# REAL TIME OBJECT DETECTION

**RAGESHWARI SHUKLA,** 1721610077**, RAGIB KHAN**, 1721610078

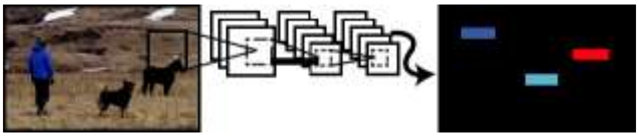
**RISHABH TIWARI** ,1721610086, **SHEHZAD ANSARI**, 1721610098

## ABSTRACT

We present YOLO, which is a new method of object recognition. Previous work on object recognition reused classifiers to perform recognition. Instead, we treat object recognition as a regression problem of spatially separated bounding boxes and related class probabilities. Display category and probability fields directly from the full screen of the score. Since the entire detection line is a single network, it can be directly optimized continuously in terms of detection efficiency. Our unified architecture is very fast. Our basic model, YOLO, processes images in real time at a speed of 45 images per second. Fast YOLO, a smaller version of the network, processes a staggering 155 frames per second and achieves twice as many access points as other real-time detectors. Thanks to modern detection systems, YOLO will generate more positioning errors, but it is unlikely to predict false positives in the background. Finally, YOLO examines very general representations of objects.By extending natural images to other fields such as art, its performance is better than other detection methods, including DPM and R-CNN.

## 1.INTRODUCTION

People look at the picture and immediately recognize which objects are in the picture, where they are and how they interact. The human visual system works quickly and accurately, which allows us to perform complex tasks such as driving a car without thinking too much. Accurate object recognition algorithms will enable computers to drive vehicles without specialized sensors, enable supporting devices to transmit scene information to human users in real time, and unleash the potential of flexible robotic systems. To identify an object, these systems use a classifier for the object and evaluate it at different points and scales on the test image. Systems such as Deformable Part Models (DPM) use sliding window methods in which classifiers work at evenly spaced positions. New methods, such as R-CNN recommended, and YOLO for image processing are simple and straightforward. Our system (1) scales the input image to 448 x 448, (2) performs a single convolutional network on the image, and (3) sets a detection threshold as a result of the model confidence program to identify the first potential bounding box. Generate images. Image, and then run the classifier on the suggested fields.After classification, post-processing is used to streamline the bounding box, remove duplicate detections and re-evaluate the frame based on other objects in the scene. These complex pipelines are time-consuming and difficult to optimize, because each individual component must be trained separately. Object detection is a single regression problem. From image pixels to bounding box coordinates and class probabilities, in our system, only the appearance of the image (YOLO) can be used to predict which objects exist and where YOLO is unusually easy: See Figure 1. The convolutional network says several bounding boxes and the class probabilities of these frames at the same time. YOLO trains the complete image and directly optimizes the recognition performance. Compared with traditional object recognition methods, this unified model has several advantages. First of all, YOLO is very fast. Frame detection is a regression problem, we don't need complicated pipelines.We only need to run the neural network on the new image during the test to predict the detection result. Our core network runs at 45 fps without batch processing on Titan X-GPU, while the fast version runs at 150 fps. Real-time streaming video processing, with a delay of less than 25 milliseconds. In addition, the accuracy of YOLO is more than twice that of other real-time systems. The demonstration of our system can be run on the webcam in real time and can be found in the second page of YOLO (Global Cause in the predicted image).



**In** contrast to methods based on region suggestions and sliding windows, YOLO sees the entire picture during training and testing, thereby implicitly encoding contextual information about the class

and its appearance. R-CNN is an excellent detection method, which makes the background blobs in the image of the object messy because it cannot see the broader context. Compared with Fast R-CNN, YOLO produces less than half of the background errors. Third, YOLO reads the general object representation.When YOLO is trained on natural images

an estimate of the reliability of each cell. The probability that these classification codes appear in the field and the degree to which the predicted field and tested on artworks, its performance is far superior to excellent detection methods such as DPM and R-CNN. Because YOLO is so versatile, it is unlikely to fail when applied to new domains or accidental records. In terms of accuracy, it still lags behind modern recognition systems. Although you can quickly identify objects in an image, it is difficult to accurately determine the exact location of certain objects (especially small objects). We will explore these trade-offs further in our experiments. open source code and various ready-made models can also be downloaded.

