ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

Aerial Small Target Detection Overview

Yifei Wang[a], Wenhua Cui[*], Ye Tao[b], and Tianwei Shi[c]

School of Computer and Software Engineering, University of Science and Technology Liaoning, Anshan, 114051, China

[a]wyf628@icloud.com, [*]cwh@systemteq.net, [b]taibeijack@163.com, [c]tianweiabbcc@163.com

Abstract

To further study the application of target detection technology in aerial remote sensing image, this paper first introduces the application background and development direction of aerial remote sensing image. Then the application and research status of small target in aerial remote sensing images are introduced. In addition, the small target detection based on traditional methods, the small target detection based on deep learning and the main means of optimization target detection at present are introduced, and the realization methods of each stage of target detection are briefly introduced. And research status. Finally, the deficiency and development prospect of aerial remote sensing image based on target detection and the significance of improving the accuracy and speed of the algorithm are summarized.

Keywords

Small Target Detection; Aerial Remote Sensing Images; Deep Learning.

1. Introduction

With the development of computer vision, target detection techniques are receiving more and more widespread attention. Target detection is an important and challenging problem in the field of computer vision. While great progress has been made in target detection in natural scenes, progress in the field of remote sensing target detection has been slow due to the lack of data from aerial scenes. Remote sensing target detection images specifically include remote sensing images acquired by using satellites as remote sensing platforms and remote sensing instruments loaded with satellites for earth observation, and aerial images acquired by using aircraft as remote sensing platforms and photographing various targets on the ground at a stable altitude at perigee. The aerial remote sensing images like Figure 1.

Figure 1. The aerial remote sensing images

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

Aerial photography images have a wide field of view, high impact, high clarity, high resolution, small area and high authenticity of geospatial information captured by UAVs. UAVs provide a remote sensing platform for aerial photography that is easier to transit, with less site restrictions on takeoff and landing, easy operation, and good safety and stability. The difference between aerial photography images and remote sensing images is that aerial photography images are high-resolution images, basically at the meter level or even the centimeter level; satellite remote sensing images are of various types from the centimeter level to the meter level and the ten-meter level. High-altitude aerial images have a small area, relatively close to the ground, and the images are clear and accurate and used in actual scenes such as aviation railway stations, large shopping malls and scenic spots.

2. Research Status

The section headings are in boldface capital and lowercase letters. Second level headings are typed as part of the succeeding paragraph (like the subsection heading of this paragraph). All manuscripts must be in English, also the table and figure texts, otherwise we cannot publish your paper. Please keep a second copy of your manuscript in your office.When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use the paper for the book or journal in question. When receiving the paper, we assume that the corresponding authors grant us the copyright to use.

To address the problems like pedestrians and other dense small targets with uneven distribution 2019 Yang et al [1] mainly conducted a study and proposed an end-to-end aerial target detection framework (ClusDet) that combines target clustering and detection. This framework contains three basic components: a cluster proposal network (CPNet) for target clustering, which generates cluster regions of targets, a scale estimation network for estimating the scale of target clusters, and a dedicated detection network (DetecNet) for target detection of cluster regions normalized to each scale. However, this approach adds additional networks, which complicates the whole structure and slows down the training time.In 2020 Li et al [2] did target detection in aerial images with the help of density maps. The proposed DMNet consists of three main steps: (1) density map generation network, (2) segmentation of the input map into foregrounds based on the density map, and (3) target detection using the generated foregrounds. Compared with ClusDet, DMNet only needs to train a simple density generation network instead of training two sub-networks (CPNet and ScaleNet), the structure is simpler and the detection performance is better. However, if a fixed number of cropping blocks is set, the problem of setting the density threshold and the quality of the density map affect the cropping results.

In the same year Wang et al [3] used a clustering algorithm to search for regions containing dense targets to solve the problems of small targets and scale variation in uneven distribution of aerial images. However, not every clustered region can bring performance improvement. Therefore, the detection speed is improved by calculating the difficulty value of each clustered region, mining difficult regions, and eliminating simple clustered regions. A Gaussian deflation function is then used to deflate the difficult clustering regions to reduce the difference between target scales. Compared with ClusDet and DMNet, the advantages of CRENet: (1) the computational resources are concentrated on the regions containing dense targets, which can improve the detection efficiency; (2) because the clustering regions have different sizes, the clustering algorithm is directly used to predict the clustering regions instead of the network, which can avoid the problems of setting anchors and overlapping clustering regions; (3) the difficulty value of each clustering region is calculated, and the The clustering region that cannot bring accuracy gain can improve the computational speed; (4) using Gaussian deflation function to reduce the difference between different image target scales. However, this method is divided into two stages with coarse to fine and cannot be end-to-end. 2021 Deng et al [4] proposed an end-to-end global-local adaptive network. It contains three main components: global-

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

local detection network (GLDN), adaptive region selection algorithm (SARSA) and local super-resolution network (LSRN). The method integrates a global-local fusion strategy into a progressively scale-varying network to perform more accurate detection. The advantages of the method are the proposed end-to-end detection framework that uses global contour information and local detail information to estimate the bounding boxes of the original downsampled and cropped images; the proposed adaptive region selection algorithm to extract very crowded regions and reduce the number of processed pixels on high-resolution images, and the adaptive capability to provide more potential value for local super-resolution and data enhancement that can improve performance. However, the method also suffers from the need to set the number of subregions and the tendency to overlook targets in the corners of the image.

