An Improved YOLOv5-Based Method for UAV Object Detection

Authors: Han, Yunfei, Aierken, Abudoula, Wang, Yi, Ma, Yupeng

Date: 2023-03-23T00:00:00+00:00

Abstract

Object detection in unmanned aerial vehicle (UAV) imagery presents significant challenges due to the multi-scale nature and high density of objects within the UAV' s field of view. To comprehensively address these challenges and unlock the potential of UAV applications, we propose the YOLOv5-STD model. First, an additional detection head is incorporated to localize extremely small objects utilizing shallow image features. Second, a transformer-based attention mechanism is employed to optimize the backbone network. Third, SPD-Conv is utilized to prevent the loss of fine-grained image feature information. Finally, extensive experiments conducted on the VisDrone 2022 dataset demonstrate that the proposed model achieves superior performance. Compared with the baseline model, our improved model exhibits an average improvement of approximately 7% in mAP@.5 metrics, and ablation studies confirm that each enhancement contributes positively to the overall model performance. This work provides valuable insights for developers and researchers engaged in the analysis and processing of UAV imagery.

Full Text

Preamble

An Improved YOLOv5-Based Method for UAV Object Detection

Yunfei Hana,b, Abudoula Aierkena,b, Yi Wanga,b,∗, and Yupeng Maa,b

hanyf@ms.xjb.ac.cn, abdl@ms.xjb.ac.cn, wangyi@ms.xjb.ac.cn, ypma@ms.xjb.ac.cn

a The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi, China
b Xinjiang Laboratory of Minority Speech and Language Information Processing, Chinese Academy of Sciences, Urumqi, China

Abstract— Object detection based on unmanned aerial vehicle (UAV) images presents significant challenges due to the multi-scale nature and high density of objects in aerial views. To address these issues and unlock the potential of UAV applications, we propose the YOLOv5-STD model. First, we add an additional detection head to localize extremely small objects using shallow image features. Second, we employ a transformer-based attention mechanism to optimize the backbone. Third, we utilize SPD-Conv to prevent the loss of fine-grained image feature information. Extensive experiments on the VisDrone 2022 dataset demonstrate that our model achieves strong performance, with an average improvement of approximately 7% in mAP@.5 metrics compared to the baseline model. Ablation studies confirm that each improvement contributes positively to the overall model. This work can help developers and researchers achieve better performance in the analysis and processing of UAV imagery.

Keywords— object detection; yolov5; transformer; space-to-depth convolution

Introduction

Intelligent applications based on unmanned aerial vehicle image processing technology have been widely adopted across various industries, including intelligent monitoring, search and surveillance [1], intelligent rescue [2], infrastructure inspection [3], geographical mapping [4], and agricultural hazard prevention [5]. These applications all require comprehensive understanding of the scene environment within images. The fundamental challenge to be addressed is object detection, which involves recognizing object categories present in the scene and localizing their positions. Due to significant variations in UAV altitude, unique shooting angles and positions, and extensive coverage of both positive and negative samples in images, object detection in UAV imagery faces substantial difficulties such as small object sizes, diverse viewing angles, complex backgrounds, and high object density. These factors create a distinctive small object detection challenge that necessitates further research to enhance detection accuracy and UAV image understanding capabilities.

In recent years, with the rise of deep learning and the availability of large-scale labeled datasets (e.g., Pascal VOC [6] and COCO [7]), numerous state-of-the-art object detectors based on deep learning have been proposed and have achieved remarkable success in computer vision. Since Convolutional Neural Networks (CNNs) [8] were first successfully applied to object detection, Sermanet et al. introduced OverFeat [9], a one-stage CNN-based detection framework. To date, excellent deep learning-based detectors have been developed, including R-CNN [10], Fast R-CNN [11], YOLO [12-15], Faster R-CNN [16], SSD [17], R-FCN [18], RetinaNet [19], CornerNet [20], and CenterNet [21, 22].

Although these models have achieved promising results on public datasets, they remain inadequate for detecting small objects in UAV images. Consequently, many scholars have conducted research from various perspectives to address the small object detection challenge. Approaches include: multi-scale or multi-level

feature fusion to enhance small object feature acquisition [22-28]; context information and attention mechanisms to improve small object perception [29-32]; super-resolution techniques to enhance small object resolution and transform them into medium or large objects for detection [33-35]; cascaded multiple detectors [36, 37]; and methods based on image or feature patches for multi-detection fusion [38, 39].

