Pedestrian Detection in Aerial Image Based on Convolutional Neural Network with Attention Mechanism and Multi-scale Prediction

Jiaxi Yang , Jiaquan Shen , Shitong Wang , Yuhang Chen , Qian Zhang , Sitao Luan *

Posted Date: 30 August 2024

doi: 10.20944/preprints202401.1672.v2

Keywords: pedestrian detection; aerial image; attention mechanism; multi-scale prediction; convolutional neural network; new benchmark dataset

Article

Pedestrian Detection in Aerial Image Based on Convolutional Neural Network with Attention Mechanism and Multi-scale Prediction

Jiaxi Yang [1], Jiaquan Shen [2], Shitong Wang [3], Qian Zhang [1], Yuhang Chen [1] and Sitao Luan [4]

[1] Department of Electrical and Computer, Concordia University, Montreal, QC H3B 0E5 Canada
[2] School of Information and Technology, Luoyang Normal University, Luoyang, Henan 471000 China
[3] School of Computer Science, Universiti Sains Malaysia, Gelugor, Penang 11800 Malaysia
[4] Department of Computer Science, McGill University, Montreal, QC H3B 0E5 Canada
* Correspondence: author: Jiaxi Yang (jiaxi.yang@mail.concordia.ca)

Abstract: Pedestrian object detection plays a significant role in intelligent systems such as intelligent traffic and monitoring. Traditional machine learning methods on pedestrian detection have shown various drawbacks, e.g., low accuracy, slow speed, etc. The Convolutional Neural Network (CNN) based object detection algorithms have demonstrated remarkable advantages in the field of pedestrian detection. However, the mainstream CNNs still face the problems of slow speed and low detection accuracy, especially on small and occluded targets from aerial perspective. In this paper, we propose Multi-Scale Attention YOLO (MSA-YOLO) detection algorithm to address the above issues. MSA-YOLO includes a Squeeze, Excitation and Cross Stage Partial (SECSP) channel attention module for CNNs to extract richer pedestrian features with a small number of extra parameters. It also contains a multi-scale prediction module to capture the information among different pedestrian scales, which can recognize the small objects with higher accuracy and significantly reduce the missed detection. To sufficiently evaluate our proposed model, we manually collect and annotate a new benchmark dataset, Aerial Pedestrian Dataset, which has much more sample annotations, features, scenes and image view angles than the existing benchmark datasets. In addition, the images in our dataset have higher resolution than most of benchmark pedestrian detection datasets, which can provide more detailed features of pedestrians and thus improve the model performance. When tested on Aerial Pedestrian Dataset, our proposed MSA-YOLO algorithm significantly outperforms the most used baseline models with almost the same model size. This shows the efficiency of our proposed model. (The code and new dataset will be released to the public later.)

Keywords: pedestrian detection; aerial image; attention mechanism; multi-scale prediction; convolutional neural network; new benchmark dataset

---

I. Index Terms Pedestrian Detection, Aerial Image, Attention Mechanism, Multi-Scale Prediction, Convolutional Neural Network, New Benchmark Dataset. Introduction

Pedestrian detection from an aerial perspective has abundant application scenarios [1]. For example, in the traffic field, the detection of the pedestrian can identify the residents who violate traffic regulations and enhance traffic safety issues [2]. In disaster relief missions, the use of pedestrian detection technology from an aerial viewpoint can assist rescue teams in quickly locating people who are trapped or in need of assistance [3].

Nowadays, there are two mainstream methods to solve detection tasks, one is traditional machine learning [4] and the other is deep learning [5]. The former approach consists of three phases) Determine the position and range of the objects in the image) Feature extractors such as HOG (Histogram of oriented gradients) are used to extract features [5]; 3) Support Vector Machine (SVM) [6] is used to classify objects according to the extracted features. This algorithm is based on appearance features, which uses the contour information of pedestrians for classification and recognition. Since pedestrian images have different scales and spatial randomness, such detection method has low accuracy and efficiency.

 doi:10.20944/preprints202401.1672.v2

Since 2006, deep learning [7] has revolutionized lots of areas [8]-[12], including object detection [13]. Two different methods, one-stage and two-stage algorithms, are used in deep learning for object detection. The one-stage approach utilizes features extracted by convolutional neural networks for classification and bounding box regression, and it is relatively fast in detection, e.g., SSD [14], YOLO [13], RetinaNet [15], etc. The two-stage method, which has higher detection accuracy, takes much higher computational cost [16] than the one-stage method. It begins by using the Region Proposal Network (RPN) [17] to extract the objective region and a Convolutional Neural Network (CNN) is used to categorize and identify the candidate region. R-CNN [18] and Faster R- CNN [17] belong to this category. The two-stage methods have good robustness and higher detection accuracy, but the model size and inference time will far exceed those of the single- stage YOLO algorithm.

Nowadays, deep learning based pedestrian detection algorithms still face the low detection accuracy challenge [19] and there are two main reasons for it. The first reason is that the size of the detected pedestrians varies widely in the image. In particular, the size of the pedestrians in some regions of the image will be relatively tiny. This makes the model vulnerable to the detection of small targets. The second reason is that the objects in the pictures are often accompanied by messy background information, e.g., being covered by buildings, trees and other pedestrians, which leads to lots of missed detection. Besides, we also lack high-quality benchmark dataset with sufficient aerial perspective images to train the detection model.

