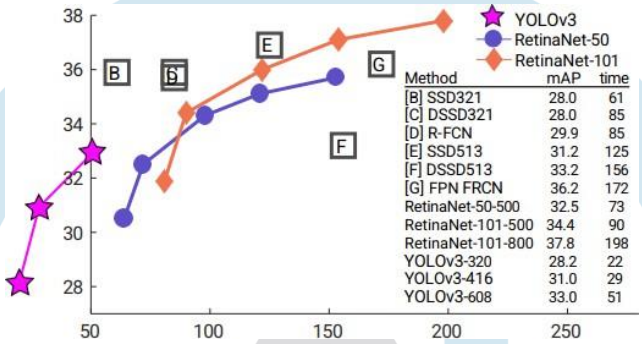
Real-Time Vision: Unleashing YOLO for Instant Object Recognition

[1]Arjun P, [2] Dhananjay P , [3]Gaurav Keshava , [4]Gowrav S Kumar , [5]T S Bhagavath Singh

[1,2,3,4,5]RNS Institute of Technology,Bengaluru-560098

Abstract: ***The following paper is a comprehensive overview of real- time object detection using the YOLO framework. YOLO has become a lot popular for its ability to obtain high speed while having high accuracy, making it suitable for real-time applications. The paper explores various versions of YOLO and examines their advancements, strengths, and limitations in real-time object detection scenarios in it.***



Method	mAP	time
[B] SSD321	28.0	61
[C] DSSD321	28.0	85
[D] R-FCN	29.9	85
[E] SSD513	31.2	125
[F] DSSD513	33.2	156
[G] FPN FRCN	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198
YOLOv3-320	28.2	22
YOLOv3-416	31.0	29
YOLOv3-608	33.0	51

1. Introduction

Object detection, a fundamental task in computer vision, plays a pivotal role in enabling machines to perceive and comprehend visual input from the surrounding environment. In real-world applications, the demand for swift and accurate identification of objects has surged, especially in scenarios where time-sensitive decisions are crucial. The breakthroughs in addressing this need is the advent of the YOLO framework. The design brought a revolutionary method to object detection: a network that could predict bounding boxes and class probabilities for many items in an image at the same time.Unlike traditional methods, it achieves impressive speed without compromising detection accuracy. This attribute makes it particularly well-suited for real-time applications, where rapid decision-making based on visual inputs is imperative. Unlike traditional methods, it achieves impressive speed without compromising detection accuracy. This attribute makes it particularly well-suited for real-time applications, where rapid decision-making based on visual inputs is imperative.

2 Analysis

2.1 The Speed

Now, here's the thing about YOLO: Most of our ideas were borrowed from others. Additionally, we trained an improved iteration for the classifier network. During the training phase, the combination inference time (ms)

Figure 1. This figure is modified from the Focal Loss study . When compared to other detection techniques with similar performance, YOLO operates much more quickly.

2.2 Bounding Box Prediction

By using dimension clusters as anchor boxes, our approach predicts bounding boxes . Every bounding box— tx, ty, tw, and th— has four predicted coordinates by the network.

Prediction of the box is a crucial element in the identification of the objects, as it is essential for accurately localizing and identifying items. The input is divided into cells using a grid-based method. Next, for every cell, the network forecasts bounding boxes and related probabilities. By dynamically adjusting the dimensions of the box, as opposed to conventional methods that rely on anchor boxes, more flexibility is available when managing objects with different sizes and aspect ratios. Each grid cell that the network produces has several bounding boxes, each of which has a menthod in which it shows the user how likely it is to contain an object. The predictions match if the bounding box previous has width and height pw, ph, and the cell is offsets.

of squared error loss is employed, particularly for the object predictions.

The gradient is calculated while working with coordinates by subtracting the expected value (t*) from the ground truth value (tˆ*).

