MTR-YOLO Multiple Transformer-Enhanced YOLO for Object Detection in Multimodal Remote Sensing Imagery

Yaqian Liu[1], Zuheng Ming[2], Bo Zhang[1], Liang He[1*1], Kaixing Zhao[1]

[1]School of Software, Northwestern Polytechnical University, China
[2] Laboratoire L2TI, Institut Galile, University Sorbonne Paris Nord, France

ABSTRACT

Real-time detection of Remote Sensing Imagery (RSI) with a wide background and small targets is challeng- ing in various fields. Multimodal data fusion and enhanc- ing CNNs with Transformers can improve detection perfor- mance. The approach combines complementary information from different modalities and leverages CNNs' feature ex- traction capabilities. Transformers capture global informa- tion and learn sequence dependency without requiring large data samples. The goal is to achieve accurate and efficient target detection in applications such as fire detection, military reconnaissance, and autonomous obstacle avoidance.

We developed MSTR-Darknet, an improved backbone network for object detection in Remote Sensing Imagery. We removed the focal module in YOLOv5 to maintain high- resolution characteristics and achieve better performance by sacrificing a small amount of speed. We replaced the last layer with a STR module for improved connectivity with global information. We also used pixel-level fusion to ex- tract information from different modalities for more effective feature representation of small objects in RSI.

Secondly, in the multi-scale feature fusion stage, we de- signed Tini-BiFPN, a more effective weighted feature fusion network, for efficient cross-scale feature fusion. Given the ex- cellent contextual relationship integration capabilities of transformers, we also integrated transformer modules into the feature fusion network to identify attention regions in scen- arios with dense objects, leading to an improvement in mAP performance

Keywords: Multimodal fusion, Transformer, RS im- age Object detection

1. INTRODUCTION

Object detection technology is widely applied in various fields, including aerial photography, fast delivery, and urban monitoring. However, in remote sensing, there are specific challenges that make accurate object detection more difficult. The challenges include a small number of labeled samples,the small size of objects in remote sensing images (typically occupying only a few dozen pixels relative to the complex background)[1], and the diverse scale and multiple categories of objects[2]. These challenges pose a significant obstacle to general object detectors based on ordinary convolutional networks. Modern detectors typically use pure convolu- tional networks as feature extractors, such as VGG[3] and ResNet[4] backbones for detectors like Faster RCNN[5] and RetinaNet[6]. The YOLO series detectors[7], on the other hand, utilize Darknet, a novel residual network that improves feature extraction efficiency.

However,convolutional networks have limitations in cap- turing global contextual information due to the locality of convolution operations. In contrast, transformers excel at cap- turing inter-dependencies among image feature patches on a global scale through multi-head self-attention. This preserves spatial information for object detection. Additionally, object detection models need improved domain adaptability and dy- namic receptive field to handle viewpoint changes in aerial

[1*]2021050018@nwpu.edu.cn

The International Conference Optoelectronic Information and Optical Engineering (OIOE2024), edited by
Yang Yue, Lu Leng, Proc. of SPIE Vol. 13513, 135134J · © 2025 SPIE · 0277-786X · © The Authors.
Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.3056820

images.domain adaptability and dynamic receptive field. The study in literature [8] showed that compared with CNN, visual transformers have stronger robustness against severe oc- clusion, disturbance, and domain shift.To improve detection performance, transformer layers can be added to pure convo- lutional backbones to incorporate more contextual information and learn better feature representations.

Currently, most object detection techniques are solely de- signed and applied for a single modality such as RGB and Infrared (IR) [9], [10]. Consequently, with respect to object de- tection, its capability to recognize objects on the Earth's surface remains insufficient due to the deficiency of complemen- tary information between different modalities[11]. As imaging technology flourishes, RSIs collected from multimodality become available and provide an opportunity to improve the detection accuracy. For example, as shown in Figure.1, the fu- sion of two different multimodalities (RGB and IR)can effectively enhance the detection accuracy in RSI.