To address the above problems, in this paper, we propose MSA-YOLO deep learning algorithm. It has the Squeeze, Excitation and Cross Stage Partial (SECSP) channel attention mechanism module, which can concentrate more on the layer of features that provide the most information and exclude the information from the less significant aspects, so that the accuracy of the detected objects can be improved. In addition, MSA-YOLO includes a multi-scale prediction module to increase the capacity on the recognition of relatively small objects in the images and decreases the rate of missed detection. These two proposed methods effectively address the problems caused by the inconspicuous pedestrian features in the images and the small size of pedestrians. To sufficiently train and evaluate our model, we collect and annotate a new benchmark dataset, Aerial Pedestrian Dataset. After comparison with baseline models on Aerial Pedestrian Dataset, we found that MSA-YOLO significantly outperform the baselines 2.3% without adding much computational cost.

In summary, the main contributions of this paper are as follows:

- We proposed the Squeeze, Excitation and Cross Stage Partial (SECSP) channel attention module, which can extract the feature more accurately and effectively.
- Then, we propose a multi-scale prediction module, which can capture multi-scale information for small and occluded pedestrians.
- To assess the pedestrian detection models, we created a new dataset, the Aerial Pedestrian Dataset, which contains 1200 aerial images with approximately 22800 labeled samples. The advantages of our proposed dataset are the richness of image samples, high image resolution, complexity of the scene. And compared to the currently existing pedestrian detection datasets, the camera angle we used is main from aerial view, which is unique and can fill the gap in the current pedestrian detection datasets.

The paper will be organized as follows: In Section II, we introduce current research on deep learning approaches to pedestrian detection, covering innovative strategies for facing occlusion detection and attention mechanisms; in Section III, we introduce the principle and structure of One-stage Neural Network and SENet in detail, which are important skeletons for our proposed model; in Section IV, we propose the SECSP Attention Module and Multi-scale Prediction Module and name our proposed algorithm as MSA-YOLO in order to enhance the feature extraction capability for small and occluded pedestrians; in Section V, we introduce our own dataset as well as the current publicly available dataset, and evaluate the capabilities of MSA-YOLO with each of these datasets. Besides, we conduct ablation study to verify the effectiveness of our proposed method; in Section VI, we summarize the merits of MSA-YOLO and Aerial Pedestrian Dataset.

 doi:10.20944/preprints202401.1672.v2

II. Related Work

A. Deep Learning-Based Pedestrian Detection

The authors in [20] propose an approach for detecting occluded pedestrians. It enhances visible pedestrian areas while suppressing occluded pedestrians by modifying full- body characteristics. Additionally, they describe the occlusion- sensitive hard sample mining strategy, which prioritizes detection failures in highly obstructed pedestrians by mining hard samples based on the degree of occlusion. To enhance pedestrian identification, Hsu et al. [21] suggest a brand-new stationary wavelet dilution residual super-resolution (SWDR- SR) network. In order to better maintain boundary features and enhance pedestrian recognition, they also suggest a novel low- to-high frequency connection technique (L2HFC). SWDR-SR performs better in identifying small-sized pedestrian pictures compared to baseline methods. In [22], the authors describe a novel Pose-Embedding Network for pedestrian identification that combines the Pedestrian Recognition Network (PRN) and the Region Proposal Network (RPN). The functions of these two networks are to produce candidate regions, raise confidence levels, and get rid of false positives. The effective- ness of their proposed method in comparison to the state-of- the-art (SOTA) method was demonstrated using the Caltech, CityPersons, and COPpersons datasets. In [23], the authors propose a new deep small-scale sense network for pedestrian detection, which can generate the proposed areas to detect small-scale pedestrians effectively. Additionally, they add a brand-new cross-entropy loss function to boost the loss contribution of minute pedestrians, which are challenging to detect. Their method shows outstanding detection performance on both the VIP pedestrian and the Caltech pedestrian datasets. In [24], the authors propose a new multi-scale network to detect pedestrians with the most suitable feature maps at a specific scale by matching their perceptual fields to the object size and introducing an adversarial hidden network to enhance the robustness. With a detection speed that is twice as quick as the original network, their technique reaches cutting-edge performance. Luo et al. [25] propose Sequential Attention- based Distinct Part Modeling to get higher classification and regression accuracies. And their proposed method improves the mean average precision against baseline models by a large margin when evaluated on Caltech and Citypersons datasets. Hsu et al. [26] propose the Ratio and Scale Aware YOLO (RSA-YOLO) strategy to address the issue of low detection accuracy due to the high small pedestrian ratio. In addition, they use intelligent segmentation to split the image into two local images to solve the problem of large differences in aspect ratio. In the results, the proposed method achieves superior performance on ETH and VOC 2012 comp4 datasets compared with baseline models.

