An efficient backbone network for target detection in aerial images

Sihan Zhang\textsuperscript{1,a}, Yaping Liu\textsuperscript{2,b}, Xile Wang\textsuperscript{1,c} and Miaohui Zhang\textsuperscript{1,3,d*}

\textsuperscript{1} School of Artificial Intelligence, Henan University, Kaifeng, Henan, 475004, China
\textsuperscript{2} School of Mathematics and Statistics, Henan University, Kaifeng, Henan, 475004, China
\textsuperscript{3} Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, Henan, 475004, China
\textsuperscript{a}email: 104754190915@henu.edu.cn, \textsuperscript{b}email: 2694364314@qq.com, \textsuperscript{c}email: wangxile208@163.com
\textsuperscript{*} Corresponding author: \textsuperscript{d}email: zhmh@henu.edu.cn

Abstract. Target detection in aerial images by high altitude unmanned aerial vehicles is one of the hot research topics. A detection that is efficient and accurate is of a high value in the military and civilian fields. However, due to the irregular distribution of targets and their various scales shown in aerial images, it is difficult for existing backbone networks to effectively extract target features based on deep learning. To address these problems, in the paper, an efficient backbone network called AerNet is proposed, to which a local feature enhancement module (LFEM) is added to fully extract discriminative features of aerial targets via multi-scale convolutional layers. The network AerNet consists of 55 layers, which can keep high spatial resolution in deeper layers and maintain a large receptive field. Experimental results show that our AerNet has achieved a satisfying detection performance on the DOTA benchmark.

1. Introduction

Due to its small size, strong mobility and low safety factor, unmanned aerial vehicle (UAV) is often used in the military field in the early days. In recent years, with the gradual maturity of UAV technology, the improvement of intelligent operation and the reduction of manufacturing cost, the civil UAV market has developed rapidly. At the same time, thanks to the fast development of computer vision technology, intelligent UAV has begun to play an increasingly important role in civil fields such as traffic monitoring\cite{1}, fishery management\cite{2}, and more.

As the deep convolution neural network (CNN) develops rapidly, the aerial target detection, in which the backbone network plays the key part, has achieved remarkable results. VGGNet\cite{3} showed that increasing the depth of neural networks is a simple way to improve model performance. ResNet\cite{4} built a novel architecture through residual blocks and skip connection to avoid the optimization issues of deep networks. DenseNet\cite{5} adopted a dense connection strategy which can significantly reduce the amount of computations and improve the learning and representational capacity of deep network. GoogleNet\cite{6} used an inception module which increases the width of network to acquire diverse features. There are still many researches on efficient backbones, such as MobileNets\cite{7} and ShuffleNet\cite{8}. However, they are usually proposed to reduce parameters and computations. The above backbone networks work well.
in natural scenes, but their results in aerial scenes are not satisfactory. In this paper, we propose the AerNet, an efficient backbone designed for aerial target detection. Specifically, to address the problems brought by irregular distribution of targets in aerial images, AerNet includes the local feature enhancement module (LFEM) which can effectively extract and cover multi-scale instance targets. For traditional backbone networks, large down-sampling factors are often used, leading to a sharp decrease in the resolution of deep feature maps. As a result, small targets become invisible. The proposed backbone network, however, consists of 55 layers, which keeps the spatial resolution of the deep feature map.

In summary, the main contribution of this work lies in the following aspects:
1. An efficient backbone named AerNet is proposed for aerial target detection. It can effectively extract features of instance targets at various scales, maintain a high spatial resolution, and offer appropriate receptive fields.
2. By comparing with other backbones, we demonstrate that AerNet delivers a better performance on DOTA benchmark.

2. Materials and Methods
In this section, we will introduce the LFEM which uses a few small filters to obtain more discriminative feature from the original convolution layer, and build the high-efficient 55-layer backbone architecture AerNet.

2.1. Local Feature Enhancement Module for Discriminative Features
The targets to be detected are always unevenly distributed and shown in various scales in aerial images. Traditional backbone networks, using the single-sized convolution kernel when extracting features, usually ignores the differences in scale features of the receptive field. As a consequence, the target instances can not be completely covered, resulting in false detection and missed detection. In response to this issue, as shown in Figure 1, the efficient LFEM is designed to extract more distinguishable features instead of single-sized filter, without incurring additional computational cost.