## 2. Unified Detection

We combine the different components of object recognition in a neural network. Our network uses the attributes of the entire image to predict eachbounding box. In addition, all bounding boxes in all image categories are predicted at the same time. This means that our network globally evaluates the entire image and all the objects it contains. YOLO is committed to continuous learning and real-time speed, with high average accuracy. Input image is divided into S*S grid by system. When grid cell is hitted by the center of the object, then cell is responsible to identify object. The bounding box B is predicted by each grid unit. The confidence of these confidence levels indicates how well the model determines that the rectangle contains objects and how accurately it generates the predicted image. For example, it predicts IOU Pr(object) is correct.If there are no objects in this cell, the confidence value must be zero; otherwise, we want the confidence value to be equal to the overlap of connections (IOU) between the prediction framework and the underlying truth. 5 predictions-x,y,w,h,& confidence is being haven by each bounding box. The edge of the grid cell is represented by coordinates x, represent the center of rectangle relative. For the entire image height & width are predicted relative. After all, predictable reliability is a promissory note. Between the predicted frame and each actual frame in the scene, each cell in the grid also predicts the probability of conditional class C, Pr(class i | object). These probabilities depend on the cells in the grid that contain the object. The category probability established for each pixel in the raster does not depend on the number of pixels B. During the test, we multiply the category-related probability by the pixel's confidence prediction.(Object) * Pr(object) * IOU truth pred = Pr(classic) * IOU truth pred 1. This gives us matches the object.

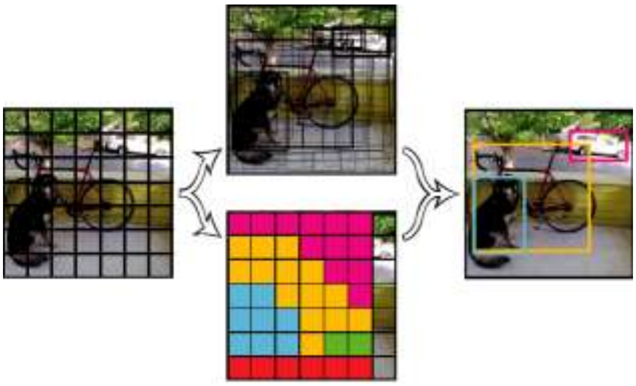


**Figure 2: The Model**. System models detection as a regression problem. It divides the image into S*S grids and predicts the bounding rectangles B, the confidence of these rectangles and the probability of class C. Predictions as S * S * tensor ((C)+5*B) are encoded by each grid unit.

To estimate YOLO in PASCAL VOC, we use S = 7, B = 2. PASCAL VOC has 20 marking levels, so C=20. Then, our final prediction is (7*7*30) tensor.

## 2.1 NETWORK DESIGN

We implementthis model as a convolutional neural network and evaluate it on the PASCAL VOC 9 recognition dataset. The initial layer of the convolutional network extracts features from the image, while the fully connected layer predicts the output coordinates and probabilities. The GoogleNet model used to classify 34 images. Our network consists of 24 folded layers and 2 fully connected layers. We do not use the initial module used by GoogleNet, but simply use 1 * 1 shrinking layer and 3 * 3 folding layer, similar to Lin et al. 22 We also trained a fast version of YOLO, which aims to break the boundaries of fast target detection. Fast YOLO uses a neural network with fewer convolutional layers and fewer filters on these layers.All training and testing parameters of the network, YOLO and Fast YOLO are the same.