B. Attention Mechanism for Pedestrian Detection

Du et al. [27] propose a synthetic aperture radar (SAR) object detection algorithm, and they enhance the network by including a multi-scale feature attention module (MFAM). By applying channel and spatial attention processes to the multi- scale feature maps, the MFAM can highlight the crucial in- formation and decrease the interference brought on by clutter. The efficacy of the proposed method has been validated by significant experimental results based on the measured SAR dataset. In [28], the authors suggest a gaze tracking technique that incorporates the local and global binocular spatial attention mechanisms (LBSAM and GBSAM, respectively) into a network model. The Gaze Capture dataset is used to validate the proposed strategy, and the results show that it performs significantly better compared to existing methods. Hu et al. [29] describe a hybrid attention method for lung cancer picture segmentation that combines a spatial attention mechanism and a channel attention mechanism. The hybrid attention module is applied in DenseNet convolutional neural network [30], and their proposed method improves 24.61% compared to the baseline method in lung tumor medical im- ages. In [31], the authors propose a residual channel attention module to suppress thin clouds in images and enhance ground scene details. The proposed method shows superiority against SOTA method in reconstructing rich ground scene details when tested on real and synthetic multi-cloud images. In [32], the authors introduce a new spatial pyramidal attention network (SPANet) that uses structural information and channel relations to better represent features. Experiments show that their proposed attention module has less parameters and is 2.3% higher in mAP compared with the baseline method. In [33], the authors introduce the Self-Attention Module (SAM) as part of the architecture of YOLOv3. When evaluated on the BDD100K and KITTI datasets,

 doi:10.20944/preprints202401.1672.v2

their proposed method shows approximately 2.6% mAP improvement compared to the original yolov3 network.

III. Preliminaries

In this section, we will introduce One-stage Object Detection Algorithm and Squeeze and Excitation Network, which are the two backbones of our proposed methods.

A. One-stage Object Detection Algorithm

One-stage algorithm is one of the most used neural network frameworks for object detection tasks. This one-stage network has the merits of fast detection speed and auto anchor, and it is favored in practical applications. As shown in Figure 1, this network consists of backbone, neck and head parts. These three parts play different roles separately: the backbone is to extract image features, the neck is to mix and combine features, and the head is to predict the results. The images will be sent to the backbone network at the initial step. If the images are not square, the border of the pictures will be filled with blank, and the size of the images will be resized to 640×640 pixels. Then, we can obtain a feature layer after each Cross Stage Partial (CSP) module [34] that can be enhanced to learn more features, for a total of four feature layers with sizes of 160 × 160, 80 × 80, 40 × 40 and 20 × 20. Then, after processing by the Spatial Pyramid Pooling-Fast (SPPF) module, feature map fusion of partial features and global features is achieved.

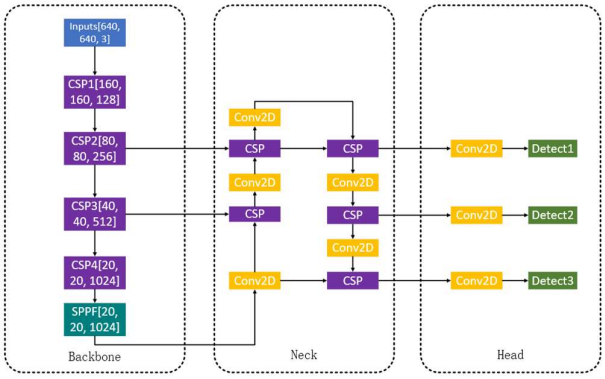

Figure 1. Structure of One-Stage Object Detection Algorithm.

In the next stage, as shown in Figure 1, the effective feature map output from the backbone part is delivered to the neck part of the network from CSP2, CSP3, and SPPF, respectively. The neck part is composed of Feature Pyramid Networks (FPN) [35] and Path Aggregation Network (PAN) structure [36]. The combination of these two structures can fuse the feature layers of different shapes to extract better features. Eventually, the three feature layers which are acquired in the neck part are fed into the head part and the results are output using the CIOU loss function [37] and the Non-Maximal Suppression (NMS) algorithm. Typically, one-stage network has three detection heads: if we input 640 × 640 size images, we can get 80×80 feature maps for detecting 8×8 size objects, 40×40 feature maps for detecting 16×16 size objects and 20×20 feature maps for 32×32 size objects.

B. Squeeze and Excitation Network

Squeeze and Excitation Network (SENet) is a convolutional neural network which incorporates the squeeze and excitation block, i.e., an attention module. The attention mechanism module only adds a tiny number of extra parameters, which has very little impact on training speed and enables the network to improve model accuracy by focusing more on features that are more important to the task at hand.

As the architecture shown in Figure 2, the input feature map $X \in R^{H' \times W' \times C'}$ is fed into the header in the attention mechanism module, where W', H' and C' stand for the feature

 doi:10.20944/preprints202401.1672.v2

width, height, and number of channels. Afterward, the output of feature layer U = {u1, u2, ..., uC} ∈ $R^{H\times W\times C}$ is produced by confounding the input feature map

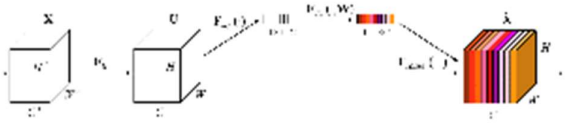

Figure 2. Squeeze and Excitation block.

$$U_C = V_C * X = \sum_{S=1}^{C} V_C^S * X^S \qquad (1)$$

where V_C is the learned filter kernel set and $*$ represents the convolution operation. The network can be made more sensitive to the information aspect with the above multi- channel convolution.

