AdaZoom: Adaptive Zoom Network for Multi-Scale Object Detection in Large Scenes

Jingtao Xu Yali Li Shengjin Wang
Department of Electronic Engineering, Tsinghua University
xjd19@mails.tsinghua.edu.cn liyali13, wgsgj@tsinghua.edu.cn

Abstract

Detection in large-scale scenes is a challenging problem due to small objects and extreme scale variation. It is essential to focus on the image regions of small objects. In this paper, we propose a novel Adaptive Zoom (AdaZoom) network as a selective magnifier with flexible shape and focal length to adaptively zoom the focus regions for object detection. Based on policy gradient, we construct a reinforcement learning framework for focus region generation, with the reward formulated by object distributions. The scales and aspect ratios of the generated regions are adaptive to the scales and distribution of objects inside. We apply variable magnification according to the scale of the region for adaptive multi-scale detection. We further propose collaborative training to complementarily promote the performance of AdaZoom and detection network. To validate the effectiveness, we conduct extensive experiments on VisDrone2019, UAVDT and DOTA datasets. The experiments show AdaZoom brings consistent and significant improvement over different detection networks, achieving state-of-the-art performance on these datasets, especially outperforming the existing methods by AP of 4.64% on VisDrone2019.

1. Introduction

In recent years, significant progress has been achieved in computer vision. Visual object detection has also been extensively studied since it is important in various applications such as video surveillance and autonomous driving. Existing detectors such as Faster R-CNN [33], YOLO [32], and CornerNet [18] achieve satisfying performance on natural images. However, in practical applications such as Unmanned Aerial Vehicle (UAV) vision, existing detectors perform poorly because the images capture large-scale scenes with wide fields of view and quite small objects.

There are several challenges for detecting objects in large-scale scenes: (1) Objects are small and dense. For example, the UAV takes images from high altitude with wide field of view, as shown in Fig. 1. Tens and hundreds of objects exist in a single image and most of them occupy quite a few pixels. Deep neural network with successive down-sampling would bring intolerable loss for semantic and positional information of small objects, resulting in poor detection performance. (2) Extreme scale variation across objects arises since images of large-scale scenes record a large span of distance and the camera-object distance varies significantly. Objects become smaller with further distance. Even objects of the same category may differ hundred times in scale. However, the receptive field of convolutional neural networks is limited. Extreme scale variation results in semantic gaps in convolution layers and brings substantial burdens in learning the powerful feature representations.

To tackle the object detection in large-scale scenes, it is urgent to design an adaptive zoomer to “focus” on objects with varying scales. Although there are some works on resizing images for multi-scale training [35, 36] and inference...
on enlarged image crops [10, 30, 42], they are inflexible for objects of various scales in large-scale scenes and separate training and inference into different pipelines.

We propose an Adaptive Zoom (AdaZoom) network based on policy gradient [37] to adaptively zoom the focus regions for further detection. Inspired by human perception [29], when we perceive a large-scale scene, we glance over the whole image for coarse cognition and zoom where objects are small and dense for a careful watch.

AdaZoom works as a selective magnifier with flexible shape and focal length. It focuses on image regions with small objects. The scale and aspect ratio of the region to be zoomed are adaptive to the scales and distribution of objects inside. For a cluster of smaller objects, AdaZoom prefers a smaller focus region enclosing them for higher magnification, just like using a magnifier with shorter focal length. Without additional annotations for regions, AdaZoom is optimized according to the reward which measures the quality of the focus region. Following the paradigm of deep reinforcement learning, we further learn the policy network to produce focus regions with the regard of visual features. In addition, we design collaborative training to iteratively promote the joint training performance of AdaZoom and the detector. The outputs of detector are introduced to the region proposals to adaptively zoom over the whole image for coarse cognition and zoom where objects are small and dense for a careful watch.

In summary, the main contributions are as follows:

- We propose a novel Adaptive Zoom (AdaZoom) network to adaptively generate and zoom the focus regions for accurate detection in large scenes, without additional annotations for the regions.
- We propose collaborative training to jointly boost the coordination between AdaZoom and the detector with a consistent pipeline of training and inference.
- Without bells and whistles, AdaZoom achieves state-of-the-art on the VisDrone2019 [47], UAVDT [9] and DOTA [41] datasets.