## 2.2 TRAINING

We train our convolutional layer on the competitive Image Net 1000 30-class dataset. We train this network for about a week and achieved a top 5 accuracy rate of 88% on the Image Net 2012 test set, which is comparable to the Google Net model on Coffee. model. Zoo 24. We use the dark web framework for all training and inference 26. Then we transform the model to perform discovery. Ren et al. They showed that adding folding and connection layers to the pre-trained network can improve performance [29]. According to his example, we added four layers of convolution and two fully connected layers, and the weights were initialized randomly. Detection usually requires detailed visual information, so we increased the input resolution of the network from 224,224 to 448,448. Our last layer predicts the probabilities of the two categories as bounding box coordinates. We normalize the width and height of the bounding box to the width and height of the image so that they are between 0 and 1.We parameterize the x and y coordinates of the bounding box so that they deviate from the position of the corresponding grid cell, so they are also limited between 0 and 1. We use the linear trigger function for the last layer, and all other levels use the linear fire leakage that is subsequently fixed.

(x)={x, if x>0 and 0.1, otherwise

We optimize the add of sq. errors within the model output. we have a tendency to use the sum of squares error as a result of it's simple to increasing the optimize, however it doesn't totally meet our goal of typical accuracy. It can't be perfect. In addition, several grid cells in every image don't contain any objects. This makes the "confidence" price of those cells nearer to zero, and infrequently exceeds the gradient of the cell containing the object. this could cause the model to become unstable, resulting in discrepancies in early training. to resolve this problem, we increase the loss of bounding box coordinate prediction and scale back the loss of confidence prediction for tables that don't contain features. For this, we have a tendency to use 2 parameters: Y-Coord and Y-noobj. we have a tendency to set Y -ordin = five and Y-noobj =: 5.The add square error additionally weights the big and little squared errors. Our error metric ought to replicate that the little deviations within the large squares aren't as vital because the small squares. To partly solve this problem, allow us to predict the root of the breadth y of the peak of the bounding box. The width and height are direct. YOLO predicts multiple bounding boxes for every grid unit. throughout training, we have a tendency to hope that every object is to blame for just one bounding rectangle. the most current is grounded IOU. This has junction rectifier to the specialization of the bounding box predictor.Each predictor improves the prediction for a selected size, facet ratio, or feature class, thereby increasing overall memory.

## 2.3. INFERENCE

Like training, the detection of predictive test images only requires network estimation. In PASCAL VOC, the network 98 predicts the bounding rectangle of each image and the class probability of each rectangle. YOLO is very fast in the test, because unlike the classifier, it only requires network evaluation-the grid design increases the spatial diversity in the bounding box prediction. It is usually clear which grid unit the object is in, and the network only predicts one frameHowever, some larger objects or objects located at the edge of multiple cells can be well positioned by multiple cells. Non-maximum suppression can be used to correct for these multiple detections. Although it is not important for performance such as R-CNN or DPM, non-maximum suppression increases MAP by 2-3%.

## 2.4 LIMITATIONS OF YOLO

YOLO imposes strict spatial constraints on bounding container prediction, due to the fact every mobileular withinside the grid can handiest are expecting rectangles and may handiest have one class. This spatial constraint limits the range of close by functions that our version can are expecting. Our version fights towards small gadgets that seem in groups, which include cows.

When our version learns to are expecting bounding bins from data, it's miles tough to generalize to gadgets with new or uncommon scales or configurations. Our version additionally makes use of fantastically widespread attributes to are expecting bounding boxesOur structure has numerous ranges of downsampling the enter image.

Finally, even though we use a loss characteristic that approximates the popularity overall performance for training, our loss characteristic can manage the mistakes of small and huge bounding bins. Small insects in huge bins are commonly harmless, however small insects in small bins have a miles more effect at the IOU. Our major supply of blunders is inaccurate location.

## 3. Compare to Other Detection Systems

Object recognition is the core problemin computer vision. The recognition pipeline usually first extracts some reliable features (haar 25, SIFT 23, HOG 4, convolution feature 6) from the input image, and then uses classifiers 36, 21, 13, 10 or locators 1, 32 to locate the object. Identify the functional space. The function of these classifiers or locators is similar to a sliding window on the entire image or a subset of regions in the image frames 35, 15, 39, so as to highlight important similarities and differences.