On the other hand, the size of objects in RSI images varies greatly, and the representation power of single-layer feature maps in convolutional neural networks is limited. Therefore, it is essential to effectively represent and process multi-scale features. A classical method is to combine low-level and high-level features through summation or concatenation op- erations, but simply summarizing or concatenating may lead to feature mismatch and performance degradation. In this regard, we introduce learnable weights to learn the impor- tance of different input features, while repeatedly applying top-down and bottom-up multi-scale feature fusion

In summary, this article proposes the following contribu- tions:

• Based on the original CSP-Darknet backbone, we introduced an improved backbone network, MSTR- Darknet, which not only removes the Focus module that hinders the definition of dense small targets but also introduces the STR multi-head self-attention block to bring more contextual information and learn more distinguishable feature representations.

• We propose a simple and efficient Tini-BiFPN structure with weighted bidirectional feature pyramid networks to reduce computational cost and parameters while enhancing multi-scale feature fusion and enriching semantic features.

• We not only incorpoate the STR attention mechanism into backbone also incorporate it into feature fusion to enhance the overall feature fusion ability of the net- work. Our exploration shows that Transformers can be flexibly used not only in feature extraction or detection heads but also in feature fusion stages with good per- formance

• We explore different fusion alternatives and choose the computation-friendly pixel-level fusion method for multimodal information combinations to further en- hance the detection accuracy. The proposed pixel-level efficiently decreases the computation cost compared with feature-level fusion.

2. RELATED WORKS

2.1. Transformer

Swim-Transformer It is an attention mechanism-based neu- ral network architecture designed for sequential data model- ing. Rather than relying on recurrent neural networks (RNNs) which can be computationally expensive, Swim-Transformer uses Transformer blocks with relative positional embeddings to effectively capture long-term dependencies in sequen- tial data. Compared to existing state-of-the-art models such as Faster R-CNN and Mask R-CNN, Swim-Transformer has shown better performance on several benchmark object detec- tion datasets including COCO and PASCAL VOC. It offers a promising alternative approach for modeling visual data with sequential structures.Cross-Transformer It introduces a self-attention mech- anism across modes, bringing the information from multiple modes together in the same model.The core idea of Cross- Transformer is to use self-attention mechanisms to capture relationships within sequence data as well as interactions across modalities. It computes the correlation within the mode by encoding each mode separately and then using the multi-mode attention mechanism. Next, in the cross-modal layer, the corresponding attention weights are used to fuse the information between the different modalities.Through self- attention computation of cross modes, Cross-Transformer can effectively mine the feature links between different modes to better understand and process multimodal data.

2.2. Muti-scale feature fusion

Multi-scale feature fusion is a computer vision technique uti- lized to enhance the performance of models in object detection tasks, particularly for detecting objects of different sizes. This technique involves the integration of features from various layers or levels of a convolutional neural network(CNN) to obtain a more complete and informative representation of an image. Different levels in a CNN capture different levels of abstraction and spatial resolution, which makes them better suited for detecting objects of different sizes.

For example, features obtained from lower layers have higher spatial resolution than those from higher layers, which can be useful for detecting small objects. Conversely, fea- tures from higher layers are better suited for detecting larger and more complex objects. By fusing these features at mul- tiple scales, the model can benefit from the strengths of each layer and achieve better overall object detection performance.

There are several ways to implement multi-scale feature fusion, including using skip connections to pass features between different layers, using feature pyramid networks (FPNs) to merge features across different resolutions, and using top-down and bottom-up pathways to aggregate infor- mation from different scales. These techniques have been proven successful in various computer vision tasks, such as object detection, semantic segmentation, and imageclassification.

