EXPERIMENT RESULTS OF EACH HYPER-PARAMETER ARE REPORTED WHEN THE OTHER HYPER-PARAMETERS ARE OPTIMAL AND CONSTANT.

hyper-parameter		AR_{100}	AR_{300}	AR_{1000}
σ	0.1	61.1	67.8	**72.4**
	0.2	**61.3**	**67.9**	**72.4**
	0.3	61.2	67.8	72.3
	0.4	60.8	67.6	72.1
	0.5	60.2	67.2	71.8
k	3	54.7	63.3	69.7
	6	56.8	64.4	70.4
	9	**61.3**	**67.9**	**72.4**
	12	60.8	67.7	**72.4**
	15	58.9	66.5	71.9
	18	58.1	65.8	71.2
N	1	60.6	67.3	71.8
	2	**61.3**	**67.9**	**72.4**
	3	60.9	67.7	**72.4**
	4	60.2	67.3	72.3
	5	59.4	66.7	72.2

TABLE VI
PERFORMANCE COMPARISONS ON TRAINING LIGHT-WEIGHT IOU-MASK BRANCH WITH DIFFERENT TRAINING SAMPLES AND IN DIFFERENT LOCATION ACCURACY CASES. WITH AND WITHOUT SADS STAND FOR THE CASES OF HIGH LOCATION ACCURACY AND LOW LOCATION ACCURACY, RESPECTIVELY.

train with SADS	training samples	test with IoU mask	AR_{100}	AR_{300}	AR_{1000}
×	without IoU mask	×	56.1	62.7	66.7
	regression samples	×	**56.5**	**63.0**	**67.0**
		✓	**56.5**	**63.0**	**67.0**
	regression samples and all negative samples	×	56.2	62.7	66.6
		✓	56.2	62.7	66.6
	classification samples	×	56.3	62.9	66.8
		✓	56.4	62.9	66.8
✓	without IoU mask	×	61.3	67.9	72.4
	regression samples	×	61.5	68.2	72.7
		✓	61.6	68.3	72.7
	regression samples and all negative samples	×	61.2	68.0	72.5
		✓	61.2	68.0	72.5
	classification samples	×	61.5	68.2	72.4
		✓	**61.7**	**68.4**	**72.8**

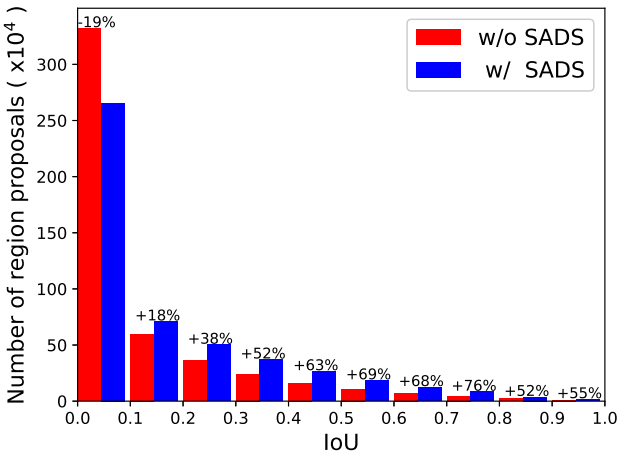


Fig. 6. The IoU distribution of region proposals with and without SADS. The incorporation of SADS significantly reduces the number of region proposals with low IoU and increases the number of high-quality region proposals under different IoUs.

2) Ablation Studies on Light-weight IoU-Mask branch. The performance comparisons on training Light-weight IoU-Mask branch with different training samples and in different location accuracy cases are shown in Table VI. With and without SADS stand for the cases of high location accuracy and low location accuracy, respectively. As we can see from the Table VI, training with regression samples in low location accuracy achieves the highest improvement, while training with classification samples in high location accuracy achieves the best performance. This phenomenon is due to the fact that there are larger difference between the quality of positive samples and negative samples in high location accuracy case, compared to that in low location accuracy case. Thus, training Light-weight IoU-Mask branch with classification samples in high location accuracy case contributes to a stronger discrimination of the model. However, training with too many negative samples will have an adverse effect in either case.