Xu et al [5] proposed a novel adaptive scaling (AdaZoom) network for problems such as dense small targets and extreme scale variation, as a selective magnifier with flexible shape and focal length that can adaptively generate and deflate focused regions for accurate target detection in large scenes without additional annotation. His proposed strategy of co-training promotes the coordination of AdaZoom and detectors together with the same training and inference process. Although reinforcement learning seems to be omnipotent with significant effect improvement. However, it also has many drawbacks, such as poor reward function setting, low sampling efficiency, unstable training results, and difficult reproduction.

3. Small Target Detection based on Traditional Method

3.1 Optimization based on Multi-scale Training

Multi-scale is similar to the image pyramid in digital image processing. The input image is scaled to multiple scales, and each scale separately calculates the feature map and carries out subsequent detection. Although this method can improve the detection accuracy to a certain extent, it takes a lot of time because multiple scales are completely parallel. Multi-scale Training (MST)[6] usually refers to setting up several different image input scales. During Training, a Scale is randomly selected from multiple scales, and the input image is scaled to this Scale and sent to the network, which is a simple and effective method to improve multi-scale object detection. Although each iteration is of a single scale, each iteration is different, which increases the robustness of the network without increasing too much computation. During the test, in order to obtain more accurate detection results, the scale of the test picture can also be enlarged, such as 4 times, so as to avoid too many small objects. Multi-scale training is a very effective method, which enlarges the scale of small objects and increases the diversity of multi-scale objects. It can be directly embedded in multiple detection algorithms.

3.2 Optimization based on Data Enhancement

Random Image Clipping and Patching (RICAP) [7]for Deep Convolutional Neural Networks (CNN)[8]. It includes three data operation steps. First, four images are randomly selected from the training set. Second, the images are clipped separately. Third, patch the cropped image to create a new one. RICAP greatly increases the diversity of images and prevents the over-fitting of many parameters with deep CNN. This tag blending works in tag smoothing and prevents the endless pursuit of hard 0 and 1 probabilities in deep CNN using Softmax functions.

RICAP shares concepts with clipping, obfuscation, and label smoothing, and has the potential to overcome their shortcomings. Clipping obscures a subregion of the image, while RICAP generates a subregion of the image. At each training step, both changed the salient features of the image. However, masking only reduces the number of features available in each sample. Instead, the proposed RICAP patch image, so patching the entire area of the image produces features that help with training.

Blending uses an alpha-blend(that is, blending the intensity of pixels) while RICAP patches four cropped images, which can be seen as a spatial blending. By mixing the two images by alpha, the mixing produces pixel-level features that the original image would never produce, greatly increasing the variety of features CNN must learn and potentially disrupting training. In contrast, images patched

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

by the RICAP method always produce pixel-level features, and in addition to edge patches, the original image also produces pixel-level features. When the boundary position (w, H) is close to four coordinates, the clipping area becomes smaller and objects are occasionally not depicted. RICAP does not check whether the object is in the clipping region. Even if there are no objects in the clipping area, CNN learns other objects from other clipping areas and enjoys the benefits of label smoothing.

4. Small Target Detection based on Deep Learning

Convolutional neural network is the main detection method based on deep learning. Because it can extract deeper information by deepening the convolutional layer and has less strict requirements on the quality of input information, it can save some work required by feature engineering compared with machine learning. For different application scenarios, convolutional neural networks usually have more parameters.

4.1 Small Target Detection based on Cascade RCNN

In recent years, with the increase of computing resources, more and more networks use cascading thinking to balance the target miss rate and error rate. The idea of cascade has a long history and has been widely used in the field of target detection. It takes a coarse-to-fine approach to detection: filter out most simple background Windows with simple calculations, and then use complex Windows to deal with the more difficult ones. With the arrival of the era of deep learning, Cai et al. [9]proposed the classic network Cascade RCNN to continuously optimize the prediction results by cascading several detection networks based on different IoU thresholds. Later, Li et al.[10] extended the Cascade RCNN to further improve the performance of small target detection. Inspired by the idea of cascade, Liu et al.[11] proposed an asymptotic positioning strategy to improve the detection accuracy of pedestrian detection by constantly increasing the IoU threshold. In addition, literature [12-14] shows the application of cascade network in difficult target detection, which also improves the detection performance of small targets to a certain extent. The network of Cascade RCNN like figure2.