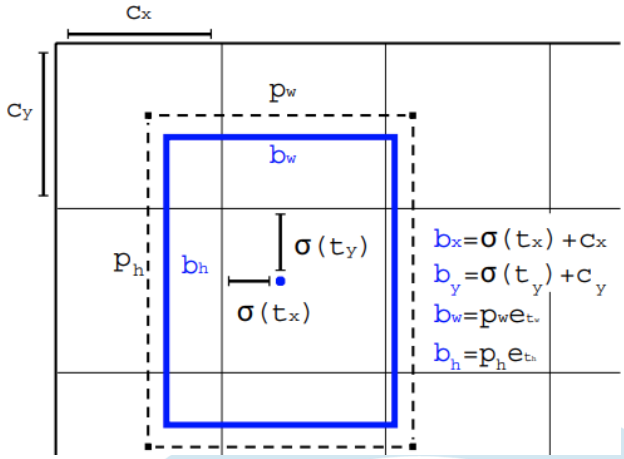


Figure 2: It shows all the boxes and its position prediction and dimension priors.

is not optimal; however, if it overlaps a ground truth object by a certain amount, we disregard the prediction and proceed as follows. We apply the criterion of.5. Every ground truth object has only one box assigned to it, unlike our system. Coordinate or class predictions are not lost if it is prior to the not assigned to a ground truth object; just objectness is lost.

2.3 Class Prediction

In the realm of YOLOv3 object prediction, the class prediction is one of the fundamental component contributing to the algorithm's ability to accurately classify objects within an image. It makes use of a neural network design has more functionalities other than to anticipate bounding boxes but also assigns class probabilities to each bounding box. This is accomplished by applying a softmax function that will activate to the output, which generates a probability among the various classes.

During training, the cross-entropy loss is commonly utilized to optimize the class predictions. The network refines its parameters that are used to reduce the discrepancy between the ground truth class labels attached to the training dataset's and its anticipated class probabilities.

2.4 Predictions Across Scales

Within the architecture, object predictions are made at three different scales. Our method uses a notion similar to the methods like pyramid to extract features from these scales.Beginning with a base feature extractor, Several convolutional layers are introduced. The last layer of these convolutions predicts a 3-dimensional tensor that includes class predictions and objectness. We forecast boxes that have a value of three in our experiments with the dataset, yielding a tensor of size N × N × [3 ∗ (4 + 1 + 80)], which corresponds to the eighty class predictions. Subsequently, we perform upsampling on the two layers prior by a factor of 2×. Furthermore, we use concatenation to combine this upsampled feature map with an earlier map from the network.. This strategy enhances semantic information from the upsampled features and captures finer-grained details from the earlier feature map. Additional convolutional layers process this amalgamated feature map, culminating in the prediction of a tensor, that is two times as big. To ensure that predictions at the third scale benefit from both previous computations and fine- grained information from the early stages of the network, this architecture is repeated once more to forecast the final boxes .

To establish bounding box priors, we continue to employ k-means clustering, selecting arbitrarily, and evenly distributing the clusters across the scales. For the COCO dataset, the chosen 9 clusters corresponded to dimensions (10 × 13), (16 × 30), (33 × 23), (30 × 61), (62 × 45),and (373 × 326). Its careful planning and clustering approach add to its flexibility and performance in object detection applications.

2.5 Extraction of Features

The feature extraction process employs a novel network also called as the Darknet-53, a hybrid architecture inspired by the Darknet-19 and incorporating elements from the emerging residual network paradigm. With its series of 3 × 3 and 1 × 1 convolutions, this novel network achieves a compromise between classic convolutional layers and it contains shortcut connections for improved information flow. Darknet-53 represents a substantial advancement, featuring 53 convolutional layers, distinguishing it from its predecessor, Darknet-19.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Table 1. **Darknet-53.**

Backone	Top-1	Top-5	Bn Ops	BFLOP/s	FPS
Darknet-19 [15]	74.1	91.8	7.29	1246	**171**
ResNet-101[5]	77.1	93.7	19.7	1039	53
ResNet-152 [5]	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	78

Table 2. **Comparison of backbones.** Accuracy, billions of operations, billion floating point operations per second, and FPS for various networks.