In the next step, the Squeeze operation is performed on the feature map U to turn the two-dimensional feature channel into a scalar number. In other words, the feature map U ∈ $R^{H\times W\times C}$ is converted into a 1×1×C output, which can have a global perceptual field to some extent. The formula is shown in Equation (2), where F_{sq} indicates the global average pooling. It not only reduces the number of parameters in the module but also avoids the negative effects of too many channels on model aggregation.

$$Z_C = F_{sq}(u_C) = \frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W} u_C(i,j) \qquad (2)$$

After we have obtained a feature layer of size 1×1×C, the feature layers will be fed into two fully-connected layers to learn an adaptive weight for each channel and thereby decide which channels are more important to focus on.

IV. The Proposed Method

In this section, inspired by the YOLO [38] network architecture and the SENet attention mechanism, we propose Multi- Scale Attention YOLO (MSA-YOLO), which contains the Squeeze, Excitation and Cross Stage Partial (SECSP) attention module and the multiscale prediction module.

A. SECSP Module

The neck of the YOLO network is made of FPN and PAN structures, however, FPN and PAN have limited feature extraction capabilities in complex scenes such as pedestrian targets that are occluded by obstacles. To enhance the feature extraction capability of FPN and PAN in complex environments and focus more on the essential features of the pedestrian objects, in this section, we introduced Squeeze, Excitation and Cross Stage Partial (SECSP) channel attention module.

The Squeeze and Excitation Network (SENet) is added to the PAN structure of the network, following by the CSP layer, so that the Occluded and small-sized pedestrian features in the image can be captured more effectively. As shown in Figure 3, in SECSP module, the input feature maps first flow through a Convolutional-BatchNorm-LeakyReLU (CBL) layer that contains convolutional operations to extract spatial features, batch normalization to stabilize the learning process and accelerate convergence, and a Leaky ReLU activation function to introduce nonlinearity and facilitate gradient propagation. The feature map is then processed through an additional convolutional layer, while at the same time a portion of the original input feature map goes directly into another convolutional layer. These two feature streams are processed through the convolutional layer and then merged at the Concat layer to integrate different levels of information. Subsequently, the merged feature maps are fed into the SENet module, where the features are recalibrated through the channel attention mechanism to highlight important information and suppress interfering information. This pipeline is demonstrated in Figure 3, which effectively improves the detection accuracy and model robustness.

 doi:10.20944/preprints202401.1672.v2

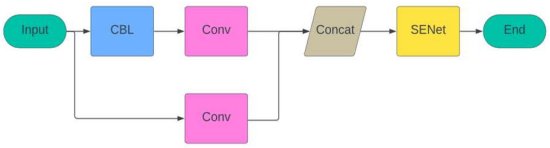

Figure 3. Structure of SECSP.

B. Implements Multi-Scale Prediction Module for Small Objects

The pedestrian detection task is unique in that the size of the pedestrians in the images varies widely. The detection results of YOLO convolutional neural network on the test set reveals that some pedestrians that occupy a relatively small portion of the image cannot be detected. To reduce the missed detection rate of the small-size pedestrian detection, we proposed Multi- Scale Attention YOLO (MSA-YOLO) pedestrian detection algorithm. The network structure is shown in Figure 4.

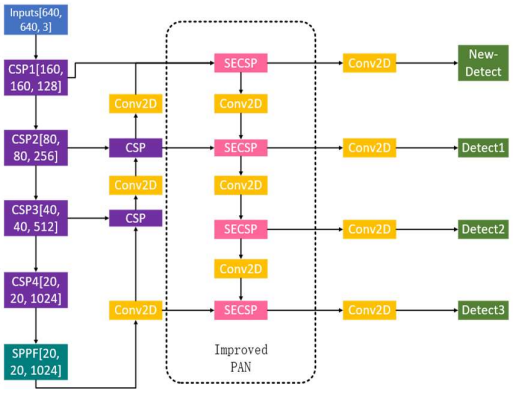

Figure 4. Structure of MSA-YOLO.

The original PAN can only output three effective feature layers that can provide three prediction scales. In addition, due to the large down-sampling multiplier and excessive perceptual field of the YOLO network, locating the feature information of small objects on deeper feature maps is quite challenging, therefore the effectiveness of small object detection is not satisfactory. To address the issues of insufficient precision and high miss detection rate brought by the scale discrepancies in the images, we introduced a multi-scale prediction module to the PAN structure. As shown in Figure 4, the new feature layer is created by merging feature maps from the backbone network's second and third CSP layers with the first CSP layer, post two CSP and convolution operations at the FPN. This process results in a 160×160 feature map, achieved through the Concat operation. The new feature layer is used to detect small objects of 4×4, which can effectively fuse the shallower feature maps with the deeper feature maps, thus enhancing the feature extraction capability and improving the detection accuracy of tiny targets. Although this method increases the computational cost and reduces the inference speed to some extent, the detection results are significantly improved, especially for the small targets, which are often missed by YOLO.