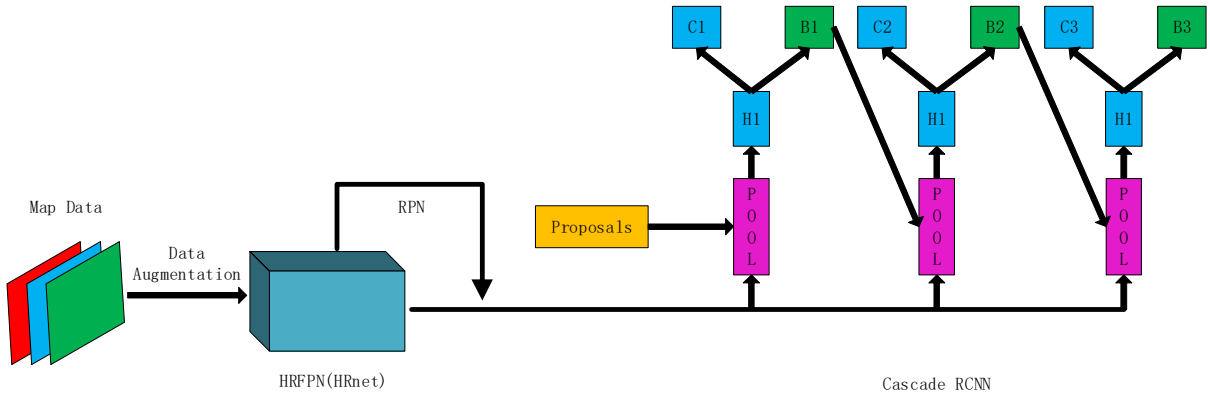

Figure 2. The network of Cascade RCNN

Cascade RCNN is a target detection algorithm based on deep learning that cascades several detection networks with different IOU thresholds on the basis of Faster RCNN. For an original input image, it first goes through a trunk network, namely the convolution layer, and extracts a series of candidate regions from the feature image extracted by the convolution layer through RPN. After passing through the detection network H1 with the threshold of 0.5, the output IOU is improved compared with the previous one after passing through the classifier and regressor. After resamping the corrected new candidate area, it is sent to detection network H2 with threshold value of 0.6, and then it is input to detection network H3 by analogization. Finally, the category and position obtained by H3 are the output result of the whole network.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

In the network structure of Cascade RCNN, I represents the input image, CONv represents the main network part such as ResNet, POOL represents the feature extraction pooling of ROI region, B0 represents the regional border information extracted by RPN, H represents the sub-network part after feature extraction, Each H is followed by C and B as the corresponding classifier and regressor respectively. The IOU threshold of each H part is different and gradually increases.

As the cascade structure of multi-threshold detection sub-network is adopted, the cascade sequence is arranged according to the threshold value from small to large. When the threshold value is 0.5, the number of positive samples is enough to ensure that the model will not overfit and ensure accuracy. In addition, when the threshold is 0.5, the correction effect of low IOU sample position is the most significant, and when the corrected high IOU is corrected by the regressor with a higher threshold, the accuracy of position can be greatly improved.

Through the cascade mechanism, the contradiction between the insufficient number of positive samples of high threshold and the decrease of correction effect of low threshold for high IOU region is solved in the traditional single network setting threshold. Through the cascade structure, the contradiction of network threshold setting can be avoided by a single module, which greatly improves the accuracy of target detection.

4.2 Nonmaximal Inhibition

Input : $\mathcal{B} = \{b_1, .., b_N\}, \mathcal{S} = \{s_1, .., s_N\}, N_t$
$\mathcal{B}$ is the list of initial detection boxes
$\mathcal{S}$ contains corresponding detection scores
N_t is the NMS threshold

begin
  $\mathcal{D} \leftarrow \{\}$
  while $\mathcal{B} \neq empty$ **do**
    $m \leftarrow$ argmax $\mathcal{S}$
    $\mathcal{M} \leftarrow b_m$
    $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{M}; \mathcal{B} \leftarrow \mathcal{B} - \mathcal{M}$
    for b_i *in* $\mathcal{B}$ **do**
      if $iou(\mathcal{M}, b_i) \geq N_t$ **then**
        $\mathcal{B} \leftarrow \mathcal{B} - b_i; \mathcal{S} \leftarrow \mathcal{S} - s_i$
      end NMS
      $s_i \leftarrow s_i f(iou(\mathcal{M}, b_i))$ Soft-NMS
    end
  end
  return $\mathcal{D}, \mathcal{S}$
end

Figure 3. Soft-NMS algorithm

Soft-nms[15]Improves the NMS algorithm by replacing a line in its code. The NMS algorithm is relatively simple, and all the boxes whose IOU is greater than the threshold are discarded. Unlike the OPERATION of the NMS algorithm, soft-NMS does not directly discard boxes that are larger than the threshold value, but reduces their score. There are also differences in the handling of checkboxes by reducing their scores rather than directly deleting high-value checkboxes.