In our comparative analysis of feature extraction networks, each model, including Darknet-53, was trained under uniform settings and evaluated at a resolution of 256 × 256, utilizing single crop accuracy metrics. Notably, Darknet-53 exhibits comparable performance to state-of- the-art classifiers while demonstrating advantages for computational efficiency and speed. Specifically, when benchmarked against ResNet-101, Darknet-53 outperforms it and is 1.5× faster. Furthermore, when compared to ResNet-152, Darknet-53 maintains similar performance while achieving a 2× speed increase. Darknet-53 is further emphasized by its highest measured floating-point.

2.6 Training

The object prediction training procedure is an essential step in making sure the model learns and categorize things in an image. During training, it uses the addition of squared error loss, concentrating on bounding box predictions. Ground truth values for bounding box coordinates, objectness, and class predictions are included in the training data.

3 Working

This method is really helpful and efficient. The average mean that is calculated for the data set has performance that is the same as the other versions of the single shot object detection systems.

Even though they are more accurate, sophisticated models like RetinaNet are still significantly faster and more efficient than the Yolo.

In evaluating of map using the traditional metric IOU=0.5 (or AP50), The results demonstrate its resilience, ranking it nearly equal to RetinaNet and far higher than Single Shot versions. These factors underscores its strength as a detector, excelling in producing accurate bounding boxes for objects. On the other hand, when the IOU threshold rises, reflecting the level of agreement between the ground truth and expected boxes, experiences a significant drop in performance, suggesting challenges in achieving precise box alignment with objects.

Historically, it faced difficulties in finding objects that are minimal in size . Yet, a notable reversal of this trend is observed with the introduction of multi-scale predictions in You Only Look Once, showcasing improved precision performance for small objects. Surprisingly, the model performs relatively worse on medium and larger-sized objects, indicating the need for additional research in this area to fully comprehend the underlying dynamics.

As seen in the graphic, we use the metric to compare speed and accuracy.

4 Future Enhancement

During procedures involving You Only Look Once, numerous experimental approaches were explored, and while some yielded successful results, others proved ineffective. In one attempt, the traditional approach of predicting offsets as multiples of box width or height using a linear activation was evaluated by including anchor box predictions. However, this approach resulted in decreased suboptimal performance and drastic decrease in the stability .

Without using logistic activation another exploration experimenting with the predictions was done. Despite the direct prediction of x, y offsets, this alteration showed noticeable decline in mAP by a few points. The linear activation did not prove to be as effective as the logistic counterpart.

Furthermore, the application of focal loss was considered in an effort to address certain challenges. Surprisingly, this modification resulted in a decrease of approximately 2 points in map. This is better and efficient.

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [5]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [8]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [6]	Inception-ResNet-v2 [21]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [20]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [15]	DarkNet-19 [15]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [11, 3]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [3]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [9]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [9]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

Table 3. In my survey paper, I've utilized tables from [9] to expedite the process. YOLOv3 excels, surpassing RetinaNet by 3.8 times in processing efficiency per image. It outperforms SSD variants and competes favorably with state-of-the-art models on the AP50 metric, underscoring its efficiency in object detection methodologies.

