[1] Garugu Jaswanth Syam Sundar

[2] Dr. B. Chaitanya Krishna

Superyolo: Super Resolution Assisted Object Detection in Multimodal Remote Sensing Imagery

Abstract :- Finding small things quickly and correctly in remote sensing pictures (RSI) is very hard because you need to use strong feature extraction and complex deep neural networks need a lot of computing power. The study introduces SuperYOLO, a novel approach to identifying objects that seeks to achieve a good combination of speed and accuracy in RSI analysis. SuperYOLO uses a multimodal data fusion method to combine information from different data sources in a way that makes it better at finding small items in RSI. This multimodal fusion (MF) process is both symmetric and compact, which makes it easy to combine data. SuperYOLO has an enhanced super-resolution (SR) learning branch in addition to MF. This SR branch lets the model make high-resolution (HR) feature representations, which lets it tell small items apart from the background when the input is low-resolution (LR). This makes recognition much more accurate without adding too much work to the computer. One great thing about SuperYOLO is that the SR branch is only used during training and is thrown away during inference. This method reduces the need for extra computing power, making sure that object recognition works quickly and efficiently. When tested on the well-known VEDAI RS dataset, SuperYOLO does better at accuracy than cutting-edge models like YOLOv5l, YOLOv5x, and YOLOrs. Additionally, SuperYOLO gets this level of accuracy while greatly lowering the model's parameter size and processing needs. Compared to YOLOv5x, SuperYOLO has 18 times fewer parameters and 3.8 times fewer GFLOPs. To sum up, SuperYOLO makes a strong case for choosing between accuracy and speed when it comes to finding small objects in RSI. The model does a better job than other options because it combines multimodal data fusion with assisted SR learning in a way that makes it more efficient and less complicated to use. This big step forward could have big effects in areas like remote sensing, where finding small things accurately is important for many jobs.

Keywords: - *Object detection, multimodal remote sensing image, super resolution, feature fusion.*

I. Introduction

Working well Object recognition is a key part of many areas of computer vision, from computer-assisted analysis to self-piloting aircraft. With the rise of deep neural network (DNN)-based object recognition systems over the last few decades, the field of computer vision has changed in amazing ways. These systems have been improved and tweaked over time to reach levels of accuracy that have never been seen before. The big steps forward in DNN-based object recognition are mostly due to training them on large natural datasets with labels that have been carefully labeled. But there are some problems that are only present when trying to find things in remote sensing images (RSIs). Finding labeled samples is difficult, which is a major obstacle in object recognition in RSIs. In natural situations, there is a lot of data available, but remote sensing pictures usually only have a few samples that have been named. This lack makes it hard to train DNNs, which makes it hard to get good recognition accuracy. The main task here is to come up with ways to make the most of this small amount of data so that DNNs can adapt well and reliably find objects. Another thing that makes RSIs unique is that the items are very small. These things usually only take up a small part of the picture, maybe tens of pixels at most. This is made even worse by the fact that backgrounds in remote sensing pictures are often very complicated and cover a lot of space. The hard part is accurately finding and locating these small things in large, complicated settings. Traditional object recognition models, which are made for bigger things that you see in real pictures, don't work well in these situations. Because of this, special methods have had to be created to fit the specific needs of RSIs. Another problem with RSIs is that they have a lot of different object sizes and types. RSIs can include a wide range of object sizes and types, unlike

[1] Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Andhra Pradesh, Vaddeswaram, 522302, India

[2] jaswanthgarugu1@gmail.com

[2*] chaitu2502@gmail.com

author Name and email ID: Garugu Jaswanth Syam Sundar, jaswanthgarugu1@gmail.com.

Corresponding author Name and email ID: Dr. B. Chaitanya Krishna, chaitu2502@gmail.com.

nature pictures where things are usually in the same size ranges and groups. Normally, object recognition systems aren't designed to deal with this much variety, and it's not easy to make them work with it. To deal with these problems, we need new methods that can work well with items in RSIs that are of different sizes and types. At the moment, most methods for finding objects are made to work best with a single mode, like RGB or Infrared (IR). This limitation is another thing that makes it hard for RSIs to accurately find objects. More and more, multi-modal image technology is being used in remote sensing. This technology combines data from different devices and sources to give a more complete picture of the Earth's surface. Object detection systems don't work well in RSIs because they can't recognize objects across multiple modes. This is because information from different modes can greatly improve accuracy. It is becoming easier to get RSIs from a variety of imaging sources as imaging technology keeps getting better. This looks like a good chance to close the gap in the accuracy of object recognition. Object recognition systems can get a lot more information by using data from multiple sources. This makes it easier for them to properly identify things on the Earth's surface. In conclusion, deep neural network-based object recognition has come a long way in computer vision, but it faces some unique problems when used with pictures from space. There isn't enough labeled data, the objects are small, there isn't enough scale variety, and there isn't any cross-modality recognition. But the growing supply of multi-modal remote sensing data shows that these problems might be able to be solved. The creation of new techniques and the combination of information from various sources can greatly enhance the accuracy of object detection in RSIs. This can help advance areas such as computer-assisted diagnosis and self-piloting in the context of remote sensing.