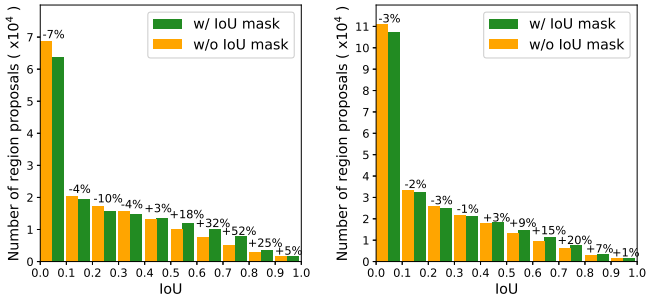
The performance comparisons on predicting IoU mask with different feature maps and in different location accuracy cases are shown in Table VII. Here, cls feature map and reg feature map denote classification feature map and regression feature map, respectively. Fused feature map is obtained by combining classification feature map and regression feature map. With and without SADS stand for the cases of high location accuracy and low location accuracy, respectively. Table VII shows similar experiment results in different location accuracy cases. Since the direct fusion of two feature maps with different properties may cause information confusion, predicting IoU mask with fused feature map only yields a slight improvement. When IoU mask is predicted by classification or regression feature map, the model obtains similar performance in either case. Finally, the model in high location accuracy case achieves 61.7, 68.4, and 72.8 in terms of AR_{100}, AR_{300}, and AR_{1000}, respectively.

Combined with experiment results in Table VI and Table VII, we can easily observe that in the case of high location accuracy, testing with IoU mask can achieve higher performance than testing without IoU mask. But in the case of low location accuracy, testing with IoU mask can hardly obtain improvement. In addition, the overall improvement of Light-weight IoU-Mask branch in high location accuracy case is more significant than that in the case of low location accuracy. Since the supervision information of Light-weight IoU-Mask branch is derived from the refined region proposals, the higher the location accuracy is, the higher the quality of supervision information will obtain. Therefore, Light-weight IoU-Mask

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

TABLE VII
PERFORMANCE COMPARISONS ON PREDICTING IOU MASK WITH DIFFERENT FEATURE MAPS AND IN DIFFERENT LOCATION ACCURACY CASES. HERE, CLS FEATURE MAP AND REG FEATURE MAP DENOTE CLASSIFICATION FEATURE MAP AND REGRESSION FEATURE MAP, RESPECTIVELY. FUSED FEATURE MAP IS OBTAINED BY COMBINING CLASSIFICATION FEATURE MAP AND REGRESSION FEATURE MAP.

train with SADS	feature map	test with IoU mask	AR_{100}	AR_{300}	AR_{1000}
×	without IoU mask	×	56.1	62.7	66.7
	fused feature map	×	56.1	62.7	66.7
		✓	56.2	62.7	66.7
	cls feature map	×	56.3	62.9	**66.9**
		✓	56.3	**63.0**	**66.9**
	reg feature map	×	56.3	62.9	66.8
		✓	**56.4**	62.9	66.8
✓	without IoU mask	×	61.3	67.9	72.4
	fused feature map	×	61.4	68.1	72.5
		✓	61.6	68.3	72.5
	cls feature map	×	61.6	68.2	72.6
		✓	**61.7**	**68.4**	72.7
	reg feature map	×	61.5	68.2	72.4
		✓	**61.7**	**68.4**	**72.8**



(a) The top 300 region proposals (b) The top 1000 region proposals

Fig. 7. The changes in the IoU distribution of region proposals with and without IoU mask. Here, (a) and (b) describe the changes in the IoU distribution of the top 300 region proposals and the top 1000 region proposals, respectively. Both (a) and (b) verify that IoU mask can refine the classification scores of region proposals and prevent the high-quality region proposals from being removed in NMS.

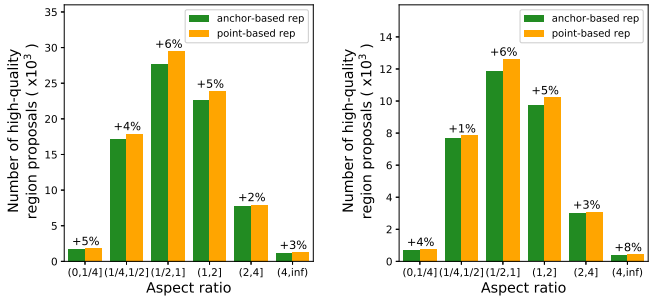
branch can effectively exploit the potential of high-quality model.

