

DynamicDet: A Unified Dynamic Architecture for Object Detection

Zhihao Lin Yongtao Wang[†] Jinhe Zhang Xiaojie Chu
 Wangxuan Institute of Computer Technology, Peking University

linzhihao@stu.pku.edu.cn, wyt@pku.edu.cn

jinhezhang17@gmail.com, chuxiaojie@stu.pku.edu.cn

Abstract

Dynamic neural network is an emerging research topic in deep learning. With adaptive inference, dynamic models can achieve remarkable accuracy and computational efficiency. However, it is challenging to design a powerful dynamic detector, because of no suitable dynamic architecture and exiting criterion for object detection. To tackle these difficulties, we propose a dynamic framework for object detection, named *DynamicDet*. Firstly, we carefully design a dynamic architecture based on the nature of the object detection task. Then, we propose an adaptive router to analyze the multi-scale information and to decide the inference route automatically. We also present a novel optimization strategy with an exiting criterion based on the detection losses for our dynamic detectors. Last, we present a variable-speed inference strategy, which helps to realize a wide range of accuracy-speed trade-offs with only one dynamic detector. Extensive experiments conducted on the COCO benchmark demonstrate that the proposed *DynamicDet* achieves new state-of-the-art accuracy-speed trade-offs. For instance, with comparable accuracy, the inference speed of our dynamic detector *Dy-YOLOv7-W6* surpasses *YOLOv7-E6* by 12%, *YOLOv7-D6* by 17%, and *YOLOv7-E6E* by 39%. The code is available at <https://github.com/VDIGPKU/DynamicDet>.

1. Introduction

Object detection is an essential topic in computer vision, as it is a fundamental component for other vision tasks, e.g., autonomous driving [26, 40, 56], multi-object tracking [52, 57], intelligent transportation [36, 55], etc. In recent years, tremendous progress has been made toward more accurate and faster detectors, such as Network Architecture Search (NAS)-based detectors [10, 25, 48] and YOLO series models [2, 9, 11, 21, 44, 45]. However, these methods need to design and train multiple models to achieve a few good trade-offs between accuracy and speed, which is not flexible enough for various application scenarios. To alleviate this

[†]Corresponding author.

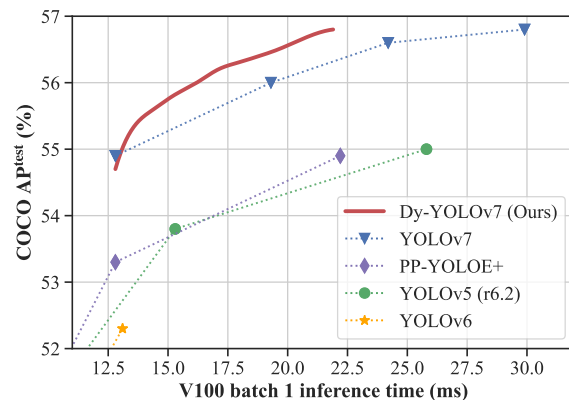


Figure 1. Comparison of the proposed dynamic detectors and other efficient object detectors. Our method can achieve a wide range of state-of-the-art trade-offs between accuracy and speed with a single model.

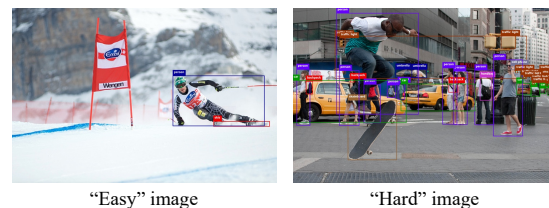


Figure 2. Examples of “easy” and “hard” images for the object detection task.

problem, we focus on dynamic inference for the object detection task, and attempt to use only one dynamic detector to achieve a wide range of good accuracy-speed trade-offs, as shown in Fig. 1.