V. Experiments

A. Hardware, Software and Hyperparameters

In this paper, we used the Windows OS version of the PyTorch framework to build our model. The hardware device environment was an NVIDIA GeForce RTX 3090 with 24GB GPU and 64 GB of RAM with 2933MHz. During the training of the model, we set the hyperparameters the same as YOLO, with a learning rate of 0.01 and a weight decay rate of 0.0005, and SGD as optimizer.

B. Datasets

- Public Dataset

 doi:10.20944/preprints202401.1672.v2

The dataset we use is called the CrowdHuman dataset. Each image in this dataset contains an average of about 23 pedestrian samples. The images in this dataset were obtained by the authors from the Internet. Most of the images were taken from a human eye-level perspective, and only a small portion of the images were taken at a slightly overhead angle. The samples of this dataset are shown by Figure 5.

Figure 5. Samples of CrowdHuman Dataset.

- Aerial Pedestrian Dataset

We collected and created a new dataset, Aerial Pedestrian Dataset (APD), to evaluate the model. The samples of APD are demonstrated in Figure 6. Our dataset elevates pedestrian detection to new heights with its aerial perspective and a high resolution of 5472 × 3078 pixels, providing a level of detail unprecedented in current public datasets. In addition, our dataset contains a variety of scenes, such as plazas, streets, outdoor stadiums, and campus business districts. And the dataset contains images of extreme situations, such as pedestrians covered by umbrellas and pedestrians exposed to strong sunlight, which can lead to the lack of distinctive features of pedestrians. Most publicly available pedestrian detection datasets are collected from a human eye-level perspective. These datasets typically feature images with relatively low resolution and exhibit a limited diversity in samples [39]. However, our aerial dataset captures a wider scene, offering a richer array of samples for superior model training. This expansive dataset, with its abundance of annotated samples, is a robust resource for developing advanced detection models that require detailed environmental understanding and can handle complex, real-world scenarios with higher accuracy.

Figure 6. Samples of APD.

The image in our dataset is collected by DJI Mavic Air 2S drone with three shooting angles of 35 degrees down, 45 degrees down and 55 degrees down. The dataset is composed of abundant sample types, with a total of 1200 images and about 22800 labeled samples and it contains two heights of 7.5m and 10m. This not only enriches the sample types but also allows the network to learn more pedestrian features for better application to real-world scenarios, which further enhances the

 doi:10.20944/preprints202401.1672.v2

robustness of the models trained on this dataset. Since our dataset is collected from an aerial perspective, it provides a unique perspective for pedestrian detection that differs from existing benchmark datasets. A key feature of our dataset is the ability to evaluate the model's ability to detect small targets comprehensively, a challenge often overlooked in traditional datasets.

C. Results

(1) Comparison with other baseline models using the public dataset

In this section, we use the publicly available dataset Crowd Human Dataset to evaluate our proposed network. As we can conclude from Table 1, our proposed algorithm slightly outperforms YOLO and Fast R-CNN in terms of mean average precision values. And the performance of the average accuracy values of the method we use in this dataset is in line with that of the Faster R-CNN and our model size is much smaller than that of Faster R-CNN. As shown in Figure 7, we evaluate YOLO and our proposed MSA-YOLO network using publicly available datasets and as can be seen from the visualization results in Figure 7, our proposed algorithm, has a slightly lower missed detection rate and false detection rate than the YOLO algorithm. The green boxes in the picture represent missed and false detection targets, while the red boxes represent correct detection results.

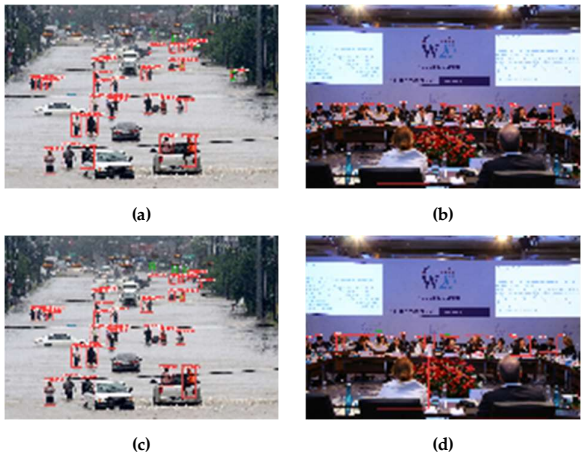

Figure 7. The visualization performance using CrowdHuman Dataset. (a) and (b) shows the detection results of YOLO, (c) and (d) shows the results of our proposed MSA-YOLO (Zoom up the figure to see the prediction confidence more clearly).

Table 1. COMPARISON RESULTS USING public dataset.