In Fig. 7, we visualize the changes in the IoU distribution of region proposals after using IoU mask. All the region proposal results are tested on MS COCO 2017 val split. Fig. 7 (a) and Fig. 7 (b) describe the changes in the IoU distribution of the top 300 region proposals and the top 1000 region proposals, respectively. These two figures consistently show a trend: when we apply IoU mask for testing, the number of region proposals with high IoU increases significantly. This phenomenon verifies that using IoU mask can refine the classification scores of region proposals and prevent the high-quality region proposals from being removed in NMS.

3) Ablation Studies on Point-based Representation. The experiment results of different components in point-based representation are shown in Table VIII. In this work, point-

TABLE VIII
ABLATION ANALYSIS OF DIFFERENT COMPONENTS IN POINT-BASED REPRESENTATION. THE FIRST LINE DENOTES THE PERFORMANCE OF ADOPTING ANCHOR-BASED REPRESENTATION.

point-based representation		AR_{100}	AR_{300}	AR_{1000}
point alignment	point description			
		55.7	62.3	66.3
	✓	56.0	62.6	66.5
✓	✓	**56.1**	**62.7**	**66.7**



(a) High-quality threshold = 0.7 (b) High-quality threshold = 0.8

Fig. 8. The performance of bounding box representation and point-based representation on objects with different aspect ratios. Here, (a) and (b) denote the high-quality threshold is set to 0.7 and 0.8, respectively. Both (a) and (b) show that with point-based representation, the model can better fit the object outline and generate more high-quality region proposals for objects with different aspect ratios.

based representation can be divided into two parts: point alignment and point description. Here, the point-based representation consistently outperforms the anchor-based representation under different region proposal numbers in terms of AR, implying that abandoning the use of anchors can better fit the object outline and achieve higher performance.

In order to further demonstrate the effectiveness of point-based representation, we delve into the performance of bounding box representation and point-based representation on objects with different aspect ratios. We firstly regard the intersection-over-union (IoU) between refined region proposals and ground-truth box as the IoU of region proposals. Then the region proposals whose IoU higher than "high-quality threshold" are regarded as high-quality region proposals. As shown in Fig. 8, x coordinate denotes different aspect ratio of objects, while y coordinate denotes the statistics number of high-quality region proposals. Fig. 8 (a) and Fig. 8 (b) denote the high-quality threshold is set to 0.7 and 0.8, respectively. All the region proposal results are tested on MS COCO 2017 val split. The figure reveals that under point-based representation, objects with different aspect ratios have more high-quality region proposals. In other words, since point-based representation is not sensitive to the object outline, it can better fit the object outline and generate more high-quality region proposals for all objects.

4) Extension With More Object Detectors. More detection results of SC-RPN combined with other object detector pipelines are reported in Table IX. Here, we investigate Double-Head R-CNN [54] and Cascade R-CNN [55] with different proposal approaches. Both the baseline of Double-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

TABLE IX
PERFORMANCE OF MORE OBJECT DETECTORS ON MS COCO 2017 TEST-DEV SPLIT.

Approach	Proposal Approach	# proposals	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Double Head R-CNN [54]	RPN	1000	39.6	59.9	43.2	23.2	42.5	49.3
	SC-RPN (ours)		**40.6**	**60.3**	**44.1**	**23.3**	**43.6**	**51.0**
Cascade R-CNN [55]	RPN	1000	40.7	59.2	44.3	23.0	43.3	51.3
	SC-RPN (ours)		**41.7**	**60.2**	**45.3**	**23.9**	**44.4**	**52.9**

TABLE X
PERFORMANCE OF REGION PROPOSAL APPROACHES ON IMAGENET-DET DATASET.

Approach	Dataset	Backbone	AR_{100}	AR_{300}	AR_{1000}	AR_S	AR_M	AR_L
RPN	ImageNet-DET	ResNet50-FPN	41.1	49.2	55.0	29.8	51.7	61.1
GA-RPN			63.6	67.5	69.7	43.0	64.1	77.3
Cascade RPN			**67.3**	71.6	74.3	46.6	68.3	82.4
SC-RPN (ours)			67.1	**71.7**	**75.0**	**46.8**	**69.4**	**82.7**

Head R-CNN and Cascade R-CNN are two-stage frameworks and adopt RPN in the first stage. Combined with SC-RPN, Double-Head R-CNN and Cascade R-CNN surge the mAP to 40.6 and 41.7, respectively. It can be concluded that the proposed SC-RPN significantly outperforms the RPN in terms of AP under different settings of thresholds and object sizes.