2. Related work

**Multi-scale object detection.** Multi-scale training and inference with image pyramid is the most straightforward idea to alleviate the problem of small objects [7, 11, 16, 34]. However, the image pyramid increases the scale variation of images. SNIP [35] proposes a training paradigm ignoring objects out of the desire size range during gradient back-propagation. Following the idea of [35], SNIPER [36] focuses on efficient multi-scale training by sampling chips of different sizes. These chips are resized to a certain size for multi-scale training. Further, AutoFocus [30] extends SNIPER [36] to a coarse-to-fine pipeline that predicts regions of interest at a coarse level and then infers on the regions at a fine level. In addition, DREN [44] and ClusDet [42] employ neural networks to estimate difficult regions for fine detection. However, these methods separate multi-scale training and inference into different pipelines which leads to inconsistency between training and inference. Apart from multi-scale process in image level, another direction is to build pyramid in feature level [13, 15, 21, 23, 24]. Feature Pyramid Network [21] is widely utilized in many SOTA detectors for cross-layer feature fusion. Due to the limitation of the receptive field of convolutional neural networks, [3, 24] assign smaller objects to shallower layers. Our method focuses on image level. We introduce deep reinforcement learning into adaptive region generation for multi-scale object detection and integrate the training and inference into the same pipeline with the regions generated by AdaZoom.

**Reinforcement-learning-based object detection.** Reinforcement learning is introduced to object detection in the following ways: (1) Focusing on objects step by step with accumulated evidence [2, 5, 12, 14, 20, 43]. In particular, an image contains specific context information and a sequence process accumulates evidence from context for detection. [12] learns a search strategy to collect context and select the next window to visit. In addition, objects in the same scene have relationship with each other, such as a person riding on a bicycle. The detected objects provide contextual cues for subsequent steps of detection [17]. (2) Selecting high-quality regions of interest. [28] proposes a sequential exploration process to select region proposals. [31] introduces deep reinforcement learning into region proposal network to filter out low-quality proposals and accumulate class-specific evidence over time step to boost detection accuracy. Besides, several works [10, 26, 38] propose selection strategies to acquire image regions from enormous candidates. Sharing the similar idea, we go beyond in the adaptability of region generation and propose effective collaborative training for region generation and detection.

3. Method

3.1. Analysis of Focus Adaptability

For effective object detection in large-scale scenes, it is essential to focus on the object-centric regions. The Uniform Partition (UP) is a straightforward strategy to “focus” on image regions in a sliding-window way. It partitions an image into several uniform regions and enlarges those regions for detection. We first conduct the UP strategy on the VisDrone2019 [47] for analysis. Various settings of UP \((1 \times 1, 2 \times 2, 3 \times 3, 4 \times 4)\) and all the combinations are evaluated \((n \times m)\) denotes uniformly partitioning the image...
into \( n \times m \) regions). The detection accuracy measured by average precision (AP) is presented in Fig. 2. The UP of appropriate settings does improve the AP, which validates the importance of focus regions in large scenes. Compared to single-scale setting (Fig. 2 left), multi-scale UP would further improve the detection accuracy (Fig. 2 right). However, when the number of regions increases, the detection accuracy drops notably. That is because the false negatives accumulate with the multi-scale repeated partitioned regions. Besides, with more cropped regions, the possibility of objects truncated by additional cropping is higher, which would cause more repeated and incomplete detections. Moreover, the UP strategy enlarges the cropped image regions with a fixed scale, which is inappropriate for scale-adaptive object detection in large scenes. In a word, the performance of UP strategy is limited by the lacking of adaptivity issue.

To tackle this, it is essential to adaptively generate focus regions in large-scene images for effective and efficient detection. Yet it is difficult to label the focus regions as true or false since the semantics is ambiguous. This makes it difficult to learn the designed network in supervised learning scheme. Based on these considerations, we propose an Adaptive Zoom network (AdaZoom) based on reinforcement learning (RL), as shown in Fig. 4. We design a continuous reward to measure the quality of focus regions based on the scales and distribution of objects. The RL agent is then encouraged to explore an adaptive region generation policy which maximizes the accumulated reward. In particular, AdaZoom can localize and zoom the focus regions with flexible shape and scale. The focus regions are adaptive to the scales and distribution of objects, which can significantly alleviate the scale variation towards high-performance object detection in large scenes.