Methods	Backbone	mAP	Size
YOLO	Darknet	77.9%	40.2M
Fast R-CNN	VGG-16	76.8%	227.5M
Faster R-CNN	ResNet50	78.3%	337.1M
MSA-YOLO (Ours)	Darknet	78.3%	45.1M

(2) Comparison with other baseline models using the APD dataset

To compare with other baseline models, we test YOLO, Fast R-CNN [40] and Faster R-CNN [17] on APD dataset. The results are reported in Table 2. It can be observed that our proposed MSA-YOLO significantly outperforms YOLO with slightly larger model size and outperforms Fast R-CNN and Faster R-CNN with significantly smaller model size. This again shows the efficiency of our proposed model.

 doi:10.20944/preprints202401.1672.v2

Table 2. COMPARISON RESULTS using apd dataset.

Methods	Backbone	mAP	Size
YOLO	Darknet	94.7%	40.2M
Fast R-CNN	VGG-16	95.8%	227.5M
Faster R-CNN	ResNet50	96.6%	337.1M
MSA-YOLO (Ours)	Darknet	97.0%	45.1M

To visualize the advantages of MSA-YOLO against YOLO on detecting small objects, we show the detection results of YOLO and MSA-YOLO in Figure 8. The green rectangular boxes in the images represent detection errors and missed detection, and the red rectangular boxes represent correct detection results. We can find that the proposed MSA-YOLO has significantly reduced the missed detection of small objects and increased the prediction confidence of the correct detection.

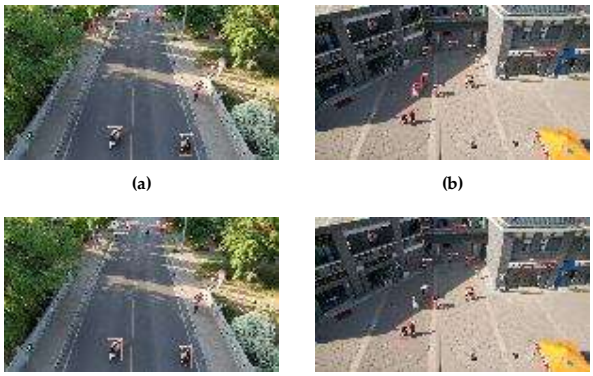

Figure 8. The visualization performance using Aerial Pedestrian Dataset. (a) and (b) shows the detection results of YOLO, (c) and (d) shows the results of our proposed MSA-YOLO (Zoom up the figure to see the prediction confidence more clearly).

The visualization of the MSA-YOLO outputs in more complicated and difficult scenarios are shown in Figure 9. The visualization results indicate that the proposed MSA-YOLO can still detect pedestrians in the places with low light intensity, obscured by foliage and in crowded square with low detection errors and high prediction confidence.

Figure 9. The visualization performance. (a) shows the detection results in a place with low light intensity and obscured by foliage, and (b) shows the detection results in a crowded square (Zoom up the figure to see the prediction confidence more clearly).

(3) Ablation Study

We conduct ablation experiments on Aerial Pedestrian Dataset to verify the effectiveness of the components in our new model. The average accuracy values for each category, i.e., the mAP values, are reported to assess our model. As shown in Table 3, the baseline YOLO model achieves an average accuracy value of 94.7%. When the attention module is added, the accuracy increases to 95.1% (↑ 0.4%) and when the multi-scale prediction module is included, the accuracy increases to 96.4% (↑ 1.7%). The performance enhancement of the two partial models shows the effectiveness of the attention and multi-scale prediction module. When both modules are added, the full model shows significant

 doi:10.20944/preprints202401.1672.v2

improvement with an average accuracy of 97.0% (↑ 2.3%), but the required memory is almost the same as the baseline YOLO. This shows the efficiency of our proposed model.

Table 3. ABLATION STUDY RESULTS.

Methods	SECSP	Multi-scale Prediction	mAP	Size
YOLO	× √ ×	× × √	94.7% 95.1% 96.4%	40.2M 40.5M 44.8M
MSA-YOLO (Ours)	√	√	97.0%	45.1M

VI. Conclusions

In this paper, we propose the MSA-YOLO detection algorithm which has a stronger and lightweight attention mechanism module, SECSP, for feature extraction. In addition, a multi-scale prediction module is added to the network for the detection of small-sized objects. The combination of these two modules leads us to the proposed MSA-YOLO.

Besides, we collect and build a new dataset, Aerial Pedestrian Dataset, which contains a great number of occluded pedestrian objects with various sizes. The ablation study, comparison with baseline and visualization of detection results on CrowdHuman Dataset and Aerial Pedestrian Dataset all show the efficiency of our proposed MSA-YOLO model.

Funding: This research was supported by the Key Scientific Research Project of Higher Education of Henan Province (No. 24A520025), and Henan Natural Science Foundation Youth Science Foundation Project (No. 232300420425), and the Henan Province Science and Technology Research Project (No. 222102210138, NO.232102220073, and No.222102110366), and the Science and Technology Innovation Team of Henan University (No. 22IRTSTHN016), and The Special project of key research and development Plan of Henan Province under Grant (No.221111111700).