5) Performance on ImageNet-DET Dataset. We present the performance of different region proposal approaches on ImageNet-DET (ILSVRC 2015) [56] dataset in Table X. To make a fair comparison, all the experimental implementation details are consistent with the experiments on MS COCO 2017 detection dataset. Regarded as the most important performance indicator, the proposed SC-RPN achieves the highest performance in AR_{1000}, surpassing all the existing region proposal approaches. Furthermore, in terms of AR under different settings of thresholds and object sizes, the proposed SC-RPN achieves the best performance except AR_{100}. And we can easily find that SC-RPN gains slighter improvement on ImageNet-DET dataset compared to that on MS COCO 2017 detection dataset. The average object number on each image of ImageNet-DET dataset is several times lower than that of MS COCO 2017 detection dataset, in other words, MS COCO 2017 detection dataset is much more complex than ImageNet-DET dataset. However, the proposed training method SADS in SC-RPN, which brings the most significant improvement, is proposed to set independent dynamic overlapping threshold for each object on an image. When training with a simple dataset, the potential of SADS isn't fully exploited, nor is the SC-RPN. Therefore, SC-RPN can achieve stronger improvement on complex dataset ($e.g.$, MS COCO 2017 detection dataset).

V. CONCLUSION

In this paper, we propose a novel two-stage strong correlation learning framework, abbreviated as SC-RPN, to generate high-quality region proposals. In order to tackle the weak correlation between location and classification of region proposals, we subtly design an extra Light-weight IoU-Mask branch to refine the classification score, which prevents the high-quality region proposals from being filtered in NMS. Besides, considering the shortcoming of anchors to represent region proposals, point-based representation is adopted in SC-RPN to generate region proposals with strong fitting ability. Furthermore, to address the weak correlation of sampling strategy between the two stages, Size-Aware Dynamic Sampling (SADS) is applied to ensure the sampling consistency during the training. Finally, we report the overall performance of the proposed SC-RPN, which surpasses all the state-of-the-art region proposal approaches. The effectiveness of the proposed approach and its components are also validated in ablation experiments. Even when combined with several object detector pipelines, SC-RPN achieves the state-of-the-art performance in terms of AP under different settings of thresholds and object sizes. As for future work, we will extend the concept of weak correlation and the proposed approach to other related tasks, $e.g.$, instance segmentation and object tracking.

REFERENCES

[1] P. Furgale, U. Schwesinger, M. Rufli, W. Derendarz, H. Grimmett, P. Mühlfellner, S. Wonneberger, J. Timpner, S. Rottmann, B. Li *et al.*, "Toward automated driving in cities using close-to-market sensors: An overview of the v-charge project," in *Proc. of Intelligent Vehicles Symposium*, 2013, pp. 809–816.

[2] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[3] T. He, W. Huang, Y. Qiao, and J. Yao, "Text-attentional convolutional neural network for scene text detection," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2529–2541, 2016.

[4] D. Conte, P. Foggia, M. Petretta, F. Tufano, and M. Vento, "Meeting the application requirements of intelligent video surveillance systems in moving object detection," in *Proc. of Pattern Recognition and Image Analysis*, 2005, pp. 653–662.

[5] W. Liao, C. Yang, M. Ying Yang, and B. Rosenhahn, "Security event recognition for visual surveillance," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 4, no. 1W1, pp. 19–26, 2017.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Itelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[9] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2017, pp. 7263–7271.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. of European Conference on Computer Vision*, 2016, pp. 21–37.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[12] R. Girshick, "Fast r-cnn," in *Proc. of IEEE International Conference on Computer Vision*, 2015, pp. 1440–1448.

[13] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2020, pp. 11 583–11 591.

[14] K. Chen, W. Lin, J. See, J. Wang, J. Zou *et al.*, "Ap-loss for accurate one-stage object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[15] Z. Chen, K. Chen, W. Lin, J. See, H. Yu, Y. Ke, and C. Yang, "Piou loss: Towards accurate oriented object detection in complex environments," in *Proc. of European Conference on Computer Vision*, 2020, pp. 195–211.