**Deformable parts model:**Deformable Part Model (DPM) uses sliding window technology to detect objects. 10. DPM uses non-overlapping channels to extract static objects, classify regions, and predict bounding boxes for high-performance regions. Maximum inhibition and neural network for situational thinking. This network is not a static function, but an online learning function and its optimization for recognition tasks. Our unified architecture serves different purposes. A faster and more accurate model than DPM.

**R-CNN:** R-CNN and its variants use region hints instead of sliding windows to find objects in the image. Selective search generates potential bounding boxes, convolutional network extracts features, SVM evaluates rectangles, and linear models correspond to bounding boxes and maximum values to suppress and prevent double detection. Each stage of this complex pipeline had to be fine-tuned independently. As a result, the system was very slow, taking more than 40 seconds per image in 14 tests.

**Deep MultiBox:** Compared with R-CNN, Szegedi and others trained an evolutionary neural network to predict regions of interest [8] instead of using selective search. Multi Box can also perform single object detection and replace trust prediction with single-class prediction. Traditional object recognition is still only part of a larger recognition pipeline that requires additional classification of image regions.Convolutional network used to predict bounding boxes in images, but YOLO is a complete detection system.

**OverFeat.**Sermanet et al. trained a convolutional neural network to perform localization and adjusted the position to perform detection 32. Over Feat effectively performs sliding window detection, but it is still a non-overlapping system. Over Feat optimizes positioning, not recognition performance. The locator only looks at local information during the forecast period. Over Feat cannot consider the global context, so a lot of post-processing is required to obtain consistent detection results.

**MultiGrapes.**In terms of design, our undertaking is similar to that of Redmon et al. [27] Our bounce prediction grid era is mainly based totally absolutely on the MultiGrasp series and regression system, but detection is lots much less complex than element detection. By which includes photographs of the product. Whether you need size, location, or searching beforehand to gadgets or their diploma restrictions, you could effects find the right area. YOLO predicts the boundary discipline and class competencies of multiple devices in multiple commands in an image.

## 4. Experiments

We compare YOLO with other real-time detection systems in PASCAL VOC 2007. In order to understand the difference between YOLO and R-CNN variants, we checked the errors in VOC 2007 between YOLO and Fast R-CNN (one of the older versions). R-CNN performance. 14. Using various error profiles, we show that YOLO can be used to re-evaluate fast R-CNN detection and reduce false positive background errors, thereby significantly improving throughput. We also showed VOC 2012 andCompare MAP with modern methods. Finally, we show that YOLO generalizes to new regions better than other detectors on the two sets of inset data.

### 4.1. Compare to Other Real-Time Systems

Many researches on object recognition have focused on making standard recognition channels faster. 5, 38, 31, 14, 1728 However, only Sadegi and others are actually developing real-time recognition systems (30 frames per second or better)31. We compared YOLO to its DPM GPU-implemented at 30Hz or 100Hz. Although other efforts failed to reach the real-time milestone, we also compared its MAP and speed to examine the trade-off between object detection accuracy and performancesystem.

YOLO is the fastest way to detect objectsPascal; as far as we know, this is the fastest object detector. At a MAP of 52.7%, this is more than twice the accuracy of previous real-time detection work. While maintaining real-time performance, YOLO increased MAP by 63.4%. Train YOLO with the help of VGG-16. This model is biggerMore accurate, but much slower than YOLO. It is useful to compare with other detection systems based on VGG-16, but since it is slower than real-time, the rest of the document will focus on our faster model.

Faster DPM effectively speeds up DPM without sacrificing more MAPs, but still loses 238 times real-time performance.