References

1. S. V. A. Kumar, E. Yaghoubi, A. Das, B. S. Harish, and H. Proenca, 'The P-DESTRE: A Fully Annotated Dataset for Pedestrian Detection, Tracking, and Short/Long-Term Re-Identification From Aerial Devices', *IEEE Trans.Inform.Forensic Secur.*, vol. 16, pp. 1696–1708, 2021, doi: 10.1109/TIFS.2020.3040881.
2. Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, 'Traffic Flow Prediction With Big Data: A Deep Learning Approach', *IEEE Trans. Intell. Transport. Syst.*, pp. 1–9, 2014, doi: 10.1109/TITS.2014.2345663.
3. S. Sambolek and M. Ivasic-Kos, 'Automatic Person Detection in Search and Rescue Operations Using Deep CNN Detectors', *IEEE Access*, vol. 9, pp. 37905–37922, 2021, doi: 10.1109/ACCESS.2021.3063681.
4. M. Bilal and M. S. Hanif, 'Benchmark Revision for HOG-SVM Pedestrian Detector Through Reinvigorated Training and Evaluation Methodologies', *IEEE Trans. Intell. Transport. Syst.*, vol. 21, no. 3, pp. 1277–1287, Mar. 2020, doi: 10.1109/TITS.2019.2906132.
5. K. Dasgupta, A. Das, S. Das, U. Bhattacharya, and S. Yogamani, 'Spatio-Contextual Deep Network-Based Multimodal Pedestrian Detection for Autonomous Driving', *IEEE Trans. Intell. Transport. Syst.*, vol. 23, no. 9, pp. 15940–15950, Sep. 2022, doi: 10.1109/TITS.2022.3146575.
6. J. Platt, 'Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines', Apr. 1998, Accessed: Aug. 19, 2024. [Online]. Available: https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/
7. G. E. Hinton, S. Osindero, and Y.-W. Teh, 'A Fast Learning Algorithm for Deep Belief Nets', *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006, doi: 10.1162/neco.2006.18.7.1527.
8. A. Krizhevsky, I. Sutskever, and G. E. Hinton, 'ImageNet classification with deep convolutional neural networks', *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
9. A. Graves, A. Mohamed, and G. Hinton, 'Speech recognition with deep recurrent neural networks', in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 6645–6649. doi: 10.1109/ICASSP.2013.6638947.
10. D. Bahdanau, K. Cho, and Y. Bengio, 'Neural Machine Translation by Jointly Learning to Align and Translate', May 19, 2016, *arXiv*: arXiv:1409.0473. doi: 10.48550/arXiv.1409.0473.
11. D. Silver *et al.*, 'Mastering the game of Go with deep neural networks and tree search', *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016, doi: 10.1038/nature16961.
12. S. Luan *et al.*, 'When Do Graph Neural Networks Help with Node Classification? Investigating the Homophily Principle on Node Distinguishability', *Advances in Neural Information Processing Systems*, vol. 36,

pp. 28748–28760, Dec. 2023, Accessed: Aug. 19, 2024. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/5ba11de4c74548071899cf41dec078bf-Abstract-Conference.html