[16] A. Shepley, G. Falzon, and P. Kwan, "Confluence: A robust non-iou alternative to non-maxima suppression in object detection," *arXiv preprint arXiv:2012.00257*, 2020.

[17] C. Guo, B. Fan, Q. Zhang, S. Xiang, and C. Pan, "Augfpn: Improving multi-scale feature learning for object detection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2020, pp. 12 595–12 604.

[18] K. E. Van de Sande, J. R. Uijlings, T. Gevers, and A. W. Smeulders, "Segmentation as selective search for object recognition," in *Proc. of International Conference on Computer Vision*, 2011, pp. 1879–1886.

[19] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[20] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized prim's algorithm," in *Proc. of IEEE International Conference on Computer Vision*, 2013, pp. 2536–2543.

[21] P. Rantalankila, J. Kannala, and E. Rahtu, "Generating object segmentation proposals using global and local search," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014, pp. 2417–2424.

[22] J. Carreira and C. Sminchisescu, "Constrained parametric min-cuts for automatic object segmentation," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 3241–3248.

[23] J. Carreira and C. Sminchisescu, "Cpmc: Automatic object segmentation using constrained parametric min-cuts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1312–1328, 2011.

[24] I. Endres and D. Hoiem, "Category independent object proposals," in *Proc. of European Conference on Computer Vision*, 2010, pp. 575–588.

[25] I. Endres and D. Hoiem, "Category-independent object proposals with diverse ranking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 222–234, 2013.

[26] A. Humayun, F. Li, and J. M. Rehg, "Rigor: Reusing inference in graph cuts for generating object regions," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014, pp. 336–343.

[27] P. Krähenbühl and V. Koltun, "Geodesic object proposals," in *Proc. of European Conference on Computer Vision*, 2014, pp. 725–739.

[28] B. Alexe, T. Deselaers, and V. Ferrari, "What is an object?" in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010, pp. 73–80.

[29] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[30] E. Rahtu, J. Kannala, and M. Blaschko, "Learning a category independent object detection cascade," in *Proc. of IEEE International Conference on Computer Vision*, 2011, pp. 1052–1059.

[31] M. B. Blaschko, J. Kannala, and E. Rahtu, "Non maximal suppression in cascaded ranking models," in *Scandinavian Conference on Image Analysis*, 2013, pp. 408–419.

[32] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014, pp. 3286–3293.

[33] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo, "Spatio-temporal closed-loop object detection," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1253–1263, 2017.

[34] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. of European Conference on Computer Vision*, 2014, pp. 391–405.

[35] M. Van den Bergh, G. Roig, X. Boix, S. Manen, and L. Van Gool, "Online video seeds for temporal window objectness," in *Proc. of IEEE International Conference on Computer Vision*, 2013, pp. 377–384.

[36] C. Szegedy, S. Reed, D. Erhan, D. Anguelov, and S. Ioffe, "Scalable, high-quality object detection," *arXiv preprint arXiv:1412.1441*, 2014.

[37] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014, pp. 2147–2154.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[39] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "Deepproposal: Hunting objects by cascading deep convolutional layers," in *Proc. of IEEE International Conference on Computer Vision*, 2015, pp. 2578–2586.

[40] Z. Jie, X. Liang, J. Feng, W. F. Lu, E. H. F. Tay, and S. Yan, "Scale-aware pixelwise object proposal networks," *IEEE Transactions on Image Processing*, vol. 25, no. 10, pp. 4525–4539, 2016.

[41] Q. Zhong, C. Li, Y. Zhang, D. Xie, S. Yang, and S. Pu, "Cascade region proposal and global context for deep object detection," *Neurocomputing*, vol. 395, pp. 170–177, 2020.

[42] J. Wang, K. Chen, S. Yang, C. C. Loy, and D. Lin, "Region proposal by guided anchoring," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2019, pp. 2965–2974.

[43] T. Vu, H. Jang, T. X. Pham, and C. Yoo, "Cascade rpn: Delving into high-quality region proposal network with adaptive convolution," in *Advances in Neural Information Processing Systems*, 2019, pp. 1432–1442.

[44] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proc. of European Conference on Computer Vision*, 2018, pp. 784–799.

[45] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2020, pp. 9759–9768.

[46] Z. Yang, S. Liu, H. Hu, L. Wang, and S. Lin, "Reppoints: Point set representation for object detection," in *Proc. of IEEE International Conference on Computer Vision*, 2019, pp. 9657–9666.