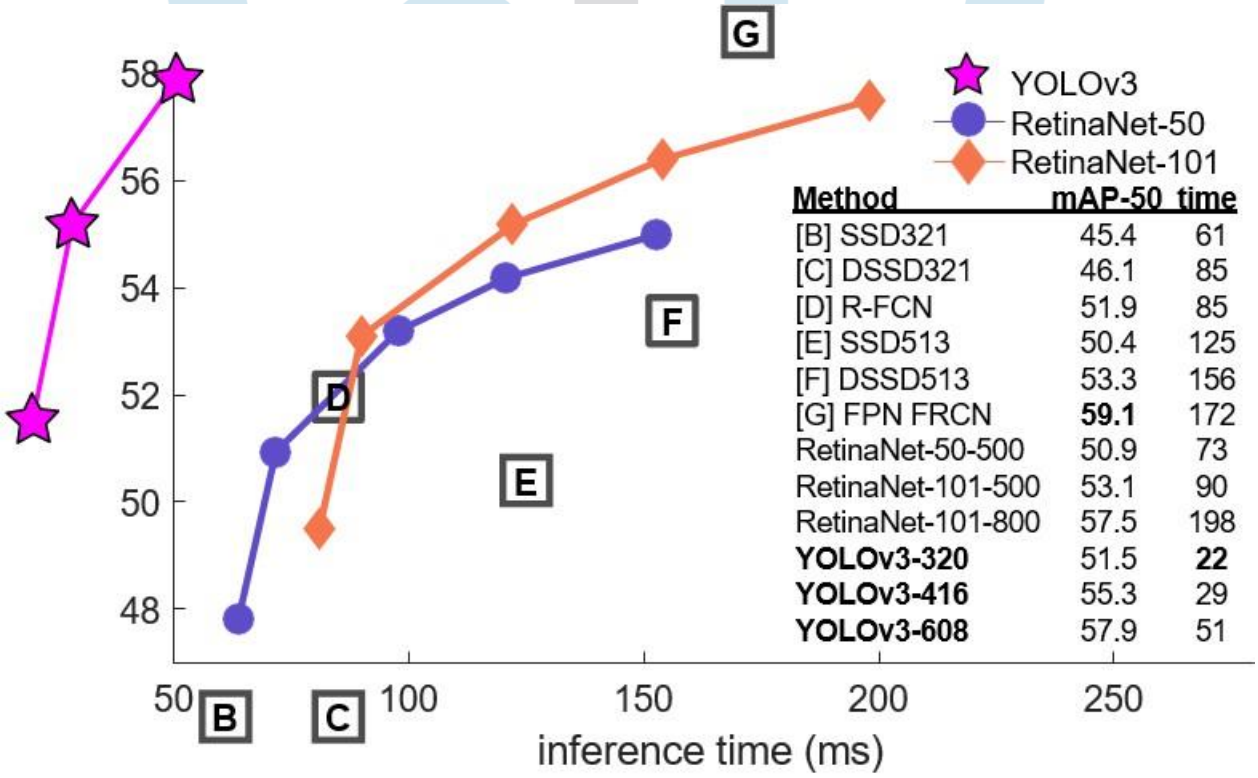


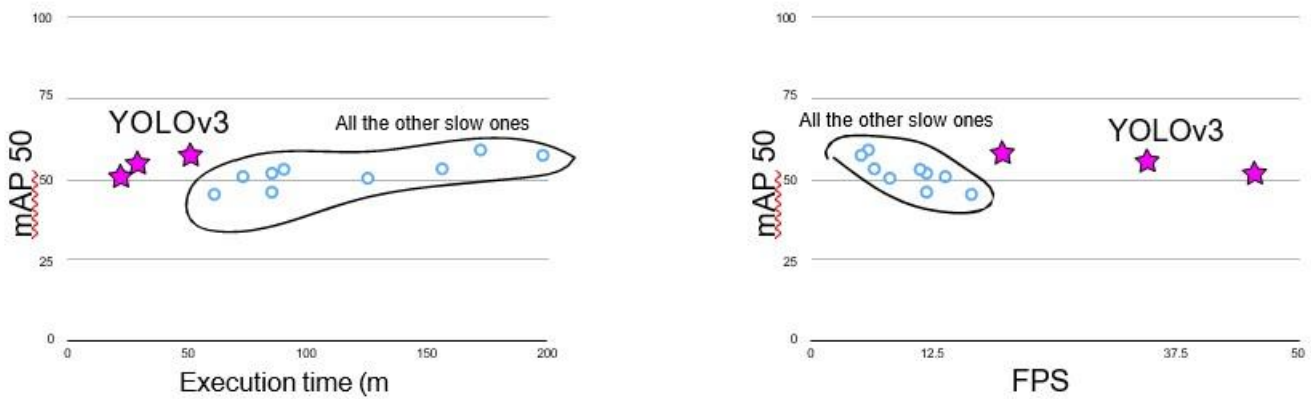
Figure 3. In my survey paper, referencing [9], I showcase the speed/accuracy tradeoff on the mAP at .5 IOU metric. YOLOv3's outstanding balance of high accuracy and speed is evident, notably positioned to the left. Citing my work as [16], I also note a significant 2 mAP improvement in YOLOv2 due to the resolution of a data loading bug, subtly integrated to maintain overall coherence.