II. LITERATURE REVIEW

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation:

Recent analysis of the PASCAL VOC dataset shows that object recognition has not made much progress in recent years. The most effective approach involves combining various image features with contextual information. Our detection method, which is both easy to implement and scalable, achieves a mean average accuracy of 53.3%, surpassing the previous best result on VOC 2012 by more than 30%. We utilize large convolutional neural networks to identify and distinguish objects through bottom-up region proposals. Additionally, our method incorporates supervised pre-training and fine-tuning to enhance performance, especially when there is limited labeled training data. Our research also provides insights into what the network learns and reveals a detailed image trait structure.

Fast R-CNN:

This paper proposes object detection using Fast R-CNN, a Fast Region-based Convolutional Network. Fast R-CNN relies on deep neural network-based item suggestion grouping. Fast R-CNN employs innovative methods to speed up testing and training and improve recognition accuracy. Fast R-CNN is a faster and more accurate method for generating the deep VGG16 network compared to other methods like SPPnet. It achieves this by using Python and C++ (Caffe) and is able to teach the network quicker and perform tests significantly faster. Additionally, Fast R-CNN has a superior performance in terms of PASCAL VOC 2012 mAP.

You Only Look Once: Unified, Real-Time Object Detection:

YOLO is a novel object-finding method. In the past, classifiers identified objects. We see object identification as a regression issue with spatially divided bounding boxes and class odds. One neural network can predict bounding boxes and class possibilities from entire photos in one test. Since the detection process is a single network, its speed may be enhanced throughout. The combination design works rapidly. Our fundamental YOLO model processes real-time 45-fps images. Fast YOLO, a smaller network, can process 155 frames per second and obtain twice as much mAP as other real-time monitors. YOLO isn't as strong at localization as the current recognition algorithms, but it also doesn't predict background phony hits. Finally, YOLO learns general images. It recognizes artwork better than DPM and R-CNN when switching from nature to other images.

Multiple Instance Detection Network with Online Instance Classifier Refinement:

Modern object recognition relies heavily on poorly directed item identification. Deep learning-based weakly guided devices are promising. But poorly monitored deep network-based device training is tougher than completely supervised training. We call poorly directed detection a MIL issue in this study. Hidden nodes in the network are instance classifiers or object detectors. MIL and instance classifier refinement are combined into a deep network in our novel online instance classifier refinement approach. Without object location knowledge, the network may be trained from start to finish using image-level direction. Specific instance labels inferred from weak guidance are supplied to instances that meet in space to enhance the instance classifier live. Deep networks employ several streams to enhance repeated instance classifiers. Each stream watches after its successor. The tough PASCAL VOC 2007 and 2012 benchmarks are utilized for weakly guided object identification. We achieve 47% mAP on VOC 2007, far better than the prior best practice.

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks:

Modern object-finding networks employ region suggestion to infer locations. These recognition networks run faster owing to SPPnet [1] and Fast R-CNN [2]. This slows region proposal processing. Our Region Proposal Network (RPN) shares full-image convolutional features with the recognition network. This makes region recommendations virtually free. RPNs are fully convolutional networks that can predict object bounds and objectness scores at each location. The RPN learns how to provide effective area recommendations for Fast R-CNN recognition from start to finish. A single network is created by combining RPN and Fast R-CNN convolutional features. The RPN portion uses the new nomenclature for neural networks with "attention" processes to direct the merged network. For the extremely deep VGG-16 model, our GPU-based system can discover objects at 5 frames per second (all stages included) [3]. With just 300 recommendations per image, it has the greatest object identification accuracy on PASCAL VOC 2007, 2012, and MS COCO datasets. Faster R-CNN and RPN were the core technologies of ILSVRC and COCO 2015 winners. Everyone can use code today.

III. METHODOLOGY

The studies use Faster R-CNN and Res-Net to improve remote sensing picture identification. They compared the mean Average Precision (mAP) of ResNet-50 (Reference Backbone) and ResNet-101 (Proposed Backbone), which employed the same anchor ratio. Another study added a CNN module that blends semantic information with fine-grained information. This module can replace the basic block in the backbone of object recognition methods to make them work better. In particular, their module has two parts: an Efficient channel attention (ECA) module for changing weights channel-wise and a double branch for getting conceptual information and fine-grained information.

Drawbacks:

- The Faster R-CNN system works well, but it can be hard on computers because it has a lot of steps, like making region proposals and then classifying them.
- There aren't any clear efforts to improve multimodal fusion in the current work, which could make it less useful when working with different types of data.
- Work that has already been done that focuses on improving recognition accuracy using various backbone networks might not naturally be able to solve this multistage object detection problem.
- The second project is mostly about mixing semantic and fine-grained data in the CNN module. However, it might not be able to combine info from different sources. In the second paper, an Efficient Channel Attention (ECA) module and a module that blends conceptual and fine-grained information are shown. This extra complexity could make it harder to apply the module and connect it to other object recognition methods that are already in use.