13. J. Redmon and A. Farhadi, 'YOLOv3: An Incremental Improvement', Apr. 08, 2018, *arXiv*: arXiv:1804.02767. doi: 10.48550/arXiv.1804.02767.
14. W. Liu *et al.*, 'SSD: Single Shot MultiBox Detector', in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 21–37. doi: 10.1007/978-3-319-46448-0_2.
15. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, 'Focal Loss for Dense Object Detection', 2017, pp. 2980–2988. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html
16. M. Carranza-García, J. Torres-Mateo, P. Lara-Benítez, and J. García-Gutiérrez, 'On the Performance of One-Stage and Two-Stage Object Detectors in Autonomous Vehicles Using Camera Data', *Remote Sensing*, vol. 13, no. 1, p. 89, Jan. 2021, doi: 10.3390/rs13010089.
17. 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 Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.
18. R. Girshick, J. Donahue, T. Darrell, and J. Malik, 'Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation', 2014, pp. 580–587. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html
19. S. Iftikhar, Z. Zhang, M. Asim, A. Muthanna, A. Koucheryavy, and A. A. Abd El-Latif, 'Deep Learning-Based Pedestrian Detection in Autonomous Vehicles: Substantial Issues and Challenges', *Electronics*, vol. 11, no. 21, p. 3551, Jan. 2022, doi: 10.3390/electronics11213551.
20. J. Xie, Y. Pang, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, 'Mask-Guided Attention Network and Occlusion-Sensitive Hard Example Mining for Occluded Pedestrian Detection', *IEEE Transactions on Image Processing*, vol. 30, pp. 3872–3884, 2021, doi: 10.1109/TIP.2020.3040854.
21. W.-Y. Hsu and P.-C. Chen, 'Pedestrian Detection Using Stationary Wavelet Dilated Residual Super-Resolution', *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–11, 2022, doi: 10.1109/TIM.2022.3142061.
22. Y. Jiao, H. Yao, and C. Xu, 'PEN: Pose-Embedding Network for Pedestrian Detection', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1150–1162, Mar. 2021, doi: 10.1109/TCSVT.2020.3000223.
23. B. Han, Y. Wang, Z. Yang, and X. Gao, 'Small-Scale Pedestrian Detection Based on Deep Neural Network', *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 3046–3055, Jul. 2020, doi: 10.1109/TITS.2019.2923752.
24. C. Lin, J. Lu, and J. Zhou, 'Multi-Grained Deep Feature Learning for Robust Pedestrian Detection', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3608–3621, Dec. 2019, doi: 10.1109/TCSVT.2018.2883558.
25. Y. Luo, C. Zhang, W. Lin, X. Yang, and J. Sun, 'Sequential Attention-Based Distinct Part Modeling for Balanced Pedestrian Detection', *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 15644–15654, Sep. 2022, doi: 10.1109/TITS.2022.3144359.
26. W.-Y. Hsu and W.-Y. Lin, 'Ratio-and-Scale-Aware YOLO for Pedestrian Detection', *IEEE Transactions on Image Processing*, vol. 30, pp. 934–947, 2021, doi: 10.1109/TIP.2020.3039574.
27. Y. Du, L. Du, and L. Li, 'An SAR Target Detector Based on Gradient Harmonized Mechanism and Attention Mechanism', *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3103378.
28. L. Dai, J. Liu, and Z. Ju, 'Binocular Feature Fusion and Spatial Attention Mechanism Based Gaze Tracking', *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 2, pp. 302–311, Apr. 2022, doi: 10.1109/THMS.2022.3145097.
29. H. Hu, Q. Li, Y. Zhao, and Y. Zhang, 'Parallel Deep Learning Algorithms With Hybrid Attention Mechanism for Image Segmentation of Lung Tumors', *IEEE Transactions on Industrial Informatics*, vol. 17, no. 4, pp. 2880–2889, Apr. 2021, doi: 10.1109/TII.2020.3022912.
30. G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, 'Densely Connected Convolutional Networks', 2017, pp. 4700–4708. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Huang_Densely_Connected_Convolutional_CVPR_2017_paper.html
31. X. Wen, Z. Pan, Y. Hu, and J. Liu, 'An Effective Network Integrating Residual Learning and Channel Attention Mechanism for Thin Cloud Removal', *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2022.3161062.

 doi:10.20944/preprints202401.1672.v2

32. X. Ma *et al.*, 'Spatial Pyramid Attention for Deep Convolutional Neural Networks', *IEEE Transactions on Multimedia*, vol. 23, pp. 3048–3058, 2021, doi: 10.1109/TMM.2021.3068576.
33. D. Tian *et al.*, 'SA-YOLOv3: An Efficient and Accurate Object Detector Using Self-Attention Mechanism for Autonomous Driving', *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4099–4110, May 2022, doi: 10.1109/TITS.2020.3041278.
34. C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, 'CSPNet: A New Backbone That Can Enhance Learning Capability of CNN', 2020, pp. 390–391. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_CVPRW_2020/html/w28/Wang_CSPNet_A_New_Backbone_That_Can_Enhance_Learning_Capability_of_CVPRW_2020_paper.html
35. T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, 'Feature Pyramid Networks for Object Detection', 2017, pp. 2117–2125. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/html/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.html
36. S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, 'Path Aggregation Network for Instance Segmentation', 2018, pp. 8759–8768. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Liu_Path_Aggregation_Network_CVPR_2018_paper.html
37. Z. Zheng *et al.*, 'Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation', *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8574–8586, Aug. 2022, doi: 10.1109/TCYB.2021.3095305.
38. A. M. Roy, R. Bose, and J. Bhaduri, 'A fast accurate fine-grain object detection model based on YOLOv4 deep neural network', *Neural Comput & Applic*, vol. 34, no. 5, pp. 3895–3921, Mar. 2022, doi: 10.1007/s00521-021-06651-x.
39. S. Zhang, R. Benenson, and B. Schiele, 'CityPersons: A Diverse Dataset for Pedestrian Detection', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 4457–4465. doi: 10.1109/CVPR.2017.474.
40. R. Girshick, 'Fast R-CNN', 2015, pp. 1440–1448. Accessed: Aug. 19, 2024. [Online]. Available: https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html

JIAXI YANG (Student Member, IEEE) received the B.E. degree in software engineering from Luoyang Normal University, Luoyang China, in 2023. He is currently pursuing the M.Eng. degree in electrical and computer engineering with the Concordia University. His research interests include computer vision, machine learning, the IoT, and signal processing.

JIAQUAN SHEN received the M.S. degree in Computer Science from Wenzhou University, in 2017, and the Ph.D. degree from Nanjing University of Aeronautics and Astronautics, in 2021. He is currently an Associate Professor with the School of Information Technology, Luoyang Normal University. His research interests include computer vision and object detection.

SHITONG WANG received the B.E. degree in software engineering from Luoyang Normal University, Luoyang, China, in 2023 . He is currently pursuing the M.S. degree in Computer Science at Universiti Sains Malaysia, Gelugor, Penang, Malaysia, focusing on computer vision, image processing and machine learning.

 doi:10.20944/preprints202401.1672.v2