During training, Faster R-CNN uses a dual IOU threshold technique, classifying predictions as positive examples if they have an overlap of 7 with the ground truth, anything below the values of .3 are examples that are negative. While we attempted a similar strategy, it did not yield satisfactory results. However, our current formulation appears to be at a local maximum and has demonstrated promise. These kinds of strategies could produce positive results if the training process is stabilized through additional adjustment.

5 Results

YOLOv3 emerges as a commendable detector, excelling in multiple factors of testing . The performance is less impressive compared to the average AP .The decision to shift metrics is shrouded in mystery, as merely alludes to a forthcoming discussion on evaluation metrics when the server is completed. Notably, human evaluators themselves find distinguishing between IOU in the system, surprisingly challenging, raising questions about the significance of such nuances in performance metrics. An important consideration emerges regarding the usage for these detectors, especially given the involvement of major tech entities like Google and Facebook. Furthermore, with significant funding for vision research coming from the military, historical concerns about the ethical implications of technological advancements resurface .While the optimism surrounding computer vision's positive applications, such as wildlife monitoring or pet tracking, acknowledges the existence of the scenario in which malicious users of this technology exist. As researchers in this field, there is a collective responsibility to critically evaluate the potential harms our work may inflict and actively seek ways to mitigate them .

In conclusion, it is imperative to reflect on the ethical dimensions of this research, acknowledging its diverse applications and the associated responsibilities. This kind of thinking is highly beneficial to the society when consider the ethical issues which arises with the widespread use of this kind of technology. As we navigate the landscape of the applications of the computer vision, it becomes evident these are not only a tool for positive endeavors, such as counting how many animals that are present in a particular area or tracking household pets , but these applications makes us think about the problems related to privacy and potential misuse, especially within the realms of data-centric companies and military applications . The inherent difficulty in distinguishing minute differences in IOU metrics, as highlighted by Russakovsky et al, prompts us to reevaluate the significance of these benchmarks and whether they truly encapsulate the practical effectiveness of detection models.



Disadvantages

While it is praised for its speed and accuracy in object detection, it is not immune to criticisms and potential limitations. Some notable points of critique include:

1. **Localization Precision:** Grid-based approach may lead to challenges in precise object localization, particularly with small or densely packed objects. Multiple detection systems that have two-stage detectors, might exhibit better localization accuracy in certain scenarios.
2. **Handling of Small Objects:** It struggle with accurately detecting objects that are tiny cause of the limitations imposed by its coarse grid cells. This can be a drawback in applications where the identification of smaller details is crucial.
3. **Training Data Imbalance:** Like many deep learning models, It can be sensitive to imbalances in the dataset that is use for training.,if certain classes or scenarios are underrepresented.
4. **Limited Context Understanding:** It processes images independently without considering contextual relationships between objects. In complex scenes where object interactions are essential for accurate understanding, You Only Look Once may face challenges compared to methods.