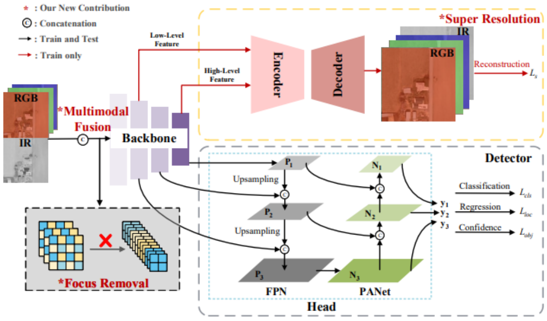
We recommend utilizing RSI to detect items quickly and accurately. SuperYOLO uses super-resolution (SR) learning and multimodal data to locate high-resolution (HR) objects on different sizes. This study focuses on improving the detection of small objects by considering both accuracy and computing cost. The researchers use a method called symmetric compact multimodal fusion (MF) to gather information from different types of data and enhance the detection of tiny objects. They also develop a simplified branch to train models that can identify small objects using low-resolution input, which improves accuracy. During testing, the simplified branch is not used to save time, and the use of low-resolution input speeds up the construction of the network model. The experiments are conducted using the popular VEDAI dataset, which is a subset of the larger AGRC dataset.

Benefits:

- Our work is mostly about using super resolution (SR) learning to find high-resolution objects on items of different sizes. This function is helpful for finding small things with lots of small details.
- Multimodal data merging is built into SuperYOLO. This can improve the accuracy of object recognition by using a variety of data sources.
- SuperYOLO seems to deal with both accuracy and computation cost by using an SR branch during training to improve accuracy and getting rid of it during inference to cut down on computation. This method might be more effective than the previous ones, which might not have taken processing speed into account as much.
- Our work does not define a set backbone network, which makes it more adaptable to different network designs. This could lead to better performance by letting you choose the backbone that's best for the job.

System Architecture :

Fig.1: System architecture



We have created the following modules in order to carry out the aforementioned project.

- Data exploration: this module will be used to load data into the system.
- Processing: data will be read for processing.
- Splitting data into train & test: data will be split into train and test using this module.
- Model generation: Model building - YoloV5S, YoloV5M, YoloV5L, YoloV5X, YoloV5x6, YoloV3, YoloV4, Super Yolo. Algorithms accuracy calculated
- User signup & login: By using this module, you may register and log in.
- User input: This module will provide input for the prediction. Prediction: the final prediction will be shown.

Note: As an extension, we used an ensemble method that combined the results of several separate models to make a final estimate that was more reliable and accurate.

There are, however, other methods we can try to improve the speed even more, such as YoloV5x6 for Detection got 0.99mPA. As an add-on, we can use the Flask framework to build the front end for user testing that includes login.

IV. IMPLEMENTATION

Here in this project, we are used the following algorithms

YoloV5S:

The YOLOv5S object recognition model is a lighter version of the YOLO (You Only Look Once) model. It strikes a good mix between speed and accuracy, making it good for real-time uses. YOLOv5S has a smaller backbone network and fewer lines than bigger versions. This makes it more efficient while still being good at finding objects.

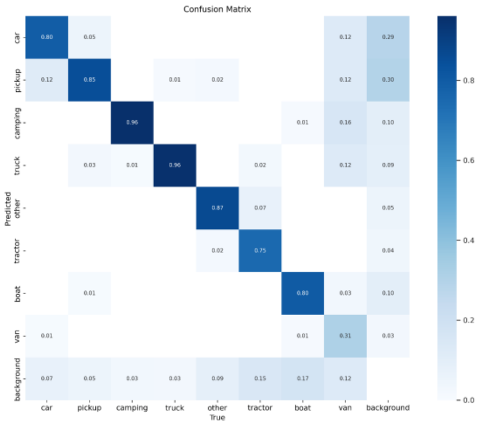


Fig 2 Confusion Matrix for YOLOV5S

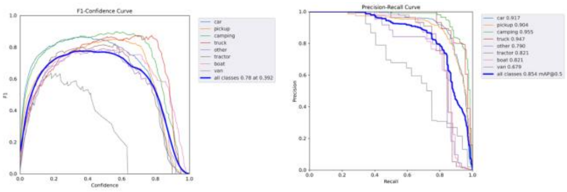


Fig 3 F1-Confidence curve graph and Precision-Recall curve graph

YoloV5M:

The YOLOv5M model for finding objects is a medium-sized version of the YOLO (You Only Look Once) model. It's fast and accurate at the same time, so it can be used in real-time situations. The backbone network and channel setup of YOLOv5M are of a modest size, which makes it efficient while still providing strong object recognition performance.