### 3.2. Problem Formulation

We construct AdaZoom with reinforcement learning framework. Based on policy gradient [37], AdaZoom is optimized according to the reward which measures the quality of the focus region. We view the sequence of region generation as a Markov Decision Process (MDP) [1] and formulate adaptive focus region generation as a reinforcement learning problem. As shown in Fig. 3, we construct the Policy Network \( \pi_\theta \) to generate the probability distribution of action space based on the state \( S_t \). The action \( A_t \) is decoupled into fixation, scale, and aspect ratio of the focus region. The object distribution is referred to guide the sampling process for better convergence. Then we derive the reward from object distribution to measure the quality of a region. The state \( S_{t+1} \) is updated according to the generated region. For each image, \( T \) time steps make up an episode and we optimize the Adazoom with policy gradient [37] to maximize the expected cumulative reward of each episode.

**State.** The state \( S_t \) consists of the base feature map \( F_t \) and the history information map \( H_t \). \( F_t \) is supposed to learn the object-wise scale and distribution information at a coarse level. It is distinct from the feature maps for object detection which focus on the fine-level features of each object. The binary history information map \( H_t \) records the regions generated at previous steps. The initial state \( S_0 \) is the concatenation of an all-zero binary map \( H_0 \) and the base feature map \( F_0 \) which is extracted from the image by a backbone network. The focus region is mapped from the image to the state, denoted as \( z_t \). The state is updated as follows:

\[
H_{t+1}(i, j) = I\{(i, j) \in z_t\} + H_t(i, j)I\{(i, j) \notin z_t\} \quad (1)
\]

\[
F_{t+1}(i, j) = \kappa F_t(i, j)I\{(i, j) \in z_t\} + F_t(i, j)I\{(i, j) \notin z_t\} \quad (2)
\]

where \( I(\cdot) \) is the indicator function and \( \kappa \) is a decay factor to suppress the response in the corresponding focus region. We set \( \kappa \) to 0.1.

**Action.** The action \( A_t \) is sampled from the probability distribution \( \pi_\theta(A_t | S_t) \). We design a three-branch Policy Network to progressively learn the fixation, scale and aspect ratio of focus regions. The first branch generates the probability distribution for fixation \( p_f(a_f | S_t) \), where \( a_f \) is
a point on the fixation probability map \( p_f \in \mathbb{R}^{h \times w} \). Inspired by the anchor-based mechanism [33], each point in the fixation probability map represents a region center location in the image. The second branch generates a scale probability map \( p_s \in \mathbb{R}^{h \times w \times n_s} \), where \( n_s \) is the predefined number of candidate region scales. \( p_s(a_s|a_f; S_t) \) is conditioned on the fixation \( a_f \), where \( a_s \) is the region scale. The third branch generates aspect ratio probability map \( p_r \in \mathbb{R}^{h \times w \times n_r \times n_s} \), where \( n_r \) is the predefined number of candidate aspect ratios. The aspect ratio probability distribution \( p_r(a_r|a_f, a_s; S_t) \) is conditioned on the fixation \( a_f \) and scale \( a_s \). The action \( A_t \) is composed of \( (a_f, a_s, a_r) \) to specify a region. We formulate the policy as follows:

\[
\pi_\theta(A_t|S_t) = p_f(a_f|S_t) \times p_s(a_s|a_f; S_t) \times p_r(a_r|a_f, a_s; S_t) \quad (3)
\]

These branches coordinate with each other by focusing on region representation from different points of view. The fixation branch tries to find the center of a cluster of objects. The scale branch is supposed to adjust the scale of the region according to the scales of objects around the fixation. The aspect ratio branch adapts to the distribution of objects around the fixation with a selected scale of the region. The network structure is detailed in the Supplementary Material.

**Reward.** There are no clear annotations to supervise the region generation since the image can be partitioned into reasonable regions in many ways and the semantics of the region is ambiguous for supervised learning. Therefore, we design the reward which measures the quality of the regions. The reward is derived from annotations for bounding boxes of objects. AdaZoom is expected to pay more attention to small objects. For the \( i_{th} \) object, it is assigned a weight \( w_i \propto \frac{1}{s_i}\), where \( s_i \) is the scale of the object. For each scale \( a_s \) of region, there is a desired object scale range \([a_s^{min}, a_s^{max}]\). The reward at step \( t \) is defined as follows:

\[
r_t(A_t) = \frac{\sum Z_t I_i(a_s) w_i}{\sum_j w_j} \quad (4)
\]

