

How to Train an Accurate and Efficient Object Detection Model on any Dataset

Galina Zaleskaya¹, Bogna Bylicka² and Eugene Liu³

¹Intel, Israel

²Intel, Poland

³Intel, U.K.

Keywords: Deep Learning, Computer Vision, Object Detection, Light-Weight Models.

Abstract: The rapidly evolving industry demands high accuracy of the models without the need for time-consuming and computationally expensive experiments required for fine-tuning. Moreover, a model and training pipeline, which was once carefully optimized for a specific dataset, rarely generalizes well to training on a different dataset. This makes it unrealistic to have carefully fine-tuned models for each use case. To solve this, we propose an alternative approach that also forms a backbone of Intel® Geti™ platform: a dataset-agnostic template for object detection trainings, consisting of carefully chosen and pre-trained models together with a robust training pipeline for further training. Our solution works out-of-the-box and provides a strong baseline on a wide range of datasets. It can be used on its own or as a starting point for further fine-tuning for specific use cases when needed. We obtained dataset-agnostic templates by performing parallel training on a corpus of datasets and optimizing the choice of architectures and training tricks with respect to the average results on the whole corpora. We examined a number of architectures, taking into account the performance-accuracy trade-off. Consequently, we propose 3 finalists, **VFNet**, **ATSS**, and **SSD**, that can be deployed on CPU using the OpenVINO™ toolkit. The source code is available as a part of the OpenVINO™ Training Extensions^a

^ahttps://github.com/openvinotoolkit/training_extensions.

1 INTRODUCTION

Over recent years, deep learning methods have been increasingly used for Computer Vision tasks, including Object Detection. While most of the architectures are optimized on well-known benchmarks like COCO (Lin et al., 2014) (Wang et al.,) (Zhang et al., 2022), further remarkable results have been obtained using CNNs for domain-specific tasks, such as medical (Yang and Yu, 2021), traffic (Ye et al., 2020), environmental (Li et al., 2020) etc. Such domain-specific solutions, however, are usually heavily optimized for a specific target dataset, starting from carefully tailored architecture to choice of training tricks. This not only makes them time-consuming and computationally heavy to obtain but also is the reason why they *lack good domain generalization*. Training models in such a way suffers from the caveat of overly adjusting the techniques to a specific dataset. That leads to, on one hand, great predictions for a given dataset, on the other, showing poor transferability to

a different one. Meaning, such a model is unlikely to train well on a new, previously unseen dataset, without additional time-consuming and computationally expensive experiments to find the optimal tweaks of the model architecture and training pipeline. Hence, such an approach, even though used in the past with great successes, is *not flexible enough* for the rapidly evolving needs across a wide variety of computer vision use cases with a focus on time to value.

In this paper we propose an alternative approach aimed at those situations, addressing the needs of fast-paced industry and real-world applications. We provide a *dataset-agnostic template for object detection that is ready to be used straight away* and guarantees a strong baseline on a wide variety of datasets, without needing additional time-consuming and computationally demanding experiments. The template also provides the flexibility for further fine-tuning, however, it is not required. This methodology includes a carefully chosen and pre-trained model, together with a robust and efficient training pipeline for further train-

ing.

To achieve this, we base this study on a carefully chosen corpus of 11 datasets that cover a wide range of domains and sizes. They differ in both number of images as well as their complexity, including the number of classes, object sizes, distinguishability from the background, and resolution. The training procedure involves parallel experiments on all the datasets from the corpus, optimizing the model with respect to average metrics over all corpora.

To summarize, we introduce:

- a training paradigm together with a choice of a corpus of datasets,
- a universal, dataset-agnostic model template for object detection training,
- additional patience parameters for ReduceOn-Plateau and Early Stopping, to correct for the different sizes of datasets and avoid over/under-fitting,
- choice of data-agnostic augmentations and specific tricks for each of our 3 models that work on a wide range of datasets

This methodology in the Intel® Geti™ computer vision platform enables platform users to accelerate their object detection training workloads for their real-world use cases across a variety of industries.