The human brain inspires many fields of deep learning, and the dynamic neural network [12] is a typical one. As two examples shown in Fig. 2, we can quickly identify all objects on the left “easy” image, while we need more time to achieve the same effect for the right one. In other words, the processing speeds of images are different in our brains [18, 34], which depend on the difficulties of the images. This property motivates the *image-wise* dynamic neural network, and many exciting works have been proposed

(*e.g.*, Branchynet [43], MSDNet [17], DVT [50]). Although these approaches have achieved remarkable performance, they are all designed specifically for the image classification task and are not suitable for other vision tasks, especially for the object detection [12]. The main difficulties in designing an *image-wise* dynamic detector are as follows.

Dynamic detectors cannot utilize the existing dynamic architectures. Most existing dynamic architectures are cascaded with multiple stages (*i.e.*, a stack of multiple layers) [17, 20, 33, 54], and predict whether to stop the inference at each exiting point. Such a paradigm is feasible in image classification but is ineffective in object detection, since an image has multiple objects and each object usually has different categories and scales, as shown in Fig. 2. Hence, almost all detectors depend heavily on multi-scale information, utilizing the features on different scales to detect objects of different sizes (which are obtained by fusing the multi-scale features of the backbone with a detection neck, *i.e.*, FPN [27]). In this case, the exiting points for detectors can only be placed behind the last stage. Consequently, the entire backbone module has to be run completely [58], and it is impossible to achieve dynamic inference on multiple cascaded stages.

Dynamic detectors cannot exploit the existing exiting criteria for image classification. For the image classification task, the threshold of top-1 accuracy is a widely used criterion for decision-making [17, 50]. Notably, it only needs one fully connected layer to predict the top-1 accuracy at intermediate layer, which is easy and costless. However, object detection task requires the neck and the head to predict the categories and locations of the object instances [3, 14, 27, 39]. Hence, the existing exiting criteria for image classification is not suitable for object detection.

To deal with the above difficulties, we propose a dynamic framework to achieve dynamic inference for object detection, named DynamicDet. Firstly, We design a dynamic architecture for the object detection task, which can exit with multi-scale information during the inference. Then, we propose an adaptive router to choose the best route for each image automatically. Besides, we present the corresponding optimization and inference strategies for the proposed DynamicDet.

Our main contributions are as follows:

- We propose a dynamic architecture for object detection, named DynamicDet, which consists of two cascaded detectors and a router. This dynamic architecture can be easily adapted to mainstream detectors, *e.g.*, Faster R-CNN and YOLO.
- We propose an adaptive router to predict the difficulty scores of the images based on the multi-scale features, and achieve automatic decision-making. In addition, we propose a hyperparameter-free optimization strategy and a variable-speed inference strategy for our dy-

amic architecture.

- Extensive experiments show that DynamicDet can obtain a wide range of accuracy-speed trade-offs with only one dynamic detector. We also achieve new state-of-the-art trade-offs for real-time object detection (*i.e.*, 56.8% AP at 46 FPS).

2. Related work

2.1. Backbone design on object detection

Backbones play a crucial role in object detectors since the performance of detectors highly relies on the multi-scale features extracted by the backbones [4]. ResNet [15] and its variants (*e.g.*, ResNeXt [51], Res2Net [8]) introduce the residual connection to neural networks, providing a high-quality backbone architecture family for all vision tasks. Further, to reduce the calculation load, CSPNet [46] cuts down the duplicate gradient information to reduce the heavy inference, improving the efficiency significantly. Its effective architecture also inspires many lightweight detectors (*e.g.*, YOLO series models [2, 9, 11, 21, 44, 45]). Then, some transformer-based backbones (*e.g.*, PVT [47], Swin Transformer [31]) are proposed to learn the global information better. In addition, many auto-designed backbones [4, 6, 19, 41] for object detection are proposed. For example, DetNAS [4] utilizes the one-shot supernet to search the optimal backbone, with the guidance of the object detection task.