3. METHODOLOGY

3.1. Overall Architecture

The proposed network architecture, called MSTR-YOLO(shown in Figure 1), is a hybrid model that combines convolution and self-attention. Firstly, we use STR-Darknet (Section3.2)as the backbone, which not only removes the Focus module but also integrates multi-head self-attention into the original CSP-Darknet to extract more Individualized features. Secondly, The Tini-BiFPN, which replaces PANet, is designed to aggregate features from different backbone levels .(Section 3.3) Lastly, we explore different fusion methods and select pixel-level fusion for high computational efficiency to fuse IR and RGB modes

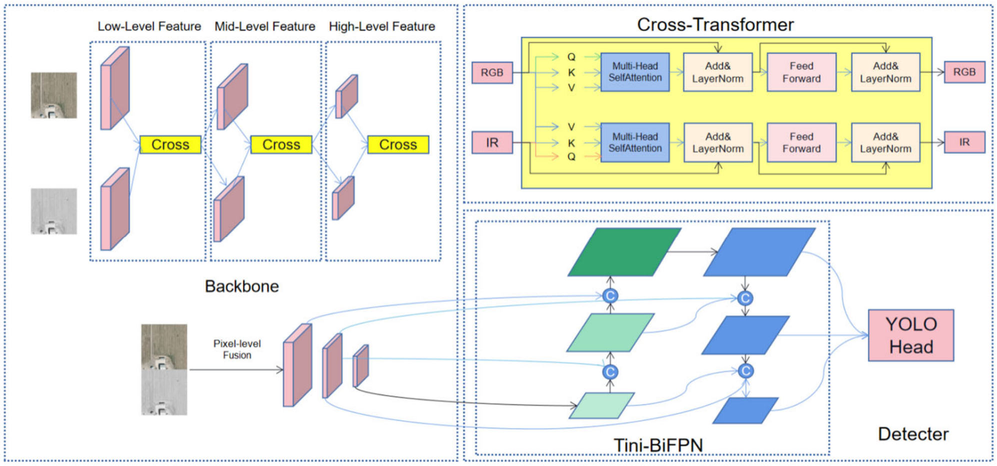

Figure. 1: The architecture of the MSTR-YOLOv5. a) MSTR-Darknet backbone with transformer encoder blocks at the end. b) The Neck use the structure Tini-BiFPN which combines the advantages of BiFPN and Transformer

3.2. STR-Darknet

The purpose of the Focus module in the YOLOv5 backbone is to gather the pixel values from the input image and then reconstruct them into smaller complementary images. The size of the reconstructed image decreases as the number of channels increases. Therefore, it will lead to a decrease in resolution and loss of spatial information for small targets. Considering that the Focus module in the YOLOv5 backbone was replaced with the CBS module to improve the detection of small targets, which relies on higher resolution.

To improve the semantic discriminability and mitigate class confusion for RSI in large-scale and complex scenes, collecting and correlating scene information from a large neighbourhood can help learn relationships between objects. However, convolutional networks have limitations in captur- ing global context information due to locality constraints of convolution operations.In contrast, transformers can globally attend to dependency relationships between image feature patches while preserving sufficient spatial information, en- abling multi-head self-attention based object detection. To enhance transferability of learned features and capture long- range contextual information, we propose the STR-Darknet backbone to extract features for detectors. The design of STR-Darknet is straightforward(as shown in Figure 3): we embed Swim-Transformer (STR) layers into the top CSPDark block to achieve global self-attention on 2D feature maps. It's worth noting that when the network is relatively shal- low and the feature maps are relatively large, early use of transformer layers to enforce boundary regression can lead to the loss of meaningful contextual information. Therefore,in STR-Darknet, transformer layers are only applied to P5 in- stead of P3 and P4, to avoid this issue.

3.3. Tini-BiFPN

The size differences of RSIs can be significant, and the rep- resentation ability of single-feature maps in conv-olutional neural networks is limited. Therefore, it is essential to ef- fectively represent and process multi-scale features. The traditional top-down FPN [16][12] is essentially limited by one-way information flow. To address this, PANet [13] adds an additional bottom-up path aggregation network, as shown in Figure 4(b). Further studies on cross-scale connections were conducted in [14][15][16]. In those works, a simple and efficient Weighted Bidirectional Feature Pyramid Network (BiFPN), as shown in Figure 4(c), achieves two optimizations for cross-scale connections.In this paper, Inspired by BiFPN, we have designed a lightweight neck network which we call Tini-BiFPN, it not only implifies BiFPN to fit the P5 structure but also introduce the Swim-Transformer into it.as shown in Figure 4(d) and Figure 5.