where \( Z_t \) is the set of objects enclosed in the \( t_{th} \) focus region and \( Z \) corresponds to the remaining regions of the image until step \( t - 1 \). In general, \( r_t \) can be regarded as a weighted recall of object weights \( w \) at step \( t \). \( I_i(a_s) \) denotes the measurement of consistency between scales of objects and regions. For the \( i_{th} \) object in the focus region with \( a_s \) scale, if the scale of the object falls beyond \([a_s^{min}, a_s^{max}]\), \( w_i \) will decay by \( I_i(a_s) \) and the mismatch between the scale of the region and object is introduced to reward. \( I_i(a_s) \) is defined as follows:

\[
I_i(a_s) = \begin{cases} 
1, & s_i \in [a_s^{min}, a_s^{max}] \\
\text{max}(0, 2 - e^{\beta \Delta s}), & \text{otherwise}
\end{cases} \quad (5)
\]

where \( s_i \) is the scale of the object, \( \Delta s = \frac{|s_i - a_s^{min}(max)|}{a_s^{max}(min)} \) is the fraction that \( s_i \) exceeds \( a_s^{max} \) or smaller than \( a_s^{min} \). \( \beta \) is a positive coefficient to adjust the decay rate for scales beyond scale range. We set \( \beta \) to 1.5.

3.3. Training and Inference

Our approach for object detection in large-scale scenes contains two components: (1) AdaZoom network to adap-
tively zoom the focus regions; (2) Detection network to locate the objects in focused regions. We train AdaZoom and detection network collaboratively to boost the performance.

**Collaborative training.** AdaZoom generates a series of focus regions with different scales and aspect ratios. We crop these focus regions from the original image and resize them to a certain scale so that smaller regions obtain higher magnification. In order to alleviate the domain shift between original images and resized regions, the detector is re-trained on the resized regions. Different from the existing work [10, 30, 36], we integrate the training and inference in the same pipeline with the focus regions generated by AdaZoom. In addition, we design collaborative training to further improve the performance of detection. During collaborative training, AdaZoom is expected to focus on the regions difficult for the re-trained detector. The detector infers on the image and outputs confidence scores for each object. For false negative, the confidence score is set to zero. Then the weight $w_i$ in Eq. 4 is modified as $w_i = 1 - c_i$. The AdaZoom is trained on the new modified weights. As a consequence, AdaZoom pay attention to difficult regions where the confidence of true positive is low. Then the difficult regions generated by the AdaZoom are used to finetune the detector (Fig. 5). This simple modification promotes the coordination between AdaZoom and the detector.

**Inference.** Inference shares the same pipeline with training. We adopt the greedy sampling to take actions:

$$\begin{align*}
a_f &= \arg \max_{a_f} p_f(a_f | S_t) \\
a_s &= \arg \max_{a_s} p_s(a_s | a_f; S_t) \\
a_r &= \arg \max_{a_r} p_r(a_r | a_f, a_s; S_t)
\end{align*}$$

(6)

The fixation $a_f$ is selected as the center of regions. Conditioned on fixation $a_f$, AdaZoom selects a scale $a_s$ that implies the scales of objects around fixation. With the selected fixation $a_f$ and scale $a_s$, an aspect ratio $a_r$ is selected to adapt to the distribution of objects around fixation. Based on the selected actions, a region is generated. The generated regions as well as the original image are resized together to a certain scale as a batch for detection. The final results of each region are merged together by non-maximum suppression (NMS) with the IoU threshold setting to 0.5.

4. Experiments

4.1. Dataset and Metric

We conduct experiments on the public detection benchmark VisDrone2019 [47], UAVDT [9] and DOTA [41] to evaluate our method. (1) **VisDrone2019** [47] consists of 10,209 images for detection task with train set of 6,471 images, val set of 548 images. The test set is split into test-dev with 1,610 images and test-challenge with 1,580 images.

![Figure 5. When training the detector, the focus regions generated by AdaZoom are resized to a certain scale to finetune the detector. When training AdaZoom, the detection results of the image are introduced to the reward for AdaZoom. During inference, the final detection results are merged from focus regions and the image.](image-url)
Table 1. Comparison with uniform partition on VisDrone2019 test-dev and UAVDT test dataset. We adopt Faster R-CNN as the detector. We also report the inference time per image.