Compared with the neural network methodology ,it's conjointly restricted by the comparatively low detection accuracy of DPM. R-CNN subtracts R and replaces selective search with a statically linked block set 20.Although it is better than

| Real-Time Detectors | Train | MAP | FPS |
|---|---|---|---|
| 100Hz DPM | 2007 | 16.0 | 100 |

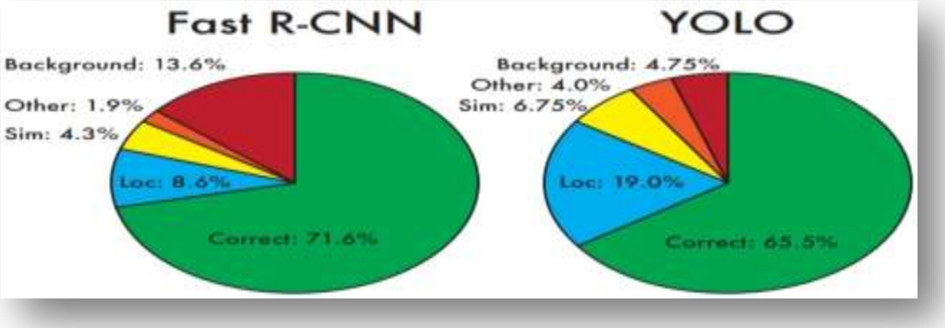
| | | | |
|---|---|---|---|
| 30Hz DPM | 2007 | 26.1 | 30 |
| Fast YOLO | 2007+2012 | 52.7 | 155 |
| YOLO | 2007+2012 | 63.4 | 45 |
| Less Than Real-Time | | | |
| Fastest DPM | 2007 | 30.4 | 15 |
| R-CNN Minus R | 2007 | 53.5 | 6 |
| Fast R-CNN | 2007+2012 | 70.0 | 0.5 |
| Faster R-CNN VGG-16 | 2007+2012 | 73.2 | 7 |
| Faster R-CNN ZF | 2007+2012 | 62.1 | 18 |
| YOLO VGG-16 | 2007+2012 | 66.4 | 21 |

**Table 1: Real time systems on PASCAL VOC - 2007**
Performance and speed comparison of fast detectors. Fast YOLO is the fastest VOC PASCAL detection detector ever, and its accuracy is still twice that of any other real-time detector. YOLO is 10 MAP more accurate than the fast version, although it is still much faster than real-time in terms of speed.

### 4.2. VOC 2007 Error Analysis:

To further investigate the differences between YOLO and the latest generation of detectors, we analyzed a detailed breakdown of the VOC 2007 results. We compared YOLO with Fast R-CNN, because Fast R-CNN

is one of the most effective detectors in PASCAL, and its detection capability is generally available.We use tools and methodologyof Hoiem et al ,for each category, we observed N key predictions for that category during the test. Classified according to the type of error, each prediction is correct.



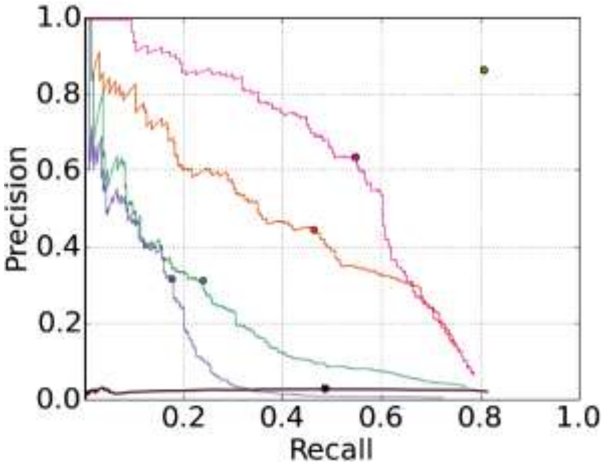
### 4.3. Combine YOLO and fast R - CNN

YOLO makes far fewer serious mistakes than Fast R-CNN. By using YOLO to remove R-CNN fast background detection, we have achieved a significant performance improvement. For each bounding box predicted by R-CNN, we check whether YOLO predicts a similar bounding box. If so, we will strengthen this prediction based on the probability predicted by YOLO and the overlap between the two fields.

| | MAP | Combine | Gain |
|---|---|---|---|
| Fast R CNN | 71.8 | - | - |
| Fast R-CNN (2007 data) | 66.9 | 72.4 | .6 |
| Fast R-CNN (VGG-M) | 59.2 | 72.4 | .6 |
| Fast R-CNN (CaffeNet) | 57.1 | 72.1 | .3 |
| YOLO | 63.4 | 75.0 | 3.2 |

Table 2: Model evaluation experiments in VOC 2007. We tested the impact of mixing numerous fashions with the quality model of Fast R-CNN. Other variations of Fast R-CNN provide simplest a small benefit, whilst YOLO gives a great

growth in performance.