In this paper, we propose YOLOv5-STD (YOLOv5 with Small Object Detection Head, Transformer, and Space-to-Depth Convolution module) based on YOLOv5, focusing on integrating multiple techniques that strengthen small object feature perception to improve accuracy. Specifically, we add an additional head for small object detection, incorporate a transformer [40] to optimize the backbone, and replace standard convolution with Space-to-Depth Convolution (SPD-Conv) [41] to explore prediction potential. Experimental results demonstrate that our method achieves significant improvements in small object detection on the VisDrone dataset and remains competitive compared to state-of-the-art methods. The overall YOLOv5-STD framework is illustrated in [Figure 1: see original paper] and will be introduced in detail in Section 3.

The contributions of this work are as follows: (1) We add an additional detection head for small objects that can localize them using shallow image features. (2) We employ an attention mechanism to optimize the backbone through transformers. (3) We utilize SPD-Conv to avoid the loss of fine-grained image feature information. (4) On the VisDrone 2022 dataset, our proposed YOLOv5-STD achieves 41.90% mAP. Experiments demonstrate that the fusion of multiple small object detection techniques is highly effective.

II. Related Work

A. Object Detectors

Most state-of-the-art object detectors have achieved breakthrough progress and excellent results on public datasets. As the first representative of two-stage object detection based on deep learning, R-CNN [10] automatically extracted features from candidate regions using CNNs. Faster R-CNN [16] established the fundamental framework for object detection, with many algorithms extending from it. Its introduction opened a new chapter in object detection, most notably by using CNNs to generate region proposals with anchor boxes throughout the inference process. However, two-stage detectors with numerous region proposals require substantial computational resources and runtime memory. In contrast, one-stage detectors such as the YOLO series [12-15], SSD [17], and RetinaNet [19] effectively alleviate inference efficiency issues by treating object detection directly as a regression problem that takes input images and learns category probabilities and bounding box coordinates. One-stage detectors typically achieve faster inference speeds than two-stage detectors. CenterNet introduced a novel anchor-free approach, predicting not only object category probabilities from image features but also bounding box keypoint coordinates directly, thereby

eliminating dependence on anchor boxes and advancing anchor-free object detection.

To date, YOLO series models have evolved to deliver excellent performance, representing exemplary work in the object detection field. In this study, we select YOLOv5 [15] as our baseline model, which comprises three components: backbone, neck, and head. The backbone uses convolutional neural networks to fully extract deep image features; the neck is designed to better utilize multi-level features extracted by the backbone; and the head predicts object classes and bounding boxes. YOLOv5 offers several advantages: it employs mosaic data augmentation to enrich small object samples and enhance detection robustness; in the backbone, YOLOv5 builds C3 layers by improving CSPBottleneck, achieving simplicity, speed, and lightness while delivering better results with nearly equivalent loss; in the neck, YOLOv5 utilizes the SPP module to fuse multi-scale features, expand the receptive field, and enrich feature map expressiveness, which proves beneficial when handling large object size variations. The YOLOv5 project features a clear architecture, rich engineering support functions, and easy, efficient deployment, making it an ideal choice for engineering projects. In summary, we select YOLOv5 as our baseline model for improving and optimizing UAV object detection.

B. Methods for Small Object Detection

Recent research has produced extensive work on small object detection optimization. Multi-level or multi-scale features are employed to enhance fine-grained feature representation of small objects. EfficientDet [26] proposed a weighted bi-directional pyramid network (BiFPN) that adds efficient bi-directional cross-scale connections and weighted feature fusion to the FPN network, enabling convenient and fast multi-scale feature fusion. Context and attention mechanisms are used to enhance small object perception. FA-SSD [31] uses feature fusion to obtain contextual information about small objects, extracting shallow features from small objects lacking semantic information, and employs an attention module to focus the network on important regions. Super-resolution techniques enable the transformation of small objects into larger ones. JCS-Net [33] investigates the relationship between large-scale and small-scale pedestrians using a super-resolution network that amplifies small objects through upsampling and recovers details of small-scale pedestrians to obtain enlarged objects. By combining super-resolution loss with classification loss, the reconstructed small-scale objects contain both original and super-resolution network output information. Cascade R-CNN [36] employs cascade regression as a resampling mechanism to progressively increase the IoU value of proposals, allowing resampled proposals from previous stages to adapt to subsequent stages with higher thresholds. MPFP-Net [39] slices features into patches and divides them into class-affiliated subsets, using bottom-up and crosswise connections to fuse multi-scale features for improved accuracy.