Table 2. Detection performance on UAVDT test dataset. We adopt Faster R-CNN and Cascade R-CNN as the detector. † denotes Cascade R-CNN.

4.3. Comparison with Baseline

We first compare the proposed AdaZoom with the Uniform Partitions (UP) as the baseline, to evaluate the effectiveness. For UP, a whole image is uniformly partitioned into \( m \times n \) regions with 50 pixels overlap. In particular, we implement comprehensive experiments of UP to provide a simple yet strong baseline. We evaluate multi-scale UP of \([1\times1, 2\times2, 3\times3]\) and multi-ratio UP of \([2\times3, 2\times2, 3\times2]\). Table 1 shows the comparative study with Faster R-CNN as the detector and ResNet-50 backbone. On VisDrone2019 dataset, the AdaZoom achieves \( AP \) of 31.22% and \( AP_{50} \) of 56.16%. Compared to UP of \( 3 \times 3 \), the AdaZoom improves \( AP \) by 2.58% with comparable inference time. Compared to the multi-scale UP and multi-ratio UP, the AdaZoom improves the \( AP \) by 1.47% \( \sim \) 2.21% and \( AP_{50} \) by 2.27% \( \sim \) 3.17%. On UAVDT dataset, the AdaZoom achieves \( AP \) of 19.6% and \( AP_{50} \) of 33.6%, outperforming UP by a large margin. Compared to multi-scale UP and multi-ratio UP, we improve \( AP \) by 4.3% \( \sim \) 4.6% and \( AP_{50} \) by 5.5% \( \sim \) 6.3%, respectively. Meanwhile, compared to multi-scale UP, we save the 31.8% and 29.8% of inference time on VisDrone2019 and UAVDT dataset, respectively. The results validate that the proposed AdaZoom is both effective and efficient of object detection in large scenes. That is mainly because our method generates focus regions in an adaptive way. Instead of sliding over the whole image, we only generate the important regions with appropriate scales and shapes. Therefore, our method is adaptive to scale varying. Besides, the AdaZoom would avoid generating too many regions, further leading to notable accuracy boosting and inference time saving.

4.4. Comparison with State-of-The-Art Models

We compare our method with the state-of-the-art methods across a wide range of datasets, such as UAVDT test (Table 2), VisDrone2019 val (Table 3) and DOTA val dataset (Table 4). From the tables we can see that our method achieves superior performance over existing methods on all the three datasets. On UAVDT dataset, we achieve 22.4% of \( AP \) and 38.6% of \( AP_{50} \) with Cascade R-CNN as the detector. Compared to GLSAN [8] with Cascade R-CNN, the \( AP \) is increased by 3.4% and \( AP_{50} \) is increased by 8.1%. On Visdrone dataset, the AdaZoom achieves 40.33% and \( AP_{50} \) of 66.94%, outperforming the SOTA performance by a large margin. On DOTA dataset, we follow the experimental setting as [42] to evaluate the detection performance in large scenes. We achieve 37.8% of \( AP \) and 63.5% of \( AP_{50} \), with ResNeXt-101 as backbone.
Comparison with pseudo annotation based region generation methods. For such methods, pseudo annotations of focus regions are produced with the object annotations. Typically, ClusDet [42] trains a region generation network by supervised learning based on pseudo generated annotations. For fair comparison, we use the Faster R-CNN with the same backbone with the ClusDet. On Visdrone dataset, our method outperforms ClusDet by 4.16% in AP even without collaborative training. It is reasonable since the region selection is difficult to be formulated as a supervised learning problem. Assigning the hard labels as true or false to focus regions is semantically vague. In contrast, we design a continuous reward to measure whether a generated region is good or bad. The RL formulation is more suitable for such problem. In addition, by integrating collaborative training, our method outperforms ClusDet in AP by 5.18% on VisDrone2019 and 5.9% on UAVVDT dataset. On DOTA [41] dataset, we set the number of focus regions of AdaZoom to 3. It should be noted from Table. 4 that our method achieves consistent improvement with stronger backbone.