Although many kinds of backbones have been proposed, almost all of them are single-pass architectures, which sequentially produce one set of multi-scale features. Thus, all stages of them cannot be skipped. Fortunately, some works propose the architectures of multiple cascaded backbones, which have the potential to be converted as a dynamic backbone for object detection. For example, CBNNet [24, 30] groups multiple identical backbones with composite connections, constructing a more powerful composite backbone. Since these backbones have multiple sub-backbones and each of them can produce intermediate multi-scale features, we can add some exiting points after each sub-backbone for dynamic inference.

2.2. Accuracy-speed trade-off on object detection

Almost all detection methods are designed for a better accuracy-speed trade-off, *i.e.*, more accurate and faster. With a given detector, the simplest way to obtain an accuracy-speed trade-off is to adopt the model scaling techniques [42, 44, 45] (*e.g.*, increasing the channel size or repeating the layers). EfficientDet [42] uniformly scales the resolution, depth, and width for all modules simultaneously, achieving remarkable efficiency on real-time detectors. Scaled-YOLOv4 [44] modifies not only the depth, width, and resolution but also the structure of the network to pursue a better trade-off. YOLOv7 [45] designs a

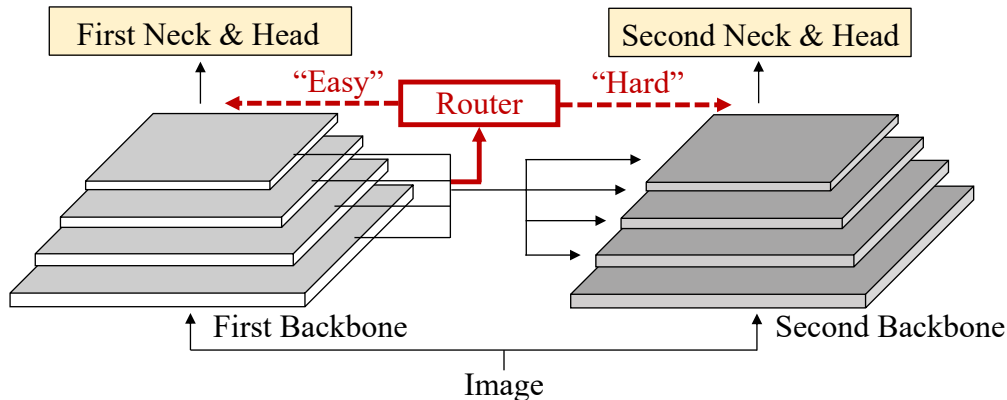


Figure 3. Illustration of the architecture of DynamicDet. The first backbone extracts the multi-scale features from the input image firstly. Then, the router will be fed with these multi-scale features to predict the difficulty of image and to decide the inference route. Notably, the “easy” images will be processed by only one backbone, while the “hard” images will be processed by two.

compound scaling method for concatenation-based models, achieving new state-of-the-art trade-offs. EAUTOdet [48] constructs a supernet and adopts Network Architecture Search (NAS) to automatically search for suitable scaling factors under different hardware constraints. However, all the above methods need to train multiple detectors for the best trade-offs (e.g., one tiny model for real-time detection and another large model for accurate detection), leading to colossal training resources. In this paper, we focus on dynamic inference, aiming to achieve a wide range of best accuracy-speed trade-offs with only one dynamic detector.

2.3. Dynamic neural network

The dynamic neural network can achieve adaptive computation for different images (i.e., *image-wise* [17, 20, 22, 23, 43, 49, 50, 54]) or pixels (i.e., *spatial-wise* [7, 13, 37]). SACT [7] is a classic *spatial-wise* dynamic network, which adaptively adjusts the number of executed layers for the regions of the image, to improve the efficiency of networks. Its practical speed-up performance highly relies on the hardware-software co-design [12]. However, the current deep learning hardware and libraries [1, 35] are not friendly to these *spatial-wise* dynamic networks [12]. On the contrary, *image-wise* dynamic networks do not rely on sparse computing and can be easily accelerated on the conventional CPUs and GPUs [5]. Branchynet [43] introduces the early exiting strategy, which enables the model to exit from the intermediate layer whenever the model is confident enough. MSDNet [17] and its variants [23, 54] develop a multi-classifier architecture for the image classification task. DVT [50] cascades multiple transformers with increasing numbers of tokens and activates them sequentially to achieve dynamic inference. However, these methods are all designed specifically for the image classification task and cannot be applied to other vision tasks, such as object detection.