2 RELATED WORK

2.1 Architectures

In this section, we will briefly describe the architectures we explore in our experiments.

SSD. Since its introduction, SSD (Liu et al., 2016) has been a go-to model in the industry, thanks to its high computational efficiency. As an anchor-based detector, it predicts detections on the predefined anchors. All generated predictions are classified as positive or negative, and then an extra offset regression on the positives is done for the final refinement of the bounding box location. However, anchor-based models present a serious disadvantage, they need manual tuning of the number of anchors, their ratio, and scale. This fact negatively affects their ability to generalize to different datasets.

FCOS. Anchor-free detectors address these concerns. Instead of anchor boxes, they use single points for detection, determining the center of the object. FCOS (Tian et al., 2019) approaches object detection

in a per-pixel way, making use of the multi-level predictions thanks to FPN (Feature Pyramid Network). Since the predictions are made for pixels, a new way to define positives and negatives is introduced to replace the traditional IoU-based strategy. This brings further improvements. Additionally, to suppress the low-quality detections during NMS, *centerness*, evaluating localization estimation, is used in combination with classification score as a new ranking measure.

YOLOX. (Ge et al., 2021) is a new anchor-free member of another classic family of models. As a basis it takes YOLOv3 (Redmon and Farhadi, 2018), still widely used in the industry, then boosting it for modern times with the newest tricks, including decoupled head, to deal with classification VS regression conflict, and advanced label assigning strategy, SimOTA.

ATSS. (Zhang et al., 2020) improves performance of detection, due to introduction of *Adaptive Training Sample Selection*. It is a method that allows for an automatic selection of positive and negative samples based on the statistical characteristics of objects.

VFNet. (Zhang et al., 2021) builds on FCOS and ATSS advances, topping it up with a new ranking measure for the NMS stage, to suppress low-quality detections even more efficiently. The proposed metric, *the IoU-aware classification score*, combines together information on the classification score and localization estimation in a more natural way and in only one variable. This not only gives better results in terms of the quality of detections but also avoids the overhead of the additional branch for localization score regression.

Faster R-CNN and Cascade R-CNN. All the models considered so far were one-stage detectors. Faster R-CNN (Ren et al., 2015) is a two-stage detector, where in the first stage the model coarse scans the whole image to generate region proposals, on which it concentrates in more detail in the second stage. A Region Proposal Network is introduced for the effective generation of region proposals. It is an anchor-based detector with a definition of positive and negative samples through IoU threshold. Cascade R-CNN (Cai and Vasconcelos, 2018) is a multi-stage detector addressing this issue. It consists of a sequence of detectors trained with increasing IoU thresholds.

2.2 Towards More Universal Models

Recently a lot of work has been done in the area of moving towards more flexible, generalizable models in computer vision. This ranges from multi-task training (Crawshaw, 2020), where a model is simultaneously trained to perform different tasks, to domain generalization (Gulrajani and Lopez-Paz, 2021), where the goal is to build models that can predict well on datasets out-of-distribution from those seen during training. Another interesting direction is a few-shot dataset generalisation (Triantafillou et al., 2021). In this approach parallel training on diverse datasets is used to build a universal template that can adapt to unseen data using only a few examples. Similar work to ours has been done in the classification domain (Prokofiev and Sovrasov, 2021), where adaptive training strategies are proposed and proved to be working by training lightweight models on a number of diverse image classification datasets.

3 METHOD

This section describes in detail our approach to train dataset-agnostic models from the architecture to the training tricks.

3.1 Architecture

To create a scope of the models that are applicable for the various object detection datasets regardless of difficulty and object size, we experimented with architectures in three categories: the light-weighted (about 10 GFLOPs), highly accurate, and medium ones. A detailed comparison of the different architectures used can be found in Table 2.

For the fast and medium models we took light and accurate enough MobileNetV2 backbone (Sandler et al., 2018), which is commonly used in the application field. For the accurate model we used ResNet50 (He et al., 2016) as a backbone.