Comparison with coarse-to-fine detection methods. For such methods, the object distribution is estimated based on a coarse-level preview detection, such as DREN [44], CRENet [40] and GLSAN [8]. Compared to CRENet with backbone of Hourglass-104, Our AdaZoom with Faster RCNN achieves 2.49% higher in AP on VisDrone2019 with ResNet-50 backbone. Compared to DREN [44] and ResNet101, the AP is increased by 4.5% with ResNet-50 backbone and 2.4% with ResNet-101 backbone on UAVVDT dataset. Compared with GLSAN [8], which uses extra super-resolution network for enlarging regions, our method outperforms it by 5.49% of AP on VisDrone2019. The performance of the coarse-to-fine methods is limited by the initial detection, which would cause the bias for region generation (i.e., small objects are easily missed in coarse-level detection). In contrast, our AdaZoom cooperates well with detector. The performance of both AdaZoom and detector are improved by collaborative training.

4.5. Ablation Study
There are two main components for AdaZoom network, such as adaptive generation of focus regions and collaborative training between AdaZoom and the detector. We evaluate the effects of different components on VisDrone2019 test-dev dataset. We use Faster R-CNN with ResNet50 for the ablation study, as shown in Table.5. SR denotes focus regions with adaptive scales of $[240^2, 350^2, 420^2]$ and adaptive aspect ratios of $[0.7, 1.0, 1.5]$. CT denotes collaborative training. The baseline uses the single scale of $350^2$ and aspect ratio of 1.0 without collaborative training, achieving 29.74% in AP.

Effect of scale/ratio adaptation. The scale and ratio adaptation of focus regions achieves the dynamic multi-scale detection. The scale adaptation alleviate the problem of scale variation. The adaptation of aspect ratio improves the recall of objects with the limited number of regions. As Table.5 shows, SR stably improves the AP. In particular, compared with baseline, SR increases the flexibility and diversity in region generation which improves AP by 0.77%. Compared with the model with CT, SR further increased the AP by 1.23% and $AP_{50}$ by 1.95%. The results prove that SR improves the performance of detection by adaptively zooming the focus regions with flexible scales and aspect ratios. The detector benefits from adaptive-scale detection.

Effect of collaborate training. CT optimizes the detect-
tor with the focus regions generated by AdaZoom. Meanwhile, it guides AdaZoom to mine the fine-level regions for detector boosting with the reward based on detection results. Table 5 shows that CT gains consistent performance improvement under different settings. Compared to baseline, the CT slightly boosts the AP by 0.25% and AP75 by 0.38%, because the scale of regions keeps the same. The collaborate training loses the benefits from multi-scale detection. When combining with SR, CT gains performance improvement as 0.71% in AP and 1.00% in AP75. The results show that involving collaboration between AdaZoom and the detector further promotes the object detection.

**Effect of number of focus regions.** The AdaZoom is motivated to adaptively propose focus regions with small and densely distributed objects, for further zoomed detection. To evaluate this, we present the recall rates (i.e., the proportion of objects that are enclosed in the generated regions) for objects of the reinforcement-learning-based AdaZoom, as shown in Fig. 6 (left). For detailed analysis, we report the recall for small, medium, and large objects, respectively. As can be seen in Fig. 6 (left), when the number of focus regions is 7, the recall of small, medium, and large achieve 95.1%, 92.9%, and 87.0%, respectively. Recall of small objects increasing fastest at the beginning. Because AdaZoom pays more attention to small objects and it prefers to focus on the regions with small objects.

With detector of ResNet50-based Faster R-CNN, the AP and AR are presented in Fig. 6(right). When the number of regions is zero, the detector is trained on the generated regions and infers on original images. The inference time almost linearly increases w.r.t. the number of focus regions. Our method can easily achieve the balance between accuracy and efficiency by setting the number of focus regions.

**Qualitative evaluation.** We visualize the detection results with and without AdaZoom in Fig. 7. We also display the top-3 focus regions for each image and the detection results on the focus regions. It can be observed from Fig. 7 (b) that the AdaZoom generates small region for small objects and the shape of the region adapts to the distribution of objects. Comparing Fig. 7 (c) with (a), the small and dense objects are well detected by our method, especially in the focus regions. Our method can adaptively focus on the important regions and achieve higher recalls of the small and dense objects.

5. Conclusion

In this work, we propose an Adaptive Zoom (AdaZoom) network to zoom the focus regions with flexible scale and aspect ratio for multi-scale object detection. We optimize AdaZoom with policy gradient algorithm without additional annotations for focus regions. Moreover, we propose collaborative training to promote the coordination between AdaZoom and the detector which further improves the performance of detection. Without bells and whistles, our method achieves state-of-the-art on the VisDrone2019, UAVDT and DOTA datasets.
References

[1] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.