3.4. Multimodal Fusion

Multimodal fusion is an effective approach for integrating di- verse information from multiple sensors.The more informa- tion is utilized to distinguish objects, the better performance can be achieved in object detection. there are three promi- nent fusion methods: decision-level fusion,feature-level fu- sion, and pixel-level fusion. However, decision-level fusion, due to its high computational requirements, is not considered in this paper. Instead, we focus on describing our proposed feature-level fusion and pixel-level fusion techniques, which enable the integration of information at different processing depths within the network.

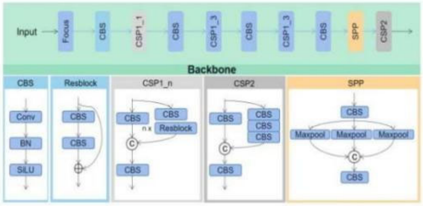

Figure. 2: The backbone structure of YOLOv5s. The low-level texture and high-level semantic features are extracted by stacked CSP, CBS, and SPP structures.

Figure 1 and Figure 2 illustrate show the feature-level fu- sion of various blocks works. To ensure a fair comparison, the IR image is expanded to three bands. Each block's fusion operation Cross-1,Cross-2,Cross-3, represent the fusion op- eration performed in the Low-Level, Mid-Level, High-Level, on the other hand, is considered to be a Feature-Level fusion operation.

When it comes to pixel-level fusion(RGB+IR), we nor- malize the input RGB and IR images to intervals of [0,1], then combine them with relatively low computational effort com- pared to the other fusion methods, which fuse the information during later procedures to speed up the inference. As Sec- tion will demonstrate, pixel-level fusion achieves better re- sults than feature-level fusion when combining different types of complementary information.

$$Q, K, V = \mathrm{Conv2D}(X, k = (1, 1)) \qquad (1)$$

when it comes to feature-level fusion ,we try to fuse three scale feature by cross transformer. In multi-head cross- atten- tion, we map the input patch sequence Xrgb to Q1 ,K1 ,V1 and Xir to Q2 ,K2 ,V2 in the headi (i = 1...h, and his the the num- ber of head), following the Q-K-V attention in transformer

As illustrated in Figure 4, the cross-attention layer plays a pivotal role in aggregating key information (K-V pairs) from two different branches to establish attention. Specifically, the K-V pairs from the two branches are concatenated to-gether. Similarly, in order to aggregate K-V pairs from the RGB branch to the IR branch, we also concatenate the K-V pairs from these two branches.

$$K_{cat} = [K_1 , K_2] \qquad (2)$$

$$V_{cat} = [V_1 , V_2] \qquad (3)$$

In this context, the operator [·, ·] refers to the concatenation along the token dimension as the default operation. The cross-attention feature map can be computed as:

$$\mathrm{Attencross}\,(Q_i , K_{cat} , V_{cat}) = \mathrm{softmax}(\frac{Q_i K_{cat}^T}{\sqrt{d_k}})V_{cat} \quad (4)$$

3.5. Loss Function

The loss function of our network consists of two parts: detec- tion lossLv and SR construction lossLs , which can be com-puted as :

$$L_{total} = \lambda\, 1\, L_o + \lambda 2\, L_s \qquad (5)$$

Where λ1 and λ2 refer to the coefficients used to balance two training tasks. In this case, the L1 loss (rather than L2 loss) is utilized to calculate the SR construction loss, denoted as Ls , between the input image X and the SR result S. This can be expressed as follows:

$$L_s = \| S - X \| 1 \qquad (6)$$

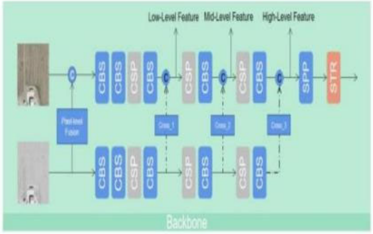

Figure. 3: The backbone structure of YOLOv5s. The low-level texture and high-level semantic features are extracted by stacked CSP, CBS, and SPP structures.