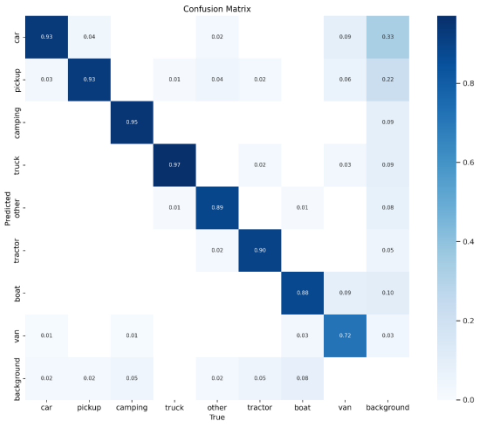


Fig 4 Confusion matrix for YOLOV5M

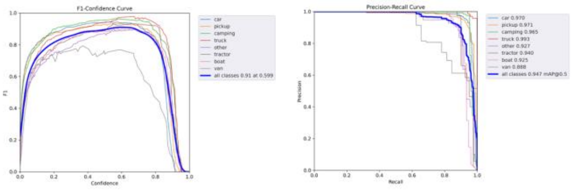


Fig 5 F1-Confidence curve graph and Precision-Recall curve graph

YOLOv5I:

YOLOv5I is a big version of the YOLO (You Only Look Once) object recognition model that is best for jobs that need to be done very accurately. It has a bigger backbone network and more lines than smaller models, which lets it find things more easily in complicated scenes. YOLOv5L is good for jobs that need to be done precisely.

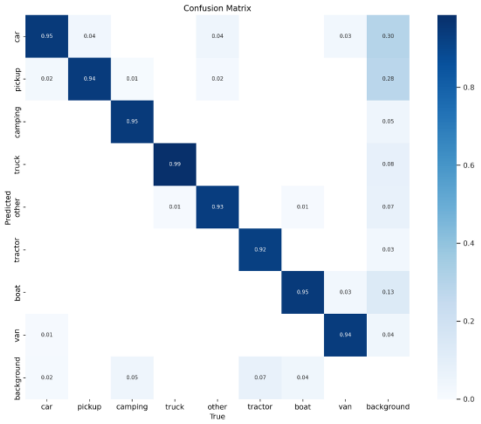


Fig 6 Confusion matrix for YOLOV5I

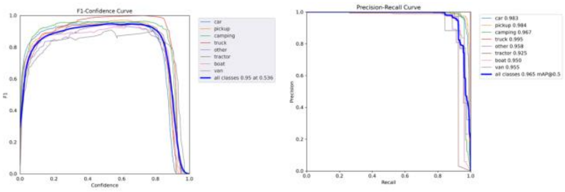


Fig 7 F1-Confidence curve graph and Precision-Recall curve graph

YoloV5X:

YOLOv5X is an extra-large version of the YOLO (You Only Look Once) object recognition model, which is made to find things in photos and movies very accurately. It has an even bigger backbone network and more lines, which makes it good for difficult jobs that need to be done with a lot of accuracy, but it also needs more computing power.

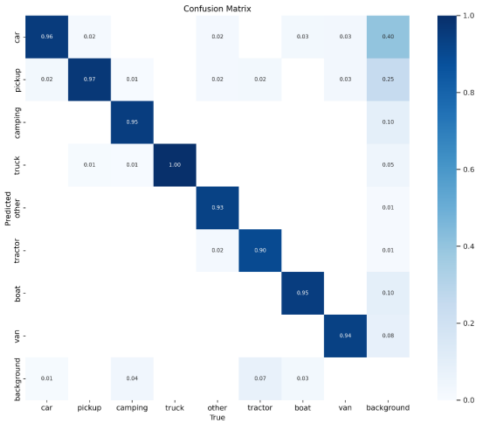


Fig 8 Confusion matrix for YOLOV5X

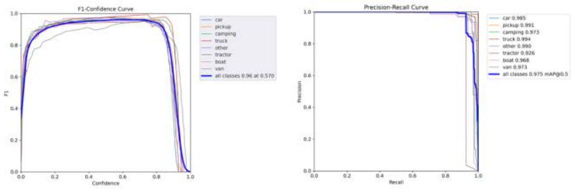


Fig 9 F1-Confidence curve graph and Precision-Recall curve graph

YoloV5x6:

The YOLO (You Only Look Once) object recognition model does not have a standard or commonly known version called YOLOv5x6. It's possible that since then, new model versions or special changes have been made. New sites or the official YOLO library are the best places to find the most up-to-date information on YOLO models.

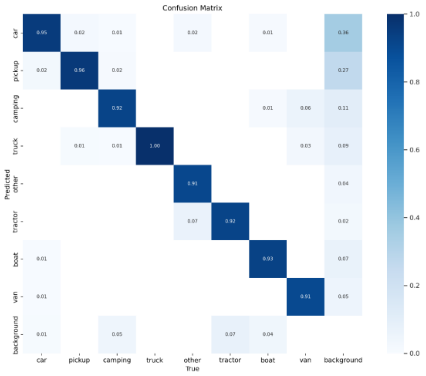


Fig 9 Confusion matrix for YOLOV5x6

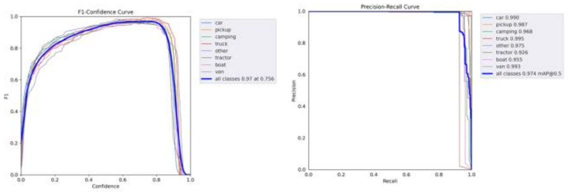
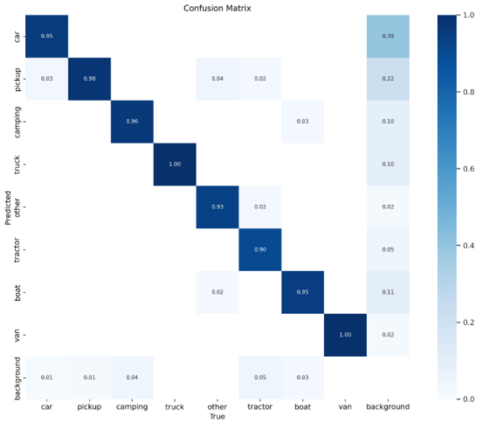