The closest work to our DynamicDet is Adaptive Feed-

ing [58]. In Adaptive Feeding [58], each image is detected by a lightweight detector (e.g., Tiny YOLO [38]) and then classified as easy or hard by a linear support vector machine (SVM) with those detected results. Then, the easy images will go through a fast detector (e.g., SSD300 [29]), while the hard images will go through a more accurate but slower one (e.g., SSD500 [29]). Adaptive Feeding [58] introduces the above multi-stage process for dynamic inference, which is inefficient and not elegant. In comparison, the proposed DynamicDet cascades two detectors and a classifier (i.e., the router), yielding a more unified and efficient dynamic detector.

3. Approach

In the following, we elaborate on our dynamic architecture for object detection. We first introduce the overall architecture in Sec. 3.1. Then, we state the proposed adaptive router, i.e., the decision maker of DynamicDet in Sec. 3.2. Finally, we introduce the optimization strategy and a variable-speed inference strategy in Secs. 3.3 and 3.4.

3.1. Overall architecture

The overall architecture of our dynamic detector is shown in Fig. 3. Inspired by CBNNet [24, 30], our dynamic architecture consists of two detectors and one router. For an input image \mathbf{x} , we initially extract its multi-scale features F_1 with the first backbone \mathcal{B}_1 as

$$F_1 = \mathcal{B}_1(\mathbf{x}) = [f_1^{\{1\}}, f_1^{\{2\}}, \dots, f_1^{\{L\}}], \quad (1)$$

where L denotes the number of stages, i.e., the number of multi-scale features. Then, the router \mathcal{R} will be fed with these features F_1 to predict a difficulty score $\phi \in (0, 1)$ for this image as

$$\phi = \mathcal{R}(F_1). \quad (2)$$

Generally speaking, the “easy” images exit at the first backbone, while the “hard” images require the further processing. Specifically, if the router classifies the input image as

an “easy” one, the followed neck and head \mathcal{D}_1 will output the detection results \mathbf{y} as

$$\mathbf{y} = \mathcal{D}_1(F_1). \quad (3)$$

On the contrary, if the router classifies the input image as a “hard” one, the multi-scale features will need further enhancement by the second backbone, instead of immediately decoded by \mathcal{D}_1 . In particular, we embed the multi-scale features F_1 into H by a composite connection module \mathcal{G} as

$$H = \mathcal{G}(F_1) = [h^{\{1\}}, h^{\{2\}}, \dots, h^{\{L\}}], \quad (4)$$

where \mathcal{G} is the DHL of CBNNet [24, 30] in our implementation. Then, we feed the input image \mathbf{x} into the second backbone and enhance the features of the second backbone via summing the corresponding elements of H at each stage sequentially, denoted as

$$F_2 = \mathcal{B}_2(\mathbf{x}, H) = [f_2^{\{1\}}, f_2^{\{2\}}, \dots, f_2^{\{L\}}], \quad (5)$$

and the detection results will be obtained by the second head and neck \mathcal{D}_2 as

$$\mathbf{y} = \mathcal{D}_2(F_2). \quad (6)$$

Through the above process, the “easy” images will be processed by only one backbone, while the “hard” images will be processed by two. Obviously, with such an architecture, trades-offs between computation (*i.e.*, speed) and accuracy can be achieved.

3.2. Adaptive router

In mainstream object detectors, different scale features play different roles. Generally, the features of the shallow layers, with strong spatial information and small receptive fields, are more used to detect small objects. In contrast, the features of the deep layers, with strong semantic information and large receptive fields, are more used to detect large objects. This property makes it necessary to consider multi-scale information when predicting the difficulty score of an image. According to this, we design an adaptive router based on the multi-scale features, that is, a simple yet effective decision-maker for the dynamic detector.