III. YOLOv5-STD

A. Overview

The basic YOLOv5 framework can be divided into four components: Input, Backbone, Neck, and Prediction. The Input component enriches the dataset through mosaic data augmentation, which requires minimal hardware and computational cost. However, this augmentation causes original small objects in the dataset to become even smaller, reducing model generalization performance. The Backbone primarily consists of CSP modules, with feature extraction performed by CSPDarknet53. The Neck employs FPN and Path Aggregation Network (PANet) to aggregate image features at this stage. Finally, the network performs object prediction and generates output predictions.

YOLOv5 has five versions: YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, with main differences in model depth and width. We modified the original YOLOv5 to specialize in small object detection. YOLOv5n is the smallest version in the YOLO series, making it more suitable for deployment across various hardware platforms with a simpler and clearer architecture. [Figure 1: see original paper] illustrates the framework of YOLOv5-STD.

B. Small Object Detection Head

We observed numerous small objects in the VisDrone 2022 dataset. As shown in [Figure 2: see original paper], objects smaller than 32$\times$32*butlargerthan*8$\times$8*constitutethelargestproportionat*5 account for 6.65%. To address the prevalence of small objects in drone imagery, adding a dedicated prediction head for small object detection better handles multi-scale object detection in UAV scenes overall. As illustrated in [Figure 1: see original paper], the new prediction head (Head 1) utilizes high-resolution image features combined with lower-level visual feature maps to perceive small objects more efficiently. Although adding a prediction head for extremely small objects increases computational complexity and resource consumption, it substantially improves detection performance.

C. Transformer

The transformer model, proposed by Google, represents a revolutionary advancement applicable not only to machine translation but also to text summarization, speech recognition, question-answering systems, dialogue systems, and computer vision. It employs a purely attention-based approach that accelerates training and inference while improving performance. The core concept is the self-attention mechanism, which automatically learns to focus on important information in input sequences, enabling interaction between information at different positions. The transformer' s self-attention mechanism uses multi-head attention to simultaneously focus on different subspaces of the input sequence, thereby enhancing model expressiveness. Compared to traditional RNN models, the transformer offers three key advantages: (1) it avoids sequential calculations in RNNs, enabling parallel processing of input sequences for much

faster training and inference; (2) it achieves information interaction across different sequence positions through self-attention, enhancing expressiveness and enabling handling of longer sequences; (3) it employs residual connections and layer normalization techniques, making the model more stable and easier to train.

Inspired by the vision transformer [40], we replaced the Bottlenecks in the last C3 blocks of the original YOLOv5 with transformer blocks. This modification enables the model to extract local features while using attention mechanisms to focus on regions containing small objects and explore feature representation potential through self-attention. Transformer encoder blocks demonstrate improved performance on occluded and high-density objects.

D. Space-to-Depth Convolution

Space-to-depth convolution is a specialized convolution operation widely used in deep learning for image processing tasks such as image classification, object detection, pose estimation, and motion recognition. It converts input image data from spatial dimensions to depth dimensions, thereby increasing network nonlinearity and sparsity while improving model representation capability and computational efficiency. Specifically, space-to-depth convolution divides input images into multiple blocks, arranges pixels from each block along the depth dimension, and combines these blocks into new image data. This operation reduces spatial dimensions while increasing depth dimensions, enabling the network to better detect objects of varying sizes and improving detection efficiency and accuracy. In this work, we replace most convolutions in YOLOv5 with space-to-depth convolution to better recognize multi-scale objects and achieve more accurate object detection in UAV imagery.

IV. Experiments

A. Datasets and Experiment Settings

The VisDrone2022-DET dataset, identical to the VisDrone2019-DET and VisDrone2018-DET datasets, comprises 7,019 static photographs captured by drone platforms at various locations and altitudes [18]. The test-dev set contains 1,610 images, while the training and validation sets contain 6,471 and 548 images, respectively. Images are annotated with bounding boxes across ten predefined classes: pedestrian, person, car, van, bus, truck, motor, bicycle, awning-tricycle, and tricycle. All models in this study are trained on the training set, validated on the validation set, and evaluated on the test-dev set. We present object detection performance on the test-dev set and compare it with baseline object detection models.