The LFEM consists of three branches, which are convolution kernels of different sizes arranged in parallel. As shown in Figure 1 (a), the three branches are respectively constituted by 1×1 convolution kernel, 3×3 convolution kernel and 5×5 convolution kernel, and then the fused features are obtained by element-wise addition on the same feature dimension. Convolution with different-sized filters can extract features of different scales, which means the final classification result is more accurate. It is worth noting that the use of 1×1 convolution kernel can not only increase the depth and width of the network, but also adjust the spatial size of output feature maps, which can fuse features of different scales in any dimension. At the same time, considering that increasing the network width will lead to massive computational costs, we adopt the GhostNet[9] to improve the computing efficiency of backbone network. Where LFEM is applied with stride, we make two simple modifications: (1) Adding 3×3 average pooling on 1×1 convolution layer; (2) adding economical 3×3 depthwise convolution[7], which makes it easy to change the channel dimension to match the 1×1 convolution path. The improved architecture is depicted in Figure 1(b).

![Figure 1](image_url)  
**Figure 1** The proposed LFEM. (a) LFEM with stride = 1; (b) LFEM with stride = 2; (c) LFEM with dilation operation.
In general, the spatial resolution of the feature map will reduce exponentially after each stage of convolution. Commonly speaking, decreasing the spatial resolution of the feature map can enlarge the receptive field of network, and reducing the size of the feature map can reduce the computational cost and convolution operation. However, these measures will also weaken the information of the detected target, which is not conducive to small target detection. Therefore, in order to achieve a good effect without decreasing the size of the feature map, as shown in Figure 1(c), the dilated convolution[10] is added to the LFEM to keep the spatial resolution of the feature map.

2.2. Building Efficient AerNet
Taking advantages of the LFEM, we design the efficient backbone network AerNet as described in Table 1. As can be seen from Figure 1, the LFEM looks like the basic residual block in ResNet, where several convolution layers are integrated. The process of AerNet is divided into five stages. Note that in the stages two to four, except the first LFEM where stride=2, the other all employ the LFEM whose stride=1, and for the next stage the output channels are doubled. In order to maintain the spatial resolution of the deep feature map, and retain more spatial information and small target information, we use the LFEM with dilation convolution in the fifth stage. In addition, considering the computation and memory consumption caused by the dilation convolution, we keep the number of channels same as before, which will not significantly increase the parameter of the backbone network. Furthermore, we apply batch normalization (BN) and ReLU nonlinearity after each convolution layer in the LFEM. Finally, the global average pooling and 1×1 convolution are used to transform the feature maps to 1000-dimensional feature vectors for classification.

The design of AerNet efficiently deals with the difficulties in aerial image target detection. Specifically, AerNet uses the LFEM to cover and extract the features of detected targets. In the fifth stage, the LFEM with dilation convolution is applied to maintain the spatial resolution of deep feature map and increase the receptive field, which makes the detection of small targets and the localization of large targets more accurate. Most importantly, AerNet can be easily integrated with current mainstream detectors.

| Layer Name | Output Size | Operator | Stride |
|------------|-------------|----------|--------|
| Conv 1     | 112 × 112 × 32 | Conv 7× 7 | 2 |
| Conv 2-x   | 56 × 56 × 64 | [LFEM]×1 | 2 |
|            | 56 × 56 × 64 | [LFEM]×2 | 1 |
| Conv 3-x   | 28 × 28 × 512 | [LFEM]×1 | 2 |
|            | 28 × 28 × 512 | [LFEM]×3 | 1 |
| Conv 4-x   | 14 × 14 × 1024 | [LFEM]×1 | 2 |
|            | 14 × 14 × 1024 | [LFEM]×5 | 1 |
| Conv 5-x   | 14 × 14 × 256 | [LFEM with dilation operation] × 3 | 1 |
|            | 14 × 14 × 256 |           | |
| Fully connected layer | 1 × 1 × 1000 | Avgpool14 × 14, Conv1 × 1, Softmax | —— |

3. Results & Discussion
In this section, the implement of the experiment will be firstly introduced, after which the feasibility of the AerNet and LFEM is verified and the feature maps of the LFEM are visualized. Next, we integrate the AerNet with mainstream detectors on the DOTA benchmark to further demonstrate its effectiveness. At last, the detection results are displayed.
3.1. Implementation Details
We mainly evaluate our backbone on the DOTA dataset. The AerNet is end-to-end trained on 2 NVIDIA Titan X GPUs of 12GB memory based on Ubuntu 16.04 by PyTorch 1.5. In addition, the weight decay and momentum of stochastic gradient descent are set to 0.0001 and 0.9, respectively. The AerNet is trained for 200 epochs, with the initial learning rate being 0.0025 and divided by 10 every 40 epochs. We apply COCO metric to evaluate our approach.