Fig 10 F1-Confidence curve graph and Precision-Recall curve graph

YoloV3:

YOLOv3, which stands for "You Only Look Once," is a program for finding things. It takes a picture and turns it into a grid. Then, it guesses the edges and classes of the things in each grid cell. Because YOLOv3 is fast and accurate, it is used a lot in real-time object recognition systems like self-driving cars and security cameras.



	car	pickup	camping	truck	other	tractor	boat	van	background
car	0.95								0.39
pickup	0.03	0.98			0.04	0.02			0.22
camping			0.96				0.03		0.10
truck				1.00					0.10
other					0.93	0.02			0.02
tractor						0.90			0.05
boat					0.02		0.95		0.11
van								1.00	0.02
background	0.01	0.01	0.04			0.05	0.03		

Fig 11 Confusion matrix for YOLOV3

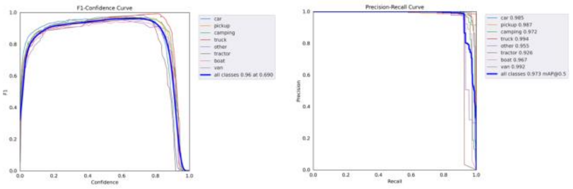
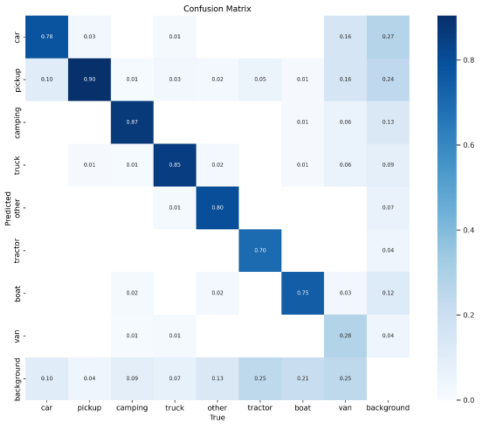


Fig 12 F1-Confidence curve graph and Precision-Recall curve graph

YoloV4:

The "You Only Look Once" object recognition model has been improved with YOLOv4, which makes it faster and more accurate. It adds a CSPDarknet53 backbone, PANet, and spatial attention, which make it easier to find objects in scenes with a lot of them. YOLOv4 is known for having great performance, which makes it useful for many computer vision tasks, such as spying and self-driving cars.



	car	pickup	camping	truck	other	tractor	boat	van	background
car	0.78	0.03		0.01				0.16	0.27
pickup	0.10	0.90	0.01	0.03	0.02	0.05	0.01	0.16	0.24
camping			0.87				0.01	0.06	0.13
truck		0.01	0.01	0.85	0.02		0.01	0.06	0.09
other				0.01	0.80				0.07
tractor						0.70			0.04
boat			0.02		0.02		0.75	0.03	0.12
van			0.01	0.01				0.28	0.04
background	0.10	0.04	0.09	0.07	0.13	0.25	0.21	0.25	

Fig 13 Confusion matrix for YOLOV4

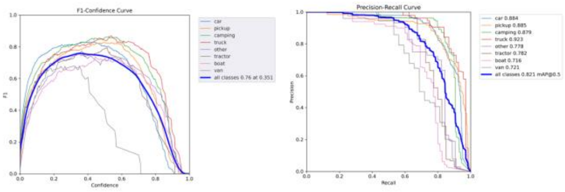
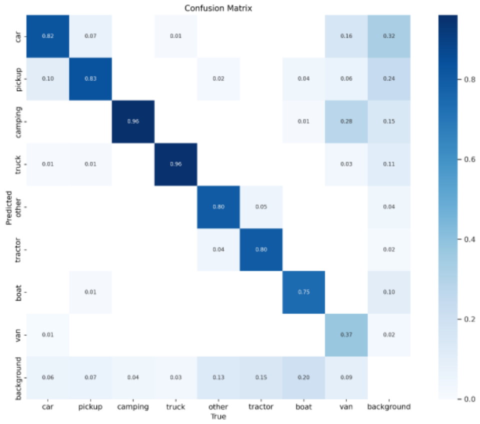


Fig 14 F1-Confidence curve graph and Precision-Recall curve graph

Super Yolo:

The object recognition method YOLO (You Only Look Once) has changed the field of computer vision in a big way. It works quickly and well, which makes it a great choice for jobs that need to find objects in real time.