For the light model we used really fast SSD with 2 heads, which can be even more optimized for CPU using OpenVINO™ toolkit; for medium - ATSS head; and for accurate - VFNet with modified deformable convolutions (DCNv2 (Zhu et al., 2019)). In our experiments, as described later, we were not limited with these architectures to achieve specific goals of performance and accuracy. For example, we also tried lighted ATSS as a candidate for the fast model, and SSD with more heads as a candidate for the medium model. But the final setup with 3 different architectures worked best.

3.2 Training Pipeline

Pretraining Weights. To achieve fast model convergence and to get high accuracy right from the start, we used pretrained weights. Firstly, we trained MobileNetV2 classification model with ImageNet21k (Ridnik et al., 2021), so a model could learn to differentiate trivial features and patterns. Afterward, we only took the trained MobileNetV2 backbone weight, cutting off the fully connected classifier head. Then, we used this pretrained backbone for our object detection model and trained ATSS-MobileNetV2 and SSD-MobileNetV2 on the COCO object detection dataset (Lin et al., 2014) with 80 annotated classes. For VFNet we used pretrained COCO weights from OpenMMLab MMDetection.

Augmentation. For augmentation, we used some classic tricks that have gained the trust of researchers. We augmented images with a random crop, horizontal flip, and brightness and color distortions. The multi-scale training was used for medium and accurate models to make them more robust. Also, after a series of experiments we carefully and empirically chose certain resolutions for each model in order to meet the trade-off between accuracy and complexity. For example, SSD performed best on the square 864x864 resolution, ATSS on the rectangle (992x736), and for VFNet we used the default high resolution without any tuning (1344x800). During resizing we did not keep the initial aspect ratio for 3 reasons:

1. to add extra regularization;
2. to avoid blank spaces after padding;
3. to avoid spending too much time on model resizing if the image aspect ratio differs a lot from the default.

The impact of these tricks on accuracy and training time can be found in the Ablation study (section 4.4).

Anchor Clustering for SSD. Despite the wide application of the SSD model, especially in production, its high inference speed, and good enough results, its dependence on careful anchor box tuning is a huge drawback. It relies on the dataset and object size characteristics. So, it becomes a challenge to build dataset-agnostic training using SSD model. For that reason, we followed the logic in (Anisimov and Khanova, 2017) and used information from the training dataset to modify the anchor boxes before training started. We collect object size statistics and cluster them using KMeans algorithm to find the optimal anchor boxes with the highest overlap among the objects. These modified anchor boxes are then passed

for training. This procedure is used for each dataset and helps to make the light model more adaptive, improving the result without additional model complexity overhead.

Early Stopping and ReduceOnPlateau. The complexity and size of the input dataset can vary drastically, making it hard to predict the number of epochs, needed to get a satisfying result and could often lead to large training times or even overfitting. To prevent it, we used Early Stopping (Prechelt, 2000) to stop training after a few epochs if those did not improve the result further. Further, we used average precision (AP@[0.5,0.95]) on the validation subset as a target metric. This problem with the unfixed number of epochs also creates uncertainty in determining learning rate schedule for the given scenario. We utilized the adaptive ReduceOnPlateau scheduler in tandem with Early Stopping, as commonly employed, to circumvent this problem. This scheduler reduces the learning rate automatically if the consecutive epochs do not lead to improvement in the target metric - average precision for the validation dataset in our case. We implemented both of these algorithms for the MMDetection (Chen et al., 2019) framework that forms the backbone of our codebase. It also powers the Intel Geti computer vision platform. It is an OpenVINO™ toolkit fork of the OpenMMLab repository.

Iteration Patience. Another challenge is that the number of iterations in the epoch differs drastically from dataset to dataset depending on its length. This creates a problem to choose an appropriate "patience" parameter for Early Stopping and ReduceOnPlateau in case of the dataset-agnostic training. For example, if we use the same patience for large datasets and small ones, the training with a small amount of data will stop too early and will suffer from undertraining. A possible solution here would be to copy the training dataset several times for every training epoch, increasing the number of iterations in the epoch. However, this will also lead to large datasets suffering from overfitting and long training times.