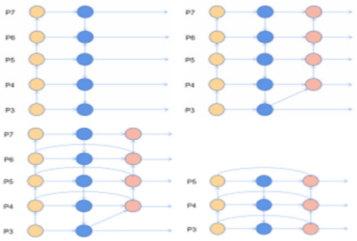

Figure. 4: Feature network design (a) FPN (b) PANet adds an additional bottom-up pathway on top ofFPN. (c) BiFPN implements two optimizations for cross-scale connections. (d) Tini-BiFPN simplifies BiFPN to fit the P5 structure and intro- duce Swim-Transformer into it

The detection loss consists of three components: the loss for determining whether there is an object (Lobj), the loss for object localization (Lloc), and the loss for object classification (Lcls). These components are used to eval-uate the loss of prediction, which can be expressed as follows:

$$Lo = \Sigma \lambda i \sum_{j=0}^{3} Xj\ Li\ , i \in (obj, loc,cls) X \in (a,b,c) \quad (7)$$

Here, in Equation 7, the variable j represents the layer of the output in the head. The weights aj, bj, and cj correspond to the weights assigned to different layers for the three loss func- tions. The weights λloc , λobj , and λcls are used to regulate the emphasis of errors among box coordinates, box dimen- sions, objectness, no-objectness, and classification.

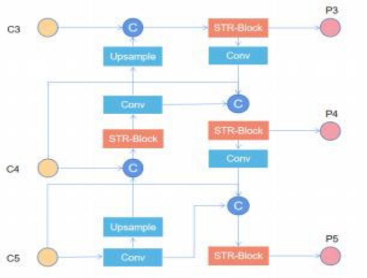

Figure. 5: Tini-BiFPN

4. EXPERIMENTS AND ANALYSIS

4.1. Datasets

The VEDAI dataset is used for multi-class vehicle detection in aerial images. It contains 3640 vehicle instances, including 9 categories such as ships, cars, campers, airplanes, shuttle buses, tractors, trucks, freight vehicles, and other categories. The dataset includes 1210 aerial images of size 1024x1024. with four uncompressed color channels, comprising three RGB color channels and one additional nearinfrared channel.

we use the VEDIA dataset to evaluate our model, and we report mAP (average of all 10 IoU thresholds, ranging from

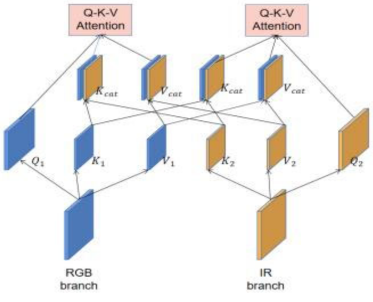

Figure. 6: The Cross-Attention in our cross-transformer feature backbone

Table 1: Distribution of Available Class Instances in the VEDAI Dataset Across 10 Folds.

Class	Tatal Instances	Distribution Across 10 Folds
car	1349	9 folds of 135; 1 fold of 134

pickup	941	9 fold of 94; 1 fold of 95
camping	390	10 folds of 39
truck	300	10 folds of 30
other	200	10 folds of 20
tractor	190	10 folds of 19
boat	170	10 folds of 17
van	100	10 folds of 10

[0.5 : 0.95]) and AP50.

4.2. Training Environment and Details

Our proposed framework is implemented in PyTorch and is executed on a workstation equipped with an NVIDIA 3090 GPU. The VEDAI dataset is utilized to train our MSTR- YOLO model.