[2] Miriam Bellver, Xavier Giro-I-Nieto, Ferran Marques, and Jordi Torres. Hierarchical object detection with deep reinforcement learning. *Advances in Parallel Computing*, 31, 2016.

[3] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *European conference on computer vision*, pages 354–370. Springer, 2016.

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[5] Juan C. Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[6] Changrui Chen, Yu Zhang, Qingxuan Lv, Shuo Wei, Xiaorui Wang, Xin Sun, and Junyu Dong. Rrnet: A hybrid detector for object detection in drone-captured images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[7] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[8] Sutao Deng, Shuai Li, Ke Xie, Wenfeng Song, Xiao Liao, Aimin Hao, and Hong Qin. A global-local self-adaptive network for drone-view object detection. *IEEE Transactions on Image Processing*, 2020.

[9] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 370–386, 2018.

[10] Mingfei Gao, Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Dynamic zoom-in network for fast object detection in large images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6926–6935, 2018.

[11] Ross Girshick. Fast r-cnn. *Computer ence*, 2015.

[12] Abel Gonzalez-Garcia, Alexander Vezhnevets, and Vittorio Ferrari. An active search strategy for efficient object class detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3022–3031, 2015.

[13] Chaohui Guo, Bin Fan, Qian Zhang, Shining Xiang, and Chunjune Pan. Augfpn: Improving multi-scale feature learning for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12595–12604, 2020.

[14] Kota Hara, Ming-Yu Liu, Oncel Tuzel, and Amir-massoud Farahmand. Attentional network for visual object detection. *arXiv preprint arXiv:1702.01478*, 2017.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–16, 2014.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[17] Xiangyu Kong, Bo Xin, Yizhou Wang, and Gang Hou. Collaborative deep reinforcement learning for joint object search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1695–1704, 2017.

[18] Hei Law and Jia Deng. CornerNet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[19] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 190–191, 2020.

[20] Yuezhang Li, Katia Sycara, and Rahul Iyer. Object-sensitive deep reinforcement learning. *arXiv preprint arXiv:1809.06064*, 2018.

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[23] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[25] Yingjie Liu, Fengbao Yang, and Peng Hu. Small-object detection in uav-captured images via multi-branch parallel feature pyramid networks. *IEEE Access*, 8:145740–145750, 2020.

[26] Yongxi Lu, Tara Javidi, and Svetlana Lazebnik. Adaptive object detection using adjacency and zoom prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2351–2359, 2016.

[27] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch. *https://github.com/facebookresearch/maskrcnn-benchmark*, 2018. Accessed: [Insert date here].

[28] Stefan Mathe, Alekxis Pirinen, and Cristian Sminchisescu. Reinforcement learning for visual object detection. In *Pro-
ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016. 2

[29] George W McConkie and Keith Rayner. The span of the effective stimulus during a fixation in reading. *Perception & Psychophysics*, 17(6):578–586, 1975. 2

[30] Mahyar Najibi, Bharat Singh, and Larry S Davis. Autofocus: Efficient multi-scale inference. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9745–9755, 2019. 2, 5

[31] Alekis Firinen and Cristian Sminchisescu. Deep reinforcement learning of region proposal networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6945–6954, 2018. 2

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 4, 5

[34] Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, and Jian Sun. Object detection networks on convolutional feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1476–1481, 2017. 2

[35] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection snip. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3578–3587, 2018. 1, 2

[36] Bharat Singh, Mahyar Najibi, and Larry S Davis. Sniper: Efficient multi-scale training. In *Advances in neural information processing systems*, pages 9310–9320, 2018. 1, 2, 5

[37] Richard S. Sutton, David Mcallester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Submitted to Advances in Neural Information Processing Systems*, 12, 1999. 2, 3

[38] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12345–12354, 2020. 2

[39] Haoran Wang, Zexin Wang, Meixia Jia, Aijin Li, Tuo Feng, Wenhua Zhang, and Licheng Jiao. Spatial attention for multi-scale feature refinement for object detection. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6

[40] Yi Wang, Youlong Yang, and Xi Zhao. Object detection using clustering algorithm adaptive searching regions in aerial images. In *European Conference on Computer Vision*, pages 651–664. Springer, 2020. 6, 7

[41] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3974–3983, 2018. 2, 5, 7