To overcome this, we adapted the classic approaches of ReduceOnPlateau and Early Stopping by adding **iteration patience** parameter. It works similarly to the epoch patience parameter but also ensures that a specific number of iterations were run during training on those epochs.

This approach increased the accuracy on the small datasets however it has some drawbacks as well. The iteration patience could vary from architecture to architecture and must be chosen manually, which

causes adding an extra hyperparameter. It leads to a less universal training pipeline and careful hyperparameter selection is required in case of trying new architectures, which could be time-consuming.

Another caveat is that on small datasets a large part of the training cycle is still spent on the validation and checkpoint creation because of the short epochs. It leads to the fact that up to 50% of the training is spent on not-so-useful checkpoints which are clearly undertrained and not really informative. Therefore, ideally, you should maximize the time spent on training compared to the time spent on validation and checkpoint creation.

4 EXPERIMENTS

4.1 Datasets

For training purposes, we used 11 public datasets varying in terms of the domains, number of images, classes, object sizes, overall difficulty, and horizontal/vertical alignment. Most of them were obtained from Roboflow public dataset storage. BCCD, Pothole, WGISD1(Santos et al., 2020), WGISD5(Santos et al., 2020) and Wildfire datasets represent the category with the small number of train images. Here WGISD1 and WGISD5 share the same images, but WGISD1 annotates all objects as 1 class (grape), while WGISD5 classifies each grape variety. Aerial, Dice, MinneApple(Häni et al., 2020) and PCB(Huang and Wei, 2019) represent datasets with small objects that average size is less than 5% of the image. The PKLOT and UNO are large datasets with more than 4000 images.

The samples of the images can be found in Figure 1. For illustration purposes, images were cropped to the square format. The detailed statistics with splitting into sections can be found in Table 1.

4.2 Evaluation Protocol

For validation, we used a COCO Average Precision with IoU threshold from 0.5 to 0.95 (AP@[0.5,0.95]) as the common metric for the object detection task.

$$AP = \sum_n (R_n - R_{n-1}) P_n$$

where P_n and R_n are the precision and recall at the n -th threshold. AP averages 10 values because IoU thresholds belong to [0.5,0.95] with an interval 0.05.

We collected train, validation, and test AP for each dataset, evaluating the final weights on the train,

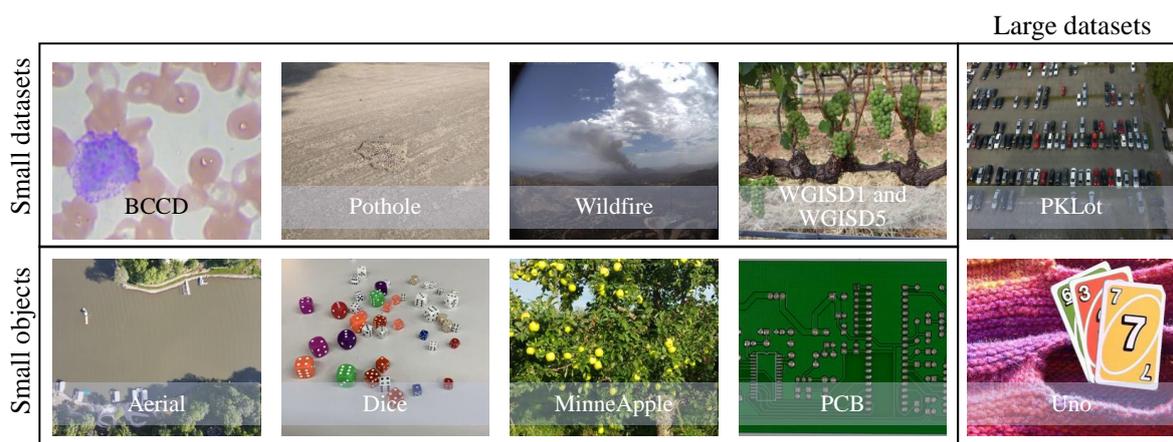


Figure 1: Samples of the Training Datasets.