For training, we employ the standard Stochastic Gradi- ent Descent (SGD) optimizer with a momentum of 0.937 and weight decay of 0.0005 for Nesterov accelerated gradients. The batch size is set to 2 (8). Initially, the learning rate is set to 0.01. The entire training process consists of 300 epochs and takes approximately 12 hours to complete.

4.3. Model Evaluation

The accuracy assessment evaluates the agreement and dis- crepancies between the detection results and the reference mask. To evaluate the performance of the methods being compared, the accuracy metrics used are recall, precision, and mAP (mean Average Precision). The calculation of precision and recall metrics is defined as follows:

$$\text{Precision} = \frac{TP}{TP+FP} \qquad (8)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP + FN}} \qquad (9)$$

TP is the count of correctly classified positive samples, FP is the count of incorrectly classified positive samples, and FN is the count of incorrectly classified negative samples. These metrics are used to evaluate the accuracy and performance of detection algorithms.

$$\text{mAP} = \frac{AP}{N} = \frac{\int_0^1 p(r)\,\mathrm{d}r}{N} \qquad (10)$$

The mAP (mean Average Precision) is a comprehensive in- dicator that averages AP values. It calculates the area under the Precision-Recall curve for all categories using an integral method. This metric quantifies the overall accuracy of an ob- ject detection model and considers the trade-off between pre- cision and recall.

4.4. Result Analysis and Ablation Experiments

We validate the effectiveness of our proposed method by con- ducting a series of ablation experiments on the first fold of the validation set. These experiments enable us to analyze the importance of each component in VDEIA (Visual Detection and Evaluation of Image Analysis).

4.4.1. Ablation of STR-Darknet

After integrating multi-head self attention into the original CSP-Darknet and remove the Focus module, the mean pre- cision and mean recall scores of this frameworks has been improved.

In particular, the mean recall score of YOLOv5s is im- proved by 21.99%(51.1% → 73.09%) Removing the Focus module not only prevents resolution degradation but also re- tains spatial interval information for small objects in RSI. Moreover, the utilization of self-attention has a significant im- pact on detecting small objects, which is widely recognized as an important and challenging task in real-world object sys- tems.

4.4.2. Ablation of Tini-BiFPN

To further validate the effectiveness of Tini-BiFPN,We com- pare the network with the original network and find that its accuracy is higher than that of the original network. The re- sults are shown in the TABLE 3.

Table 2: Ablation of STR-Darknet

Class	CSP-Darknet	STR-Darknet(ours)
car	92.17	88.90
pickup	86.22	87.12
camping	78.87	76.80
truck	76.91	81.01
other	54.26	64.62
tractor	80.12	83.22
boat	60.21	71.34
van	76.2	78.23
all	75.62	78.91

Table 3: Ablation of Tini-BiFPN

Class	PANet	Tini-FPN(ours)
car	91.64	88.90
pickup	86.77	87.12
camping	77.02	76.80
truck	82.35	81.01
other	67.53	64.62
tractor	79.64	83.22
boat	61.83	71.34
van	70.23	78.23
all	77.13	78.91

4.4.3. Ablation of Multimodal Fusion

After evaluating the devised fusion methods, we conducted experiments using pixel-level and feature-level fusion techniques, as described in Section 3.4 of the paper.

The results are presented in TABLE 5,6,7. The pixel-level fusion method achieved the best performance among all the compared methods, with a parameter size of 9.5084M and an mAP50 of 75.86%, respectively, which are the best among all the compared methods. Therefore, we choose the pixel-level fusion as our final fusion strategy, which exhibits relatively competitive performance for the VEDAI multimodal dataset with objects that are difficult to distinguish.

Table 4: Ablation of Multimodal Fusion

Class	RGB	IR	RGB+IR(ours)
car	89.93	82.72	88.90
pickup	82.86	77.43	87.12
camping	71.59	72.41	76.80
truck	75.35	67.95	81.01
other	67.64	39.65	64.62
tractor	80.56	60.15	83.22
boat	58.19	49.66	71.34

van	72.56	88.81	78.23
all	74.82	67.35	78.91

Table 5: The Comparison Result of Pixel-level and Feature- level Fusions in MTR-YOLO for Multimodal Dataset on the First Fold of the Validation Set.