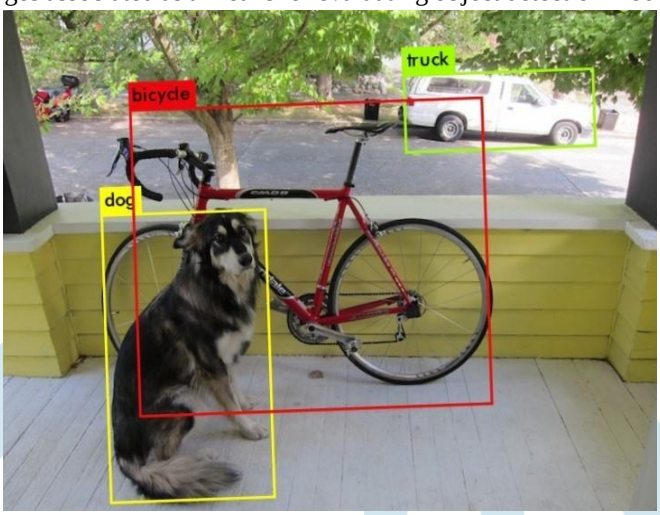
The accuracy in placing these boxes but this increase in accuracy is obtained at the expense of accuracy in which classification of the different objects are done.

The decision to shift metrics is shrouded in mystery, the paper that is used in the algorithm merely alludes to a forthcoming discussion on evaluation metrics upon its completion. Notably, human evaluators themselves find distinguishing surprisingly challenging, raising questions about the significance of such nuances in performance metrics. We understand the concern regarding the competitive landscape, often revolving around incremental gains in the 2-3% range. We appreciate the scrutiny and remain committed to addressing these considerations for the improvement of the robustness and transparency of our methodology. However, we assert that You Only Look Once brings significant advancements to the field of map and also in real-time processing capabilities.

.

There are disadvantages in the use of mAP as it's two- dimensional and all that matters for it is the ranking order of the objects that is determined before the classification.

In reconsideration of the challenges associated as a metric for evaluating object detection models. Based on our observations



there are multiple discrepancies between the metrics used in research and the real-world concerns of end-users. The question arises as to whether our current metrics truly capture what matters most in practical applications. An interesting proposition for new metrics involves shifting the focus from per-class AP to a more holistic evaluation. Specifically, one could explore the efficacy of a global average precision calculated across all classes, or even an AP calculation per image followed by averaging across all images.

Moreover, it seems reasonable to argue that the conventional bounding box approach may have limitations, and a paradigm shift towards instance segmentation, specifically mask-based evaluation, could be more representative of real-world scenarios. However, it's acknowledged that transitioning to mask-based models, like You Only Look Once might pose different issue for the training of the system.

References

[1] Analogy. *Wikipedia*, Mar 2018. 1M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and Zisserman. The pascal visual object classes (voc) chal- lenge. *International journal of computer vision*, 88(2):303– 338, 2010. 6

[2] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. Dssd: Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017. 3

[3] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi. Iqa: Visual question answering in interactive environments. *arXiv preprint arXiv:1712.03316*, 2017. 1

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learn- ing for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[5] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara,Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. 3

I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit,

[6] . Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik,

[7] D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Open- images: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 2

[8] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and

[9] S. Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 2, 3

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dolla´r. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017. 1, 3, 4

[11] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ra- manan, P. Dolla´r, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[12] V ɔr. In *E*

[13] Springer, 2016. 3

I. Newton. *Philosophiae naturalis principia mathematica*.William Dawson & Sons Ltd., London, 1687. 1

J. Parham, J. Crall, C. Stewart, T. Berger-Wolf, and D. Rubenstein. Animal population censusing at scale with citizen science and photographic identification. 2017. 4

K. Redmon. Darknet: Open source neural networks in c.

[14] http://pjreddie.com/darknet/, 2013–2016. 3

J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 6517–6525. IEEE, 2017. 1, 2, 3

K. Redmon and A. Farhadi. Yolov3: An incremental improve- ment. *arXiv*, 2018. 4

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: To- wards real-time object detection with region proposal net- works. *arXiv preprint arXiv:1506.01497*, 2015. 2

[16] O. Russakovsky, L.-J. Li, and L. Fei-Fei. Best of both worlds: human-machine collaboration for object annotation.

[17] In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2121–2131, 2015. 4

[18] Scott. Smart camera gimbal bot scanlime:027, Dec 2017. 4

Keywords

COCO - Common Objects in Context MAP - Mean Average Precision

IoU - Intersection over Union GPU - Graphics Processing Unit