B. Implementation Details

We implement YOLOv5-STD using PyTorch 1.7.1 and train all models on an NVIDIA K80 GPU. During training, we utilize partial pre-trained weights from YOLOv5 (yolov5n, yolov5s, yolov5m, yolov5l, yolov5x) since YOLOv5-STD shares portions of the backbone and head with YOLOv5. We use the SGD optimizer with hyperparameters set as follows: initial learning rate of 1e-2, momentum of 0.98, weight decay of 0.001, warm-up epochs of 5, and warm-up momentum of 0.95. The Non-Maximum Suppression (NMS) threshold is set to 0.6 for all experiments. Input images are resized to 640$×$640 pixels. Batch sizes are set to 32, 16, 16, 8, and 4 for YOLOv5n-STD, YOLOv5s-STD, YOLOv5m-STD, YOLOv5l-STD, and YOLOv5x-STD, respectively. All models are trained on the VisDrone2022 training set for 500 epochs with early stopping patience of 100 epochs.

C. Comparison with Base Models

Due to submission limits on the VisDrone2022 competition server, we obtained results for five baseline models and five improved models on the testset-challenge. The results are presented in [TABLE:I]. As expected, larger models capture richer image features and achieve better detection performance. The improved models demonstrate approximately 5 percentage point improvements over baseline models in both mAP@.5 and mAP@.5:.95 metrics, indicating that the STD-based enhancements provide effective and stable improvements.

[TABLE:I]
BASE MODELS AND IMPROVED MODELS TEST RESULTS

Methods	mAP @.5(%)	mAP @.5:.95(%)
YOLOv5n	31.6	
YOLOv5s	36.1	
YOLOv5m	39.0	
YOLOv5l	40.4	
YOLOv5x	41.9	
YOLOv5n-STD		
YOLOv5s-STD		
YOLOv5m-STD		
YOLOv5l-STD		
YOLOv5x-STD		

D. Ablation Study

To thoroughly verify the effectiveness of the small object detection head, transformer, and space-to-depth convolution modules, we conducted ablation studies on these three major components. Leveraging the stability of STD and to reduce computational overhead while improving experimental efficiency, we

verified each module' s contribution using the minimal YOLOv5n model. By recombining the three modules and testing on the testset-challenge, we obtained the results shown in [TABLE:II], with detection result comparisons between YOLOv5n and YOLOv5n-STD models presented in [Figure 3: see original paper].

[TABLE:II]
TEST RESULTS OF IMPROVED MODELS BASED ON DIFFERENT COMBINATION METHODS

No.	Methods	mAP @.5(%)	mAP @.5:.95(%)
8	YOLOv5n-STD		

1) Effect of Small Object Detection Head: The new small object detection head utilizes high-resolution image features combined with lower-level visual feature maps to efficiently perceive small objects. Adding this head to YOLOv5n, YOLOv5n-spd, YOLOv5n-tr, and YOLOv5n-spd-tr resulted in average mAP@.5 improvements of approximately 5% and mAP@.5:.95 improvements of about 4% (see [TABLE:III]). The small object detection head provides the greatest positive impact on model performance, with its computational overhead and resource consumption being well justified.

[TABLE:III]
ABLATION STUDY ON SMALL OBJECT DETECTION HEAD

No.	Base Method	With Small Object Detection Head	mAP @.5(%)	mAP @.5:.95(%)
4	YOLOv5n-small YOLOv5n-spd-tr YOLOv5n-small-tr	YOLOv5n-STD		

2) Effect of Transformer: Transformer blocks can extract local features while using attention mechanisms to focus on regions containing small objects. Due to GPU memory limitations, we added only one transformer block to the last C3 block. We improved four models: YOLOv5n, YOLOv5n-spd, YOLOv5n-small, and YOLOv5n-small-spd. As shown in [TABLE:IV], the improved models achieved average mAP@.5 improvements of approximately 1% and mAP@.5:.95 improvements of about 0.5%. Although only a single transformer module was added, it still provides a slight positive impact, demonstrating the utility of transformer blocks in small object detection. More significant results could be achieved with larger GPU memory.