Method	Parameters	mAP50
$Pixel-level Fusion$	9.5168M	78.91
$Cross\_Fusion1$	11.8873M	76.59
$Cross\_Fusion2$	11.8873M	76.20
$Cross\_Fusion3$	11.8873M	75.53

4.4.4. Comparisons with Previous Methods

The results clearly demonstrate that MSTR-YOLO outper- forms other frameworks, achieving higher AP and mAP50 scores. Notably, in multimodal mode, MSTR-YOLO sur- passes YOLOv5x by a significant 13.19% mAP50 score. The detection performance for boat, truck, van, and other cat- egories is notably improved in MSTR-YOLO compared to other methods.

5. CONCLUSION AND FUTURE WORK

In summary, this paper proposes a MSTR-YOLOv5 algo- rithm , realizing the organic combination of Transformer and CNN,meanwhile,achieving a balance of efficiency and per- formance. MSTR-YOLOv5 uses YOLOv5n-p5 as the base- line, STR Block to strengthen the connection between the backbone network and the global information, Tini-BiFPN to strengthen the network feature extraction and lighten the network. We also tried different methods of pixel-level fu- sion and feature-level fusion, and ultimately chose pixel-level fusion with better results.

In the future, We will continue to explore effective meth- ods of pixel-level fusion combined with feature-level fusion to improve detection performance. Exploring more possibili- ties for multimodal fusion.

Table 6: Class-wise Average Precision AP, Mean Average Precision mAP50, Parameters and GFLPs for Proposed MST-YOLO, YOLOv3, YOLOv4,YOLOv5s-x (IR modal ConFigureurations on VEDAI Dataset)

Methods	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	All	Params
YOLOv3	80.21	67.03	65.55	47.78	25.86	40.11	32.67	53.33	51.54	61.5351M
YOLOv4	80.45	67.88	68.84	53.66	30.02	44.23	25.40	51.41	52.75	52.5082M
YOLOv5s	77.31	65.27	66.47	51.56	25.87	42.36	21.88	48.88	49.94	7.0728M
YOLOv5m	79.23	67.32	65.43	51.75	26.66	44.28	26.64	56.14	52.19	21.0659M
YOLOv5l	80.14	68.57	65.37	53.45	30.33	45.59	27.24	61.87	54.06	46.6383M
YOLOv5x	79.01	66.72	65.93	58.49	31.39	41.38	31.58	58.98	54.18	87.2458M
SuperYOLO	87.90	81.39	76.90	61.56	39.39	60.56	46.08	71.00	65.60	4.8256M
Ours	88.63	81.69	76.31	67.26	44.86	67.86	44.77	71.08	68.43	9.5084 M

Table 7: Class-wise Average Precision AP, Mean Average Precision mAP50, Parameters and GFLPs for Proposed MST-YOLO, YOLOv3, YOLOv4,YOLOv5s-x (RGB modal ConFigureurations on VEDAI Dataset)

Methods	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	All	Params
YOLOv3	83.06	71.54	69.14	59.30	48.93	67.34	33.48	55.67	61.06	61.5351M
YOLOv4	83.73	73.43	71.17	59.09	51.66	65.86	34.28	60.32	62.43	52.5082M

YOLOv5s	80.07	68.01	66.12	51.52	46.78	66.69	36.24	49.87	58.80	7.0728M
YOLOv5m	81.14	70.26	65.53	53.98	46.78	66.69	36.24	49.87	58.80	21.0659M
YOLOv5l	81.36	71.70	68.25	57.45	45.77	70.68	35.89	55.42	60.81	46.6383M
YOLOv5x	81.66	72.23	68.29	59.07	48.47	66.01	39.15	61.85	62.09	87.2458M
SuperYOLO	90.30	82.66	76.69	68.55	53.86	79.48	58.08	70.30	72.49	4.8256M
Ours	91.18	82.75	69.50	59.48	57.51	72.64	57.49	57.49	75.86	9.5084 M