	car	pickup	camping	truck	other	tractor	boat	van	background
car	0.82	0.07		0.01				0.16	0.32
pickup	0.10	0.83			0.02		0.04	0.06	0.24
camping			0.96				0.01	0.28	0.15
truck	0.01	0.01		0.96				0.03	0.11
other					0.80	0.05			0.04
tractor					0.04	0.80			0.02
boat		0.01					0.75		0.10
van	0.01							0.37	0.02
background	0.06	0.07	0.04	0.03	0.13	0.15	0.20	0.09	

Fig 16 Confusion matrix for SUPER YOLO

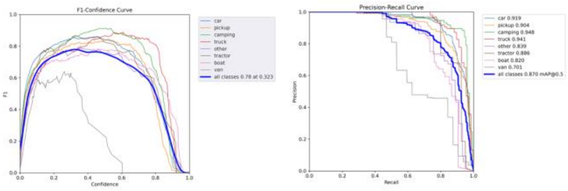
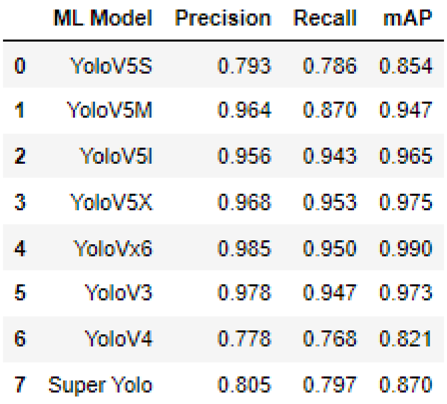


Fig 15 F1-Confidence curve graph and Precision-Recall curve graph

EXPERIMENTAL RESULTS

Comparison graphs for all algorithms used in this project. All graphs are shown in below:



	ML Model	Precision	Recall	mAP
0	YoloV5S	0.793	0.786	0.854
1	YoloV5M	0.964	0.870	0.947
2	YoloV5l	0.956	0.943	0.965
3	YoloV5X	0.968	0.953	0.975
4	YoloVx6	0.985	0.950	0.990
5	YoloV3	0.978	0.947	0.973
6	YoloV4	0.778	0.768	0.821
7	Super Yolo	0.805	0.797	0.870

Fig 16 Performance Evaluation Table For all algorithms

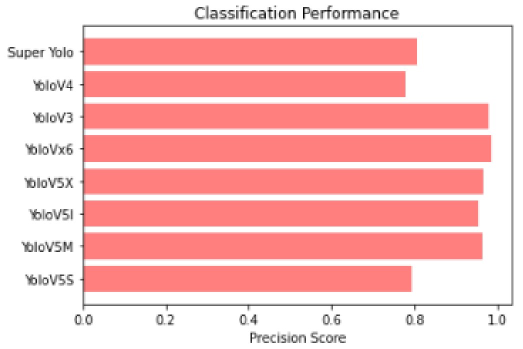


Fig 17 Precision graph for all algorithms

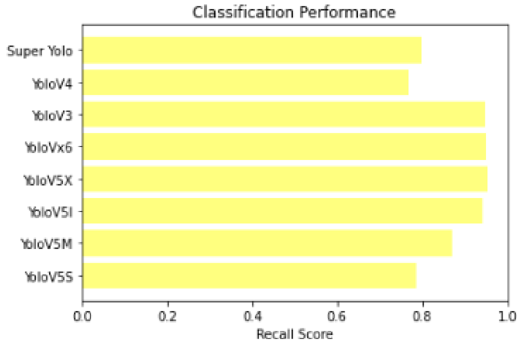


Fig 18 Recall graph for all algorithms

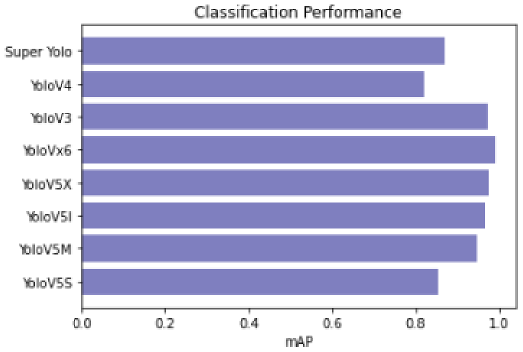
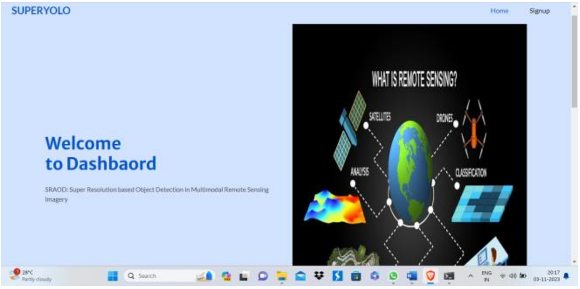


Fig 19 mAP (Mean Average Precision) graph for all algorithms



SUPERYOLO

Welcome
to Dashbaord

SRAOD: Super Resolution based Object Detection in Multimodal Remote Sensing Imagery