[47] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, "Unitbox: An advanced object detection network," in *Proc. of the 24th ACM International Conference on Multimedia*, 2016, pp. 516–520.

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. of European Conference on Computer Vision*, 2014, pp. 740–755.

[49] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu *et al.*, "Mmdetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[50] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. of European Conference on Computer Vision*, 2016, pp. 75–91.

[51] H.-F. Lu, X. Du, and P.-L. Chang, "Toward scale-invariance and position-sensitive region proposal networks," in *Proc. of European Conference on Computer Vision*, 2018, pp. 168–183.

[52] S. Gidaris and N. Komodakis, "Attend refine repeat: Active box proposal generation via in-out localization," *arXiv preprint arXiv:1606.04446*, 2016.

[53] H. Li, Y. Liu, W. Ouyang, and X. Wang, "Zoom out-and-in network with map attention decision for region proposal and object detection," *International Journal of Computer Vision*, vol. 127, no. 3, pp. 225–238, 2019.

[54] Y. Wu, Y. Chen, L. Yuan, Z. Liu, L. Wang, H. Li, and Y. Fu, "Double-head rcnn: Rethinking classification and localization for object detection," *arXiv preprint arXiv:1904.06493*, vol. 2, p. 7, 2019.

[55] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proc. of IEEE Computer Vision and Pattern Recognition*, 2018, pp. 6154–6162.

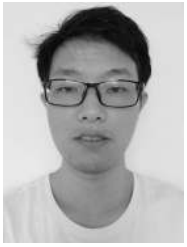
[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2021.3069547, IEEE Transactions on Image Processing

Wenbin Zou received the M.E. degree in software engineering with a specialization in multimedia technology from Peking University, China, in 2010, and the Ph.D. degree from the National Institute of Applied Sciences, Rennes, France, in 2014. From 2014 to 2015, he was a Researcher with the UMR Laboratoire d'informatique Gaspard-Monge, CNRS, and Ecole des Ponts ParisTech, France. He is currently an Associate Professor of College of Electronics and Information Engineering, Shenzhen University, China. His current research interests include saliency detection, object segmentation, and semantic segmentation.

Zhengyu Zhang received the B.E. degree in electronic and information science and technology from Guangzhou University, Guangzhou, China, in 2018. He is currently pursuing the M.E. degree with Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Institute of Artificial Intelligence and Advanced Telecommunication, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include object detection, semantic segmentation and image quality assessment.

Yingqing Peng received the B.E. degree in communication engineering from Wuyi University, Jiangmen, China, in 2018. She is currently pursuing the M.E. degree with Shenzhen Key Laboratory of Advanced Machine Learning and Applications, Institute of Artificial Intelligence and Advanced Telecommunication, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. Her research interests include semantic segmentation, object detection and image enhancement.



Canqun Xiang received the B.S. and the integrated M.S. degrees from the Department of Electric and Information Engineering, Hunan Institute of Science and Technology, Hunan, China, in 2015 and 2018, respectively. He is pursuing a doctoral degree at College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. His research interests include capsule network and image processing in various applications.

Shishun Tian received the B.Sc. degree from Sichuan University, Chengdu, China, the M.Sc. degree from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, and the Ph.D. degree from National Institute of Applied Sciences, Rennes, France in 2012, 2015 and 2019 respectively. He is currently an assistant professor of College of Electronics and Information Engineering of Shenzhen University, Shenzhen, China. His research interests include image quality assessment, visual perception and machine learning.

Lu Zhang is an associate professor at National Institute of Applied Sciences (INSA) of Rennes in France. She received the B.S degree from Southeast University and the M.S. degree from Shanghai Jiaotong University in China in 2004 and 2007, respectively. From October 2009 to November 2012, she was a PhD student of the LISA and CNRS IRCCyN labs in France, working on the model observers for the medical image quality assessment. She received the Excellent Doctoral Dissertation of France awarded by IEEE France Section, SFGBM, AGBM and GdR CNRS-Inserm Stic-Santé in 2013. Then she worked on the quality of experience (QoE) in telemedicine before she joined INSA in September 2013, as a member of the VAADER research group of the IETR lab. She is a member of the international VQEG (Video Quality Experts Group). She works on human perception understanding, image analysis, saliency detection and image quality assessment.