### 4.4. VOC 2012 Results

In the 2012 VOC test suite, YOLO achieved 57.9% of MAP. This is lower than the current one and closer to the original R-CNN with VGG-16, as shown in Table 3. In categories such as TV, monitor, sheep and bottle YOLO scores 8-10% less than R-CNN or editing functions. However, in other categories such as cats and trains, YOLO scores higher. This model is the most effective method for detecting. Fast R-CNN improved by 2.3% over YOLO and gained 5 positions in the public rankings.

### 4.5. Generalizability- In artwork, Person detection

The academic object detection data set extracts training and test data from the same distribution.In sensible applications, it's troublesome to predict all potential use cases, and therefore the take a look at knowledge is also totally different from what the system has seen before. The recognition system in Picasso Dataset 12 and People-Art Dataset 3. These two data sets are used to verify the recognition of people in artworks.

**Figure 5:** Shows the comparative performance and other detection methods of YOLO. For reference, we personally provide an AP for VOC 2007 detection, in which all models are trained on VOC 2007 data only. In Picasso, the model is trained in VOC 2012, and the character type is applied to art work training in VOC 2010 and VOC 2007. R-CNN has been significantly reduced. R-CNN uses selective search for bounding box clues matching natural images. The classifier step in R-CNN can only see small areas and requires good suggestions. DPM supports your AP well when applied to graphics. Previous work has shown that DPM works well because it contains solid space models of shapes and structures. Although DPM is not as degraded as R-CNN, it starts with a lower AP. YOLO achieved very high scores in VOC 2007. For art works, the drop in AP is less than other methods.

## 5. Real time detection in the wild .

YOLO is a quick and correct item detector, very appropriate for system imaginative and prescient applications. We connect YOLO to the webcam and check whether it works in real time.

| | VOC 2007 | | Picasso | People-Art |
|---|---|---|---|---|
| | AP | AP | Best $F_1$ | AP |
| YOLO | 59.2 | 53.3 | 0.590 | 45 |
| R-CNN | 54.2 | 10.4 | 0.226 | 26 |
| DPM | 43.2 | 37.8 | 0.458 | 32 |
| **Poselets** 2 | 36.5 | 17.8 | 0.271 | |
| D&T 4 | - | 1.9 | 0.051 | |

(a) People-Art Datasets

(b) Picasso Dataset

Figure 5- Generalization: Results on People Art Datasets and Picasso Dataset.