Table 1: Object Detection Training Datasets.

Dataset	Number of classes	Average object size, % of image	Number of Images	
			Train	Validation
BCCD	3	17x21	255	73
Pothole	1	23x14	465	133
WGISD1	1	7x16	110	27
WGISD5	5	7x16	110	27
Wildfire	1	18x15	516	147
Aerial	5	4x6	52	15
Dice	6	5x3	251	71
MinneApple	1	4x2	469	134
PCB	6	3x3	333	222
PKLOT	2	5x8	8691	2483
UNO	14	7x7	6295	1798

validation, and test subset respectively. The difference between train and validation accuracy was used to monitor if the model was overfitted. In addition, the test AP showed how used techniques affected the accuracy on the unseen parts of the datasets and thus its generalization ability. The validation AP acted as the main metric for comparison and choosing the best approach.

To rank the different training strategies and architectures, the train, validation, and test AP scores for each dataset were averaged across the data corpus. Each of these AP scores contributes equally to the final metric without any weighting strategy. As an auxiliary metric we used the same average AP scores but on different subsets:

- on datasets with small objects where average object size is less than 5% of the image - to monitor how the network handles a hard task of the small object detection
- on large datasets that contain more than 4000 images - to monitor how the network uses the poten-

tial of the large dataset and how its size impacts the training time.

4.3 Results

After more than 300 training rounds, we have chosen the 3 best architectures in terms of accuracy-performance trade-off: the lightweight and fast, the medium, and the heavy but highly accurate. These are SSD on the MobileNetV2, ATSS on the MobileNetV2, VFNet on the ResNet50 respectively. To increase the overall accuracy and keep training time within acceptable ranges we applied some common and architecture-aware practices, which are listed below and described in detail in Training pipeline section 3.2. The contribution of each of these techniques in both improving the accuracy and reducing the training time could be found in the Ablation study (section 4.4).

To choose the best models, we also experimented with well-known architectures in the scope of the lightweight, medium, and heavy models, such as

Table 2: Architecture Comparison.

Model	AVG AP*, %	GFLOPs	AVG AP on the		
			small datasets, %	small object datasets, %	large datasets, %
<i>Fast models</i>					
YOLOX	49.6	6.45	48.8	35.1	80.7
SSD	52.0	9.36	46.9	38.7	91.2
<i>Medium models</i>					
FCOS	56.7	20.86	52.2	43.8	93.8
ATSS	60.3	20.86	55.2	49.7	94.1
<i>Accurate models</i>					
Faster R-CNN	61.8	433.40	55.4	53.7	94.0
Cascade R-CNN	62.5	488.54	56.2	55.0	93.6
VFNet	64.7	347.78	59.5	56.1	95.2

*The average AP on the validation subsets was reported here

Table 3: Detailed Results of the Chosen Models on All Datasets.

Model	Small datasets					Datasets with small objects				Large datasets	
	BCCD	Pothole	WGISD1	WGISD5	Wildfire	Aerial	Dice	MinneApple	PCB	PKLOT	UNO
<i>Fast models</i>											
YOLOX	63.5	44.7	44.1	47.7	47.7	24.2	58.2	23.0	34.9	80.9	80.5
SSD	60.0	37.5	45.9	40.2	51.0	22.7	61.3	31.1	39.8	94.7	87.7
<i>Medium models</i>											
FCOS	63.5	42.4	55.2	55.0	44.7	29.3	59.8	39.5	46.7	97.3	90.2
ATSS	63.5	47.9	57.5	55.4	51.5	36.8	73.8	41.0	47.3	97.7	90.5
<i>Accurate models</i>											
Faster R-CNN	64.2	50.2	59.4	52.3	51.1	45.3	77.8	40.9	50.8	97.3	90.7
Cascade R-CNN	65.5	51.2	60.5	53.2	50.4	45.9	79.5	42.7	51.7	95.9	91.3
VFNet	65.2	55.1	63.8	59.5	53.9	46.3	81.2	44.4	52.4	98.6	91.8

The average AP on the validation subsets was reported here

YOLOX, FCOS, Faster R-CNN, and Cascade R-CNN. We have tuned parameters, resolution, training strategy, and augmentation individually to achieve the best accuracy with comparable complexity. The accuracy and complexity of these architecture experiments could be found in Table 2.