Table 8: Class-wise Average Precision AP, Mean Average Precision mAP50, Parameters and GFLPs for Proposed MST-YOLO, YOLOv3, YOLOv4,YOLOv5s-x (multi modal ConFigureurations on VEDAI Dataset)

Methods	Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	All	Params
YOLOv3	84.57	72.68	67.13	61.96	43.04	65.24	37.10	58.29	61.26	61.5354M
YOLOv4	85.46	72.84	72.38	62.82	48.94	68.99	34.28	54.66	62.55	52.5085M
YOLOv5s	80.81	68.48	69.06	54.71	46.76	64.29	24.25	45.96	56.79	7.0739M
YOLOv5m	82.53	72.32	68.41	59.25	46.20	66.23	33.51	57.11	60.69	21.0677M
YOLOv5l	82.83	72.32	69.92	63.94	48.48	63.07	40.12	56.46	62.16	46.6046M
YOLOv5x	84.33	72.95	70.09	61.15	49.94	67.35	38.71	56.65	62.65	87.2487M
SuperYOLO	90.86	84.35	78.11	68.11	53.26	82.33	60.95	70.94	73.61	4.8259M
Ours	90.15	83.88	78.77	72.46	60.28	79.78	60.05	73.35	75.86	9.5084 M

6. REFERENCES

[1] Zhuo Zheng, Yanfei Zhong, Junjue Wang, and Ailong Ma, "Foreground-aware relation network for geospa- tial object segmentation in high spatial resolution re- mote sensing imagery," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4096–4105.

[2] Zhipeng Deng, Hao Sun, Shilin Zhou, Juanping Zhao, Lin Lei, and Huanxin Zou, "Multi-scale object detec- tion in remote sensing imagery with convolutional neu- ral networks," ISPRS journal of photogrammetry and remote sensing, vol. 145, pp. 3–22, 2018.

[3] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recogni- tion," arXiv preprint arXiv:1409.1556, 2014.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in

[5] Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[6] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in neural in- formation processing systems, vol. 28, 2015.

[7] Tsung-YiLin, Priya Goyal, Ross Girshick,Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[8] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE confer- enceon computer vision and pattern recognition, 2016, pp. 779–788.

[9] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan,Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang, "Intriguing properties of vision transformers," Advances in Neural Information Process- ing Systems, vol. 34, pp. 23296–23308, 2021.

[10] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu, "Learning roi transformer for oriented ob- ject detection in aerial images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2849–2858.

[11] Zikun Liu, Hongzhen Wang, Lubin Weng, and Yiping Yang, "Ship rotated bounding box space for ship extrac- tion from high-resolution optical satellite images with complex backgrounds," IEEE geoscience and remote sensing letters, vol. 13, no. 8, pp. 1074–1078, 2016.

[12] Danfeng Hong, Lianru Gao, Naoto Yokoya, Jing Yao, Jocelyn Chanussot, Qian Du, and Bing Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," IEEE Transactionson Geoscience and Remote Sensing, vol. 59, no. 5, pp. 4340–4354, 2020.

[13] Mingxing Tan, Ruoming Pang, and Quoc V Le, "Ef- ficientdet: Scalable and efficient object detection," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10781–10790.

[14] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, et al., "ultralyt- ics/yolov5: v7. 0-yolov5 sota realtime instance segmentation," Zenodo, 2022.

[15] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He, "Fcos: Fully convolutional one-stage object detection," in Proceedings of the IEEE/CVF international confer- enceon computer vision, 2019, pp. 9627–9636.

[16] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun, "Yolox: Exceeding yolo series in 2021," arXiv preprint arXiv:2107.08430, 2021.

[17] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jiten- dra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.