Fig 20 Home Page

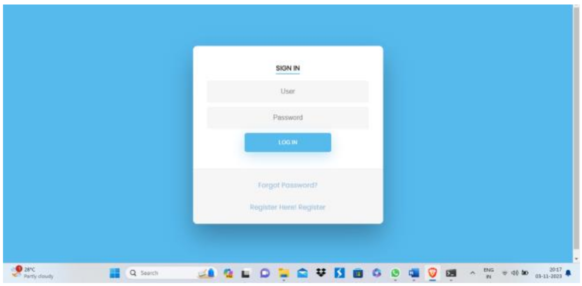


Fig 21 Signin Page

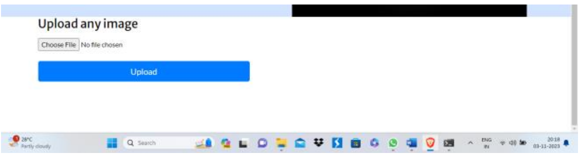


Fig 22 Input Image Page

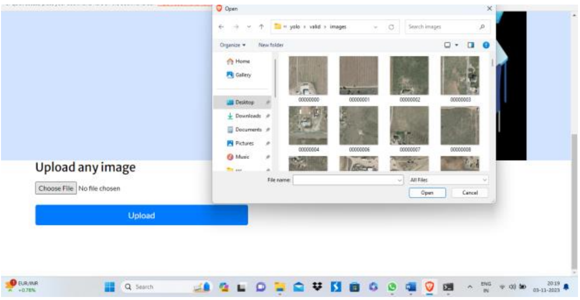


Fig 23 Upload any image page

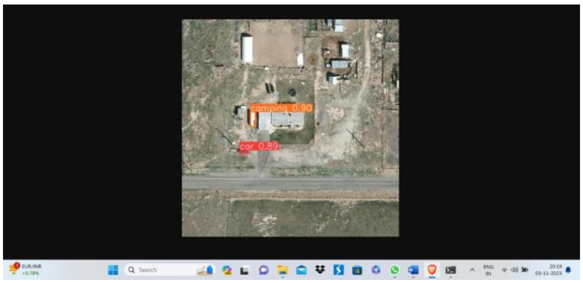


Fig 24 Prediction result for given input

V. Conclusion and Future Scope

In this study project, we are pleased to present SuperYOLO, a revolutionary real-time lightweight network that shows how dedicated we are to improving the ability to find small items in the difficult field of Remote Sensing Imagery (RSI). We worked hard to build SuperYOLO on top of the strong base of YOLOv5s, while also adding new features and making changes in a smart way to push the limits of object recognition. Changing the basic network is one of the most important things we've done in this project. By carefully getting rid of the Focus module, we were able to keep the resolution from dropping, which greatly improved the baseline's abilities. By making this important change, SuperYOLO has been able to solve the problem of losing small items in RSI, which is a big step forward from previous ways. We have also started a path of study fusion, which means using the power of multimodality to improve recognition performance by looking at information from different sources. This combination of different methods has worked well to improve the accuracy and dependability of object recognition, which is another important step forward in our study. The addition of a simple but very flexible SR

(Super-Resolution) branch may be one of the most important changes in SuperYOLO. This addition makes it possible for the network's backbone to create a High-Resolution (HR) model of features. This is a key part of being able to easily and accurately identify small objects against complex backgrounds, even when you give it Low-Resolution (LR) input data. In particular, we made the SR branch to work with the rest of the network without affecting its efficiency in terms of GFLOPs. This way, its use in the inference stage doesn't require changing the original network structure. Because of all of these new ideas and carefully planned methods, SuperYOLO has become a real leader in object recognition. Our score on the VEDAI dataset, which shows how well our method works, is truly amazing. It beats the YOLOv5s standard in terms of mAP50 while also lowering the cost of processing. This double accomplishment shows that we are dedicated to both correctness and speed. When we think about the future, our main goal is to keep making SuperYOLO better. We are excited to work on creating a low-parameter mode that can improve the extraction of HR features even more. This will make sure that our solution not only meets but also exceeds the needs of real-time, high-accuracy object recognition in the ever-changing world of RSI apps. Our journey goes on, and our goal is to push the limits of what is possible in object recognition so that we can better meet the needs of people in the remote sensing community and beyond.

REFERENCES

[1] R. Girshick, D. Jeff, D. Trevor, and M. Jitendra, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 580–587.

[2] R. Girshick, "Fast r-cnn," in Proc. IEEE Int. Conf. Comput. Vis., 2015, pp. 1440–1448.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016, pp. 779–788.

[4] P. Tang, X. Wang, X. Bai, and W. Liu, "Multiple instance detection network with online instance classifier refinement," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 3059–3067.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, 2016.

[6] D. Jia, D. Wei, R. Socher, J. Lili, K. Li, and F. Li, "ImageNet: A largescale hierarchical image database," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2009, pp. 248–255.

[7] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, ´ and C. L. Zitnick, "Microsoft coco: Common objects in context," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," Int. J. Comput. Vis., vol. 88, no. 2, pp. 303–338, 2010.