For all of the models we used the following training setup:

- start from the weights pretrained on the COCO;
- augment images with a crop, flip, and photo distortions;
- use ReduceOnPlateau learning rate scheduler with iteration patience;
- use the Early Stopping to prevent overfitting on the large datasets and iteration patience to prevent underfitting on the small ones.

Additional special techniques included:

- for SSD: anchor box reclustering and pretraining COCO weights on ImageNet-21k(Ridnik et al., 2021)
- for ATSS: multiscale training and using ImageNet-21k weights
- VFNet - multiscale training

Detailed results of the validation accuracy of described models on all datasets are in Table 3 below.

4.4 Ablation Study

To estimate the impact of each training trick on the final accuracy, we conducted ablation experiments, by removing each of them from the training pipeline. Based on our experiments, each of these tricks im-

Table 4: Impact of Each Training Trick on the Accuracy of the Candidate Models.

Configuration	SSD	ATSS	VFNet
	AVG AP*, %	AVG AP, %	AVG AP, %
Final solution	52.0	60.3	64.7
w/o pretraining on the COCO	51.2	56.8	47.7
w/o Iteration patience	44.6	58.6	64.3
w/o ReduceOnPlateau LR scheduler	50.9	58.9	63.7
w/o Multiscale training	-	59.2	64.0
<i>Architecture aware features</i>			
w/o Anchor reclustering	48.2	-	-
w/o modified DCN	-	-	64.1

*The average AP on the validation subsets was reported here

Table 5: Impact of Each Training Trick to the Training Time* of the Candidate Models.

Configuration	SSD		ATSS		VFNet	
	AVG training time, min	AVG epochs	AVG training time, min	AVG epochs	AVG training time, min	AVG epochs
Final solution	45.18	145	36.27	92	230.91	71
w/o Early Stopping	245.20	300	306.25	300	986.96	300
w/o Tuning resolution	56.64	136	50.64	111	-	-

*GeForce RTX 3090 was used to benchmark training time.

proved the accuracy of the target metric approximately by 1 AP See results in Table 4. The iteration patience parameter showed great improvement (7.4 AP) for the fast SSD model, which stopped too early without the new parameter. The best trick for medium ATSS is using COCO weights, which improved accuracy by 3.5 AP. For the VFNet the most productive trick was to use COCO weights which increased the accuracy by 17 AP as well.

Since we focused not only on the overall accuracy but also on getting results within an acceptable time range and fast inference of the trained model, we also used some tricks to reduce the training time and complexity of the model, while preserving the accuracy of predictions. We measured time spent on training in a single GPU setup, using GeForce 3090. See Table 5 for certain numbers.

The first thing it is seen here, which is counter-intuitive, is that the small SSD model averagely requires more epochs and more time to fulfill its potential and train enough than the medium ATSS model. Talking about training time tricks, the most powerful was to use Early Stopping, which reduced the training time by 4.2-8.6 times. Also, using the standard high resolution (1344x800) of the input images increased the training time for SSD and ATSS architectures by 25-40%. However, for SSD architecture, the accuracy decreased by 0.9 AP due to the image resolution being too high for such a small model. And, for ATSS, the accuracy increased by 1.7 AP, mostly due to datasets with small objects. But it slowed inference time by

1.5 times, rendering it too slow for our criteria for a medium model.