Soft-nms can be easily used in the target detection algorithm, the code is simple and does not add extra calculation, easy to integrate; Soft-nms algorithm is used only in the reasoning process. NMS is a specialization of soft-NMS, and is interlinked when using binarization functions, whereas soft-NMS is more general.

In addition, there is Softer NMS in THE NMS strategy. As the starting point of OU - Net, the scores of two-stage detectors using NMS strategy are only classification scores, which cannot reflect the positioning accuracy of bounding box. Therefore, the Softer NMS algorithm is proposed. Since the implementation of Softer-NMS is a little complicated and requires network retraining to predict the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

confidence of the coordinates of the four coordinates (X1, X2, Y1,Y2) and then use this confidence as the weight weighting, soft-NMS algorithm is considered in this paper. The Soft-NMS algorithm like figure 3.

Object overlap is caused by the fact that some objects in the output boxes are actually another object, but also accidentally removed by NMS. The solution to this problem ultimately comes down to the "delete a candidate box" step, and we need to find a way to delete the box in S more carefully, rather than violently delete all and highest score boxes. For the box with the highest score box greater than the threshold value, it is not to remove it directly, but to reduce its confidence, so that more boxes can be retained, so as to avoid overlap to a certain extent.

If you just lower the confidence, you may not be able to remove multiple boxes that represent the same object. When the same object around the box there are a lot of Soft - NMS each choose the box score the highest, inhibit the surrounding box, box its ious in the highest score, the greater the degree of inhibition, in general, said the same object frame's IoU is will is bigger than another box of ious, therefore, like other objects the box will be preserved, And the same object is removed from the box.

5. Conclusion

For distant small target detection performance has increased, on his way to the distance, the benchmark model has a lot of undetectable phenomenon, the difficulties of the small target detection is usually targeted small, low precision, and the difficulty in the uav image is more, including the influence of shooting Angle, different weather and time lead to different images and bright degree, In addition, a large number of objects block each other in the UAV image, and there are only a part of the target in the image.

Acknowledgments

This research is supported by Iot and AI Innovation Team Liaoning(601009889-04), Joint fund project of National Natural Science Foundation of China(U1908218), and Natural Science Foundation project of Liaoning Province (2021-KF-12-06).

References

[1] Yang F, Fan H , Chu P , et al. Clustered Object Detection in Aerial Images[J]. IEEE, 2020.

[2] Li C, Yang T , Zhu S , et al. Density Map Guided Object Detection in Aerial Images[J]. 2020.

[3] Wang Y, Yang Y, X Zhao. Object Detection Using Clustering Algorithm Adaptive Searching Regions in Aerial Images[M]. 2020.

[4] Deng S, Li S, Xie K, et al. A Global-Local Self-Adaptive Network for Drone-View Object Detection[J]. IEEE Transactions on Image Processing, 2021.

[5] Xu J, Li Y, Wang S. AdaZoom: Adaptive Zoom Network for Multi-Scale Object Detection in Large Scenes[J]. 2021.

[6] Ding L, Zhang J, Bruzzone L. Semantic Segmentation of Large-Size VHR Remote Sensing Images Using a Two-Stage Multiscale Training Architecture[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, PP(99):1-10.

[7] Mario, De, Vincenzi, et al. The Proceedings of the Third Roma International Conference on Astroparticle Physics (RICAP'11)[J]. Nuclear Instruments & Methods in Physics Research, 2012.

[8] Ketkar N. Convolutional Neural Networks[J]. Springer International Publishing, 2017.

[9] Cai Z, Vasconcelos N. Cascade R-CNN: Delving into High Quality Object Detection[J]. 2017.

[10] Li A, Yang X, Zhang C. Rethinking Classification and Localization for Cascade R-CNN[J]. 2019.

[11] Liu W, Liao S, Hu W, et al. Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting[J]. Springer, Cham, 2018.

[12] Yang B, Yan J, Lei Z, et al. CRAFT Objects from Images[J]. IEEE, 2016.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202207_8(7).0008

[13] Fan Y, Choi W, Lin Y. Exploit All the Layers: Fast and Accurate CNN Object Detector with Scale Dependent Pooling and Cascaded Rejection Classifiers[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016.

[14] Gao M, Yu R, Li A, et al. Dynamic Zoom-in Network for Fast Object Detection in Large Images[J]. IEEE, 2018.

[15] Bodla N, Singh B, Chellappa R, et al. Soft-NMS -- Improving Object Detection With One Line of Code[J]. 2017.