[9] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 4096–4105.

[10] J. Pang, C. Li, J. Shi, Z. Xu, and H. Feng, "R2 -CNN: Fast tiny object detection in large-scale remote sensing images," IEEE Trans. on Geosci. and Remote Sens., vol. 57, no. 8, pp. 5512–5524, 2019.

[11] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "Multi-scale object detection in remote sensing imagery with convolutional neural networks," ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 3–22, 2018.

[12] J. Ding, N. Xue, Y. Long, G. Xia, and Q. Lu, "Learning RoI transformer for oriented object detection in aerial images," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 2844–2853.

[13] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," IEEE Geosci. Remote Sens. Lett., vol. 13, no. 8, pp. 1074–1078, 2016.

[14] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remotesensing imagery classification," IEEE Trans. on Geosci. and Remote Sens., vol. 59, no. 5, pp. 4340–4354, 2021.

[15] G. J. et al., "ultralytics/yolov5: v5.0," 2021. [Online]. Available: https://github.com/ultralytics/yolov5

[16] S. Zhang, M. Chen, J. Chen, F. Zou, Y.-F. Li, and P. Lu, "Multimodal feature-wise co-attention method for visual question answering," Inf. Fusion, vol. 73, pp. 1–10, 2021.

[17] Y. Chen, J. Shi, C. Mertz, S. Kong, and D. Ramanan, "Multimodal object detection via bayesian fusion," arXiv, 2021. [Online]. Available: https://arxiv.org/abs/2104.02904

[18] Q. Chen, K. Fu, Z. Liu, G. Chen, H. Du, B. Qiu, and L. Shao, "EF-Net: A novel enhancement and fusion network for rgb-d saliency detection," Pattern Recognit., vol. 112, p. 107740, 2021.

[19] H. Zhu, M. Ma, W. Ma, L. Jiao, S. Hong, J. Shen, and B. Hou, "A spatialchannel progressive fusion resnet for remote sensing classification," Inf. Fusion, vol. 70, pp. 72–87, 2021.

[20] Y. Sun, Z. Fu, C. Sun, Y. Hu, and S. Zhang, "Deep multimodal fusion network for semantic segmentation using remote sensing image and lidar data," IEEE Trans. on Geosci. and Remote Sens., 2021.

[21] W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Asymmetric feature fusion network for hyperspectral and sar image classification," IEEE Trans. on Neural Netw. and Learn. Syst., 2022.

[22] Y. Gao, W. Li, M. Zhang, J. Wang, W. Sun, R. Tao, and Q. Du, "Hyperspectral and multispectral classification for coastal wetland using depthwise feature interaction network," IEEE Trans. on Geosci. and Remote Sens., 2021.

[23] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos, and E. Saber, "Yolors: Object detection in multimodal remote sensing imagery," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens., vol. 14, pp. 1497–1508, 2021.

[24] L. Gomez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal ´ classification of remote sensing images: A review and future directions," Proc. IEEE, vol. 103, no. 9, pp. 1560–1584, 2015.

[25] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019, pp. 9725–9734.

[26] M. Haris, G. Shakhnarovich, and N. Ukita, "Task-driven super resolution: Object detection in low-resolution images," arXiv, 2018. [Online]. Available: https://arxiv.org/abs/1803.11316

[27] J. Shermeyer and A. Van Etten, "The effects of super-resolution on object detection performance in satellite imagery," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2019, pp. 1432– 1441.

[28] L. Courtrai, M. Pham, and S. Lefevre, "Small object detection in remote ` sensing images based on super-resolution with auxiliary generative adversarial networks," Remote Sens., vol. 12, no. 19, p. 3152, 2020.

[29] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Smallobject detection in remote sensing images with end-to-end edgeenhanced gan and object detector network," Remote Sens., vol. 12, no. 9, p. 1432, 2020.

[30] H. Ji, Z. Gao, T. Mei, and B. Ramesh, "Vehicle detection in remote sensing images leveraging on simultaneous super-resolution," IEEE Geosci. Remote Sens. Lett., vol. 17, no. 4, pp. 676–680, 2019.

[31] L. Wang, D. Li, Y. Zhu, L. Tian, and Y. Shan, "Dual super-resolution learning for semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 3773–3782.

[32] C. Wang, H. Mark Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh, "CSPNet: A new backbone that can enhance learning capability of cnn," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2020, pp. 1571–1580.

[33] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," Neural Netw., vol. 107, pp. 3–11, 2018.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 9, pp. 1904–1916, 2015.

[35] T. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, ´ "Feature pyramid networks for object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017, pp. 2117–2125.

[36] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018, pp. 8759–8768.

[37] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in Proc. Eur. Conf. Comput. Vis., 2014, pp. 818–833.

[38] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW), 2017, pp. 136–144.

[39] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," IEEE Trans. on Comput. Imaging, vol. 3, no. 1, pp. 47–57, 2016.

[40] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," J. Vis. Commun. Image Represent., vol. 34, pp. 187–203, 2016.

[41] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Proc. 19th Int. Conf. Comput. Statist., 2010, pp. 177–186.