| Methods          | Backbone   | mAP  | AP50 | AP75 | AP5 | APM | APL |
|------------------|------------|------|------|------|-----|-----|-----|
| Faster R-CNN[11] | ResNet-50  | 43.3 | 58.9 | 45.1 | 29.3| 42.7| 49.6|
| Faster R-CNN     | ResNet-101 | 44.1 | 59.6 | 45.4 | 30.7| 43.2| 51.1|
| RetinaNet[12]    | ResNet-101 | 40.6 | 55.4 | 41.8 | 26.4| 39.8| 47.2|
| YOLOv3[13]       | DarkNet-53 | 30.9 | 52.1 | 32.8 | 21.3| 34.9| 44.9|
| Cascade R-CNN[14]| ResNet-50  | 45.7 | 60.4 | 47.2 | 32.2| 45.1| 53.6|
| Faster R-CNN     | AerNet     | 51.4 | 62.3 | 49.7 | 37.5| 46.2| 55.3|

3.2. The AerNet Analysis
We use Faster R-CNN with ResNet-50 backbone as the baseline. In order to demonstrate the effectiveness of AerNet, we utilize AerNet instead of ResNet-50 backbone and retain the consistency of other structures. As can be seen from Table 2, using AerNet as the backbone can obtain a significant improvement (up to 8.1% mAP) over the baseline. Meanwhile, our proposed backbone improves the detection performance of the baseline by 8.2%, 3.5%, 5.7%, respectively, on AP5, APM, APL. Compared with other methods, the detection performance is also improved.

The feature maps of AerNet and ResNet are visualized in Figure 2. It can be clearly seen that the features extracted by AerNet are more discriminative, which indicates AerNet is more suitable for complex aerial detection tasks.

![Figure 2](image)

3.3. Detection Results on DOTA Dataset
Some of the detection results are shown in Figure 3. It proves that our proposed AerNet backbone can effectively extract discriminative features to solve the problems of aerial target detection in complex scenes.
4. Conclusions
In this paper, the backbone AerNet is proposed for aerial image target detection. Different from traditional backbone structures, a plug-and-play local feature enhancement module is added, which can effectively extract the features of the detected target while maintaining the spatial resolution of the deep feature map. The suggested backbone network can detect aerial targets even if the target size changes greatly. The experimental results show that the backbone has a good performance even in complex aerial scenes.

Acknowledgments
This research was funded by the National Natural Science Foundation of China (62002100) and the Foundation of Henan Educational Committee (19A520002).

References
[1] Leitloff J, Hinz S and Stilla U 2010 Vehicle detection in very high resolution satellite images of city areas IEEE Trans. Geosci. Remote Sens. 48 2795–806
[2] Zhang S, Xin M, Wang X and Zhang M 2021 Anchor-free network with guided attention for ship detection in aerial imagery J. Appl. Remote Sens. 15
[3] Simonyan,Karen and Zisserman,Andrew 2015 Very deep convolutional networks for large-scale image recognition 3rd Int. Conf. Learn. Represent. ICLR 2015 - Conf. Track Proc.
[4] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 770–8
[5] Huang G, Liu Z, Van Der Maaten L and Weinberger K Q 2017 Densely connected convolutional networks Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1 2261–9
[6] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A and Zeiler M D 2015 Going deeper with convolutions Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 40 1–9
[7] Howard A G, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M and Adam H 2017 MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications Computer Vision and Pattern Recognition vol 14 pp 53–7
[8] Zhang X, Zhou X, Lin M and Sun J 2018 ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 6848–56
[9] Han K, Wang Y, Tian Q, Guo J, Xu C and Xu C 2020 GhostNet: More features from cheap
operations Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. 1577–86
[10] Yu F and Koltun V 2016 Multi-scale context aggregation by dilated convolutions 4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.
[11] Ren S, He K, Girshick R and Sun J 2017 Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks IEEE Trans. Pattern Anal. Mach. Intell. 39 1137–49
[12] Lin T Y, Goyal P, Girshick R, He K and Dollar P 2017 Focal Loss for Dense Object Detection Proc. IEEE Int. Conf. Comput. Vis. 2999–3007
[13] Redmon J and Farhadi A 2018 YOLOv3: An Incremental Improvement
[14] Cai Z and Vasconcelos N 2018 Cascade R-CNN: Delving into High Quality Object Detection Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 6154–62