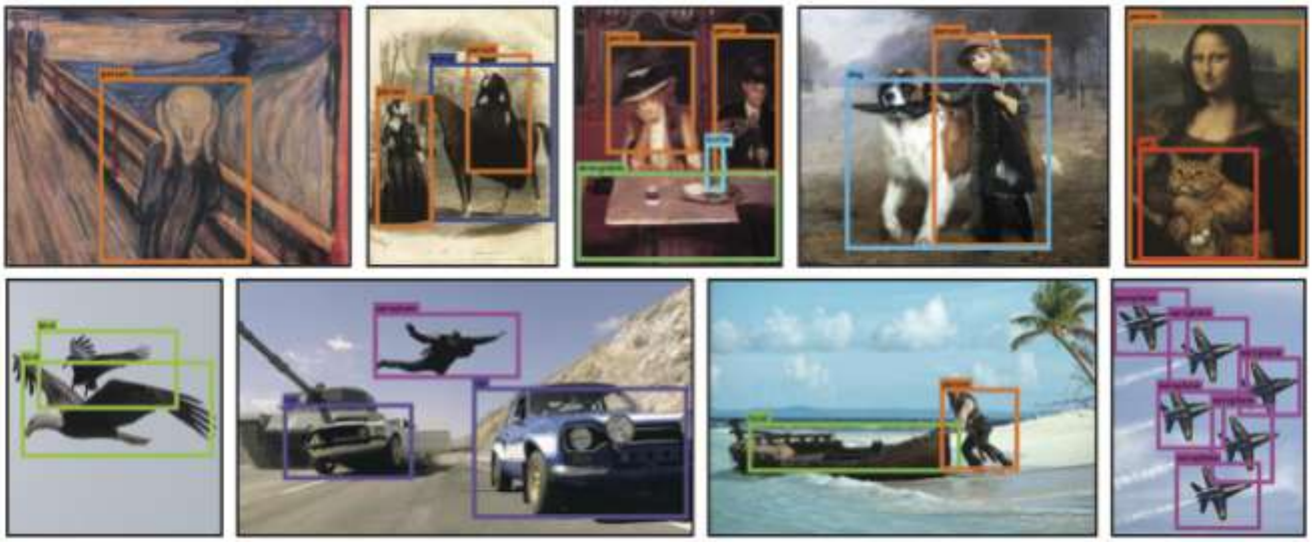


**Figure 6.** Qualitative results: YOLO is based on illustrations and natural images from the Internet. Although he believes that a person is an airplane, he is mostly accurate.

Including camera image search and screen recognition time. The resulting system is interactive and engaging. Although YOLO processes images separately, when connected to a webcam, it can act as a tracking system, detecting moving objects and changing their appearance.

**6. Conclusion:**We introduce YOLO, a unified object recognition model. Our model is easy to build and can be trained directly on the full screen. Compared with the classifier-based method, YOLO is trained on the loss function directly corresponding to the recognition performance, and the entire model is very fast. YOLO is the fastest general-purpose object detector in the literature, and YOLO is the leader in cutting-edge real-time object technology. YOLO can also be extended to new areas, so it is very suitable for applications that rely on fast and reliable object recognition.

## References :

[1] M. B. Blaschko and C. H. Lampert. Learning to localize ob-jects with structured output regression. In Computer Vision– ECCV 2008, pages 2–15. Springer, 2008.

[2] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In International Conference on Computer Vision (ICCV), 2009.

[3] H. Cai, Q. Wu, T. Corradi, and P. Hall. The cross-depiction problem: Computer vision algorithms for recog-nising objects in artwork and in photographs. arXiv preprint arXiv:1505.00110, 2015.

[4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recogni-tion, 2005. CVPR 2005. IEEE Computer Society Conference on, volume 1, pages 886–893. IEEE, 2005.

[5] T. Dean, M. Ruzon, M. Segal, J. Shlens, S. Vijaya-narasimhan, J. Yagnik, et al. Fast, accurate detection of 100,000 object classes on a single machine. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Confer-ence on, pages 1814–1821. IEEE, 2013.

[6] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional acti-vation feature for generic visual recognition. arXiv preprint arXiv:1310.1531, 2013.

[7] J. Dong, Q. Chen, S. Yan, and A. Yuille. Towards unified object detection and semantic segmentation. In Computer Vision–ECCV 2014, pages 299–314. Springer, 2014.

[8] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov. Scalable object detection using deep neural networks. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Confer-ence on, pages 2155–2162. IEEE, 2014.

[9] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual ob-ject classes challenge: A retrospective. International Journal of Computer Vision, 111(1):98–136, Jan. 2015.

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ra-manan. Object detection with discriminatively trained part based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.

[11] S. Gidaris and N. Komodakis. Object detection via a multi-region & semantic segmentation-aware CNN model. CoRR, abs/1505.01749, 2015.

[12] S. Ginosar, D. Haas, T. Brown, and J. Malik. Detecting peo-ple in cubist art. In Computer Vision-ECCV 2014 Workshops, pages 101–116. Springer, 2014.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich fea-ture hierarchies for accurate object detection and semantic segmentation. In Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, pages 580–587. IEEE, 2014.

[14]R. B. Girshick. Fast R-CNN. CoRR, abs/1504.08083, 2015.

[15] S. Gould, T. Gao, and D. Koller. Region-based segmenta-tion and object detection. In Advances in neural information processing systems, pages 655–663, 2009.