[TABLE:IV]
ABLATION STUDY ON TRANSFORMER

No.	Base Method	With Transformer	mAP @.5(%)	mAP @.5:.95(%)
1	YOLOv5n	YOLOv5n-tr	+0.9	
2	YOLOv5n-small	YOLOv5n-small-tr		
3	YOLOv5n-spd	YOLOv5n-spd-tr		
4	YOLOv5n-small-spd	YOLOv5n-STD		

3) Effect of Space-to-Depth Convolution: Space-to-depth convolution increases network nonlinearity and sparsity, enhancing representation ability and computational efficiency while better identifying multi-scale objects. We replaced most convolutions in four models—YOLOv5n, YOLOv5n-spd, YOLOv5n-small, and YOLOv5n-small-spd—with space-to-depth convolutions. As shown in [TABLE:V], improved models achieved average mAP@.5 improvements of approximately 2.5% and mAP@.5:.95 improvements of about 1.5%. Space-to-depth convolution effectively improves model performance while optimizing computational complexity, proving highly effective for small object detection.

[TABLE:V]
ABLATION STUDY ON SPD-CONV

No.	Base Method	With Space-to-Depth Convolution	mAP @.5(%)	mAP @.5:.95(%)
	YOLOv5n-spd			
3	YOLOv5n-tr	YOLOv5n-spd-tr		
4	YOLOv5n-small-tr	YOLOv5n-STD		

V. Conclusion

This paper proposes YOLOv5-STD, an improved small object detector based on YOLOv5. We incorporated several techniques to address small object detection challenges, including a small object detection head, transformer, and space-to-depth convolution. YOLOv5-STD demonstrates particular strength in object detection for UAV imagery. Extensive experiments on the VisDrone2022-DET dataset show that our improved models achieve superior detection results, confirming the feasibility of our approach. Furthermore, ablation studies thoroughly demonstrate that the small object detection head, transformer, and space-to-depth convolution each contribute positively to small object detection, validating the effectiveness and stability of our improvements. This work can help developers and researchers achieve better performance in the analysis and processing of UAV imagery.

Acknowledgment

This research was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region of China (2020D01B55) and the West Light Foundation of the Chinese Academy of Sciences (2019-XBQNXZ-B-009).

References

[1] Y. LI, Han, et al. Multi-block SSD based on small object detection for UAV railway scene surveillance[J]. Chinese Journal of Aeronautics, 2020, v.33;No.171(06):179-187.
[2] L. HONG, Y. WANG, Y. DU. Xin CHEN, Yujun ZHENG. UAV search-and-rescue planning using an adaptive memetic algorithm. Front. Inform. Technol. Electron. Eng, 2021, 22(11): 1477 149.
[3] H. Cheng, J. Yang. Solar Power Plant Maintenance with Thermal UAV Inspection Technology[J]. Power: The Magazine of Power Generation and Plant Energy Systems, 2022(6):166.
[4] H. Jia, L. Wang, D. Fan. The application of UAV LiDAR and tilt photography in the early identification of geo-hazards[J]. The Chinese Journal of Geological Hazard And Control, 2022, 32(2):60-65.
[5] Lnl A, Jls A, Yc B, et al. Using UAV-based thermal imagery to detect crop water status variability in cotton - ScienceDirect[J]. Smart Agricultural Technology, 2021.
[6] Everingham M, Gool L, Williams C K, et al. The Pascal Visual Object Classes (VOC) Challenge[J]. Springer US, 2010(2).
[7] Lin, T.-Y. et al. Microsoft COCO: Common Objects in Context. in Computer Vision –ECCV 2014 (eds. Fleet, D., Pajdla, T., Schiele, B. & Tuytelaars, T.) 740–755 (Springer International Publishing, 2014).
[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," Neural Information Processing Systems, pp. 1097–1105, 2012.
[9] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. Lecun, "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks," arXiv preprint arXiv:1312.6229, 2013.
[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ' 14), pp. 580-587, Columbus, Ohio, USA.
[11] R. Girshick, "Fast R-CNN," in Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV ' 15), pp. 1440–1448, Santiago, Chile, 2015.
[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, (CVPR ' 16), pp. 779–788, Las Vegas, Nev, USA, 2016.
[13] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in

Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, (CVPR ' 17), pp. 6517-6525, Honolulu, Hawaii, USA, 2017.
[14] J. Redmon, and A. Farhadi, "YOLOv3: an incremental improvement (2018)." arXiv preprint arXiv:1804.02767, 2018.
[15] Glenn Jocher, Alex Stoken, Ayush Chaurasia, et al., ultralytics/yolov5, https://github.com/ultralytics/yolov5, 2022.
[16] 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, 2017.
[17] W. Liu, D. Anguelov, D. Erhan, et al., "SSD: single shot multibox detector," in Proceedings of the Computer Vision –ECCV 2016, vol. 9905 of Lecture Notes in Computer Science, pp. 21–37, 2016.
[18] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks," arXiv preprint arXiv:1605.06409, 2016.
[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. ICCV, 2017.
[20] H. Law, J. Deng, "CenterNet: Keypoint Triplets for Object Detection" , European Conference on Computer Vision, 2018.
[21] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang and Q. Tian, "CenterNet: Keypoint Triplets for Object Detection" , arXiv preprint arXiv:1904.08189, 2019.
[22] X. Zhou, D. Wang, P. Krähenbühl, "Objects as Points" , arXiv preprint arXiv:1904.07850, 2019.
[23] G. Hu, Z. Yang, L. Hu, et al. Small object detection with multiscale features. International Journal of Digital Multimedia Broadcasting, 2018, 2018.
[24] S. Liu, L. Qi, H. Qin, et al. Path aggregation network for instance segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.
[25] Q. Zhao, T. Sheng, Y. Wang, et al. M2det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the AAAI conference on artificial intelligence. 2019, 33(01): 9259-9266.
[26] M. Tan, R. Pang, Q. Le. Efficientdet: Scalable and efficient object detection. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.
[27] X. Yu, Y. Gong, N. Jiang, et al. Scale match for tiny person detection. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2020: 1257-1265.
[28] Y. Gong, X. Yu, Y. Ding, et al. Effective Fusion Factor in FPN for Tiny Object Detection. Proceedings of the IEEE/CVF Winter Conference on Workshop on Applications of Computer Vision. 2021.
[29] L. Guan, Y. Wu, J. Zhao. Scan: Semantic context-aware network for accurate small object detection. International Journal of Computational Intelligence Systems, 2018, 11(1): 951-961.
[30] Y. Yuan, Z. Xiong, Q. Wang. VSSA-NET: Vertical spatial sequence attention network for traffic sign detection. IEEE transactions on image processing, 2019, 28(7): 3423-3434.

[31] J. S. Lim, M. Astrid, H. J. Yoon, et al. Small object detection using context and attention. 2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC). IEEE, 2021.
[32] N. Carion, F. Massa, G. Synnaeve, et al. End-to-end object detection with transformers, European conference on computer vision. Springer, Cham, 2020: 213-229.
[33] Y. Pang, J. Cao, J. Wang, et al. JCS-Net: Joint classification and super-resolution network for small-scale pedestrian detection in surveillance images. IEEE Transactions on Information Forensics and Security, 2019, 14(12): 3322-3331.
[34] Y. Bai, Y. Zhang, M. Ding, et al. SOD-MTGAN: Small Object Detection via Multi-Task Generative Adversarial Network, Proceedings of the European Conference on Computer Vision (ECCV), pp. 206-221, 2018.
[35] Y. Bai, Y. Zhang, M. Ding, et al. Finding Tiny Faces in the Wild With Generative Adversarial Network, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 21-30, 2018.
[36] Z. Cai, N. Vasconcelos. Cascade R-CNN: Delving into High Quality Object Detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 6154-6162.
[37] Q. Lin, Y. Ding, H. Xu, et al. ECASCADE-RCNN: Enhanced cascade RCNN for multi-scale object detection in UAV images[C]//2021 7th International Conference on Automation, Robotics, and Applications (ICARA). IEEE, 2021: 268-272.
[38] A. Wang, W. Li, X. Wu, et al. MPANet: Multi-Patch Attention For Infrared Small Target Object Detection. arXiv preprint arXiv:2206.02120, 2022.
[39] P. Shamsolmoali, J. Chanussot, M. Zareapoor, et al. Multi-patch feature pyramid network for weakly supervised object detection in optical remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-13.
[40] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv preprint arXiv:2010.11929, 2020.
[41] R. Sunkara, T. Luo. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. arXiv preprint arXiv:2208.03641, 2022.

Note: Figure translations are in progress. See original paper for figures.