5 CONCLUSIONS

In this research, we have outlined an alternative approach to training deep neural network models for fast-paced real-world use cases, across a variety of industries. This novel approach addresses the need for achieving high accuracy without performing time-consuming and computationally expensive training or fine-tuning for individual use cases. This methodology forms the backbone of the Intel Geti computer vision platform that enables rapidly building computer vision models for real world use cases.

As a first step, we have proposed a corpus of publicly available 11 datasets, spanning across various domains and dataset sizes. Subsequently, we have performed more than 300 experiments, each consisting of parallel trainings – one per each dataset in the selected corpus. We have selected average AP results over all datasets in the corpus as a target metric to optimize our final models.

Using this methodology, we have analyzed a number of architectures, concretely, SSD and YOLOX as fast model architectures for inference, ATSS and FCOS as medium model architectures, and VFNet, Cascade-RCNN, and Faster-RCNN as accurate models. In the process, we have identified tricks and partial optimization that helped us optimize the average

AP scores across the dataset corpus.

We encountered a significant challenge adjusting the optimal training time while building a universal template for a wide range of possible real-world use cases. This also affected the learning rate scheduler. To solve this we introduced an additional iteration patience parameters for early stopping and ReduceOnPlateau along with the epoch patience parameter. This enabled us to balance the requirements for small and large datasets while maintaining optimal training times irrespective of the dataset sizes.

Our efforts result in 3 (one per 3 different performance-accuracy regimes) dataset-agnostic templates for object detection training, that provide a strong baseline on a wide variety of datasets and can be deployed on CPU using the OpenVINO™ toolkit.

REFERENCES

- Anisimov, D. and Khanova, T. (2017). Towards lightweight convolutional neural networks for object detection. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–8. IEEE.
- Cai, Z. and Vasconcelos, N. (2018). Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162.
- Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. (2019). MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.
- Crawshaw, M. (2020). Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796*.
- Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv: Arxiv-2107.08430*.
- Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. In *International Conference on Learning Representations*.
- Häni, N., Roy, P., and Isler, V. (2020). Minneapple: a benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters*, 5(2):852–858.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.
- Huang, W. and Wei, P. (2019). A pcb dataset for defects detection and classification. *arXiv preprint arXiv:1901.08204*.
- Li, S., Yan, Q., and Liu, P. (2020). An efficient fire detection method based on multiscale feature extraction, implicit deep supervision and channel attention mechanism. *IEEE Transactions on Image Processing*, 29:8467–8475.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. (2016). Ssd: Single shot multi-box detector. In Leibe, B., Matas, J., Sebe, N., and Welling, M., editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham. Springer International Publishing.
- Prechelt, L. (2000). Early stopping - but when?
- Prokofiev, K. and Sovrasov, V. (2021). Towards efficient and data agnostic image classification training pipeline for embedded systems.
- Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv: Arxiv-1804.02767*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.
- Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. (2021). Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.
- Santos, T. T., de Souza, L. L., dos Santos, A. A., and Avila, S. (2020). Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Computers and Electronics in Agriculture*, 170:105247.
- Tian, Z., Shen, C., Chen, H., and He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635.
- Triantafyllou, E., Larochelle, H., Zemel, R., and Dumoulin, V. (2021). Learning a universal template for few-shot dataset generalization. *arXiv preprint arXiv:2105.07029*.
- Wang, C., Bochkovskiy, A., and Liao, H. Scaled-yolov4: Scaling cross stage partial network. *arxiv 2020. arXiv preprint arXiv:2011.08036*.
- Yang, R. and Yu, Y. (2021). Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis. *Frontiers in Oncology*, 11:573.
- Ye, T., Zhang, X., Zhang, Y., and Liu, J. (2020). Railway traffic object detection using differential feature fusion convolution neural network. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1375–1387.
- Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., and Shum, H.-Y. (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

- Zhang, H., Wang, Y., Dayoub, F., and Sunderhauf, N. (2021). Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8514–8523.
- Zhang, S., Chi, C., Yao, Y., Lei, Z., and Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Zhu, X., Hu, H., Lin, S., and Dai, J. (2019). Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9308–9316.