Inspired by the squeeze-and-excitation (SE) module [16], we first pool the multi-scale features F_1 independently and concatenate them all as

$$\tilde{F}_1 = \mathcal{C}([\mathcal{P}(f_1^{\{1\}}), \mathcal{P}(f_1^{\{2\}}), \dots, \mathcal{P}(f_1^{\{L\}})]), \quad (7)$$

where \mathcal{P} denotes the global average pooling and \mathcal{C} denotes the channel-wise concatenation. With this operation, we compress the multi-scale features F_1 into a vector $\tilde{F}_1 \in \mathbb{R}^d$ of dimension d . Then, we map this vector to a difficulty score $\phi \in (0, 1)$ via two learnable fully connected layers as

$$\phi = \sigma(W_2(\delta(W_1\tilde{F}_1 + b_1)) + b_2), \quad (8)$$

where δ, σ denote the ReLU and Sigmoid activation functions respectively, and W_1, W_2, b_1, b_2 are learnable parameters. Following [59], we reduce the feature dimension to

$\lfloor d/4 \rfloor$ in the first fully connected layer, and exploit the second fully connected layer with a Sigmoid function to generate the predicted score. It is worth noting that the computational burden of our router can be negligible since we first pool all multi-scale features to one vector.

3.3. Optimization strategy

In this section, we describe the optimization strategy for the above dynamic architecture.

Firstly, we jointly train the cascaded detectors, and the training objective is

$$\min_{\Theta_1, \Theta_2} (\mathcal{L}_{det}^{\{1\}}(\mathbf{x}, \mathbf{y}|\Theta_1) + \mathcal{L}_{det}^{\{2\}}(\mathbf{x}, \mathbf{y}|\Theta_2)), \quad (9)$$

where \mathbf{x}, \mathbf{y} denote the input image and the ground truth respectively, Θ_i denotes the learnable parameters of the detector i and $\mathcal{L}_{det}^{\{i\}}$ denotes the training loss for detector i (*e.g.*, bounding box regression loss and classification loss). After the above training phase, these two detectors will be able to detect the objects, and we freeze their parameters Θ_1, Θ_2 during the later training.

Then, we train the adaptive router to automatically distinguish the difficulty of the image. Here, we assume the parameters of the router are $\Theta_{\mathcal{R}}$ and the predicted difficulty score obtained from the Eq. (8) is ϕ . We hope the router can assign the “easy” images (*i.e.*, with lower ϕ) to the faster detector (*i.e.*, the first detector) and the “hard” images (*i.e.*, with higher ϕ) to the more accurate detector (*i.e.*, the second detector).

However, it is non-trivial to implement that in practice. If we directly optimize the router without any constraints as

$$\min_{\Theta_{\mathcal{R}}} ((1 - \phi)\mathcal{L}_{det}^{\{1\}}(\mathbf{x}, \mathbf{y}|\Theta_1) + \phi\mathcal{L}_{det}^{\{2\}}(\mathbf{x}, \mathbf{y}|\Theta_2)), \quad (10)$$

the router will always choose the most accurate detector as it allows for a lower training loss. Furthermore, if we naively add hardware constraints to the training objective as

$$\min_{\Theta_{\mathcal{R}}} ((1 - \phi)\mathcal{L}_{det}^{\{1\}}(\mathbf{x}, \mathbf{y}|\Theta_1) + \phi\mathcal{L}_{det}^{\{2\}}(\mathbf{x}, \mathbf{y}|\Theta_2) + \lambda\phi), \quad (11)$$

we will have to adjust the hyperparameter λ by try and error, leading to huge workforce consumption.

To overcome the above challenges, we propose a hyperparameter-free optimization strategy for our adaptive router. First, we define the difficulty criterion based on the corresponding training loss difference between two detectors of an image, as shown in Fig. 4. Specifically, we assume that if the loss difference of an image between two detectors is small enough, this image can be classified as an “easy” image. Instead, if the loss difference is large enough, it should be classified as a “hard” image. Ideally, for a balanced situation, we hope the easier half of all images go through the first detector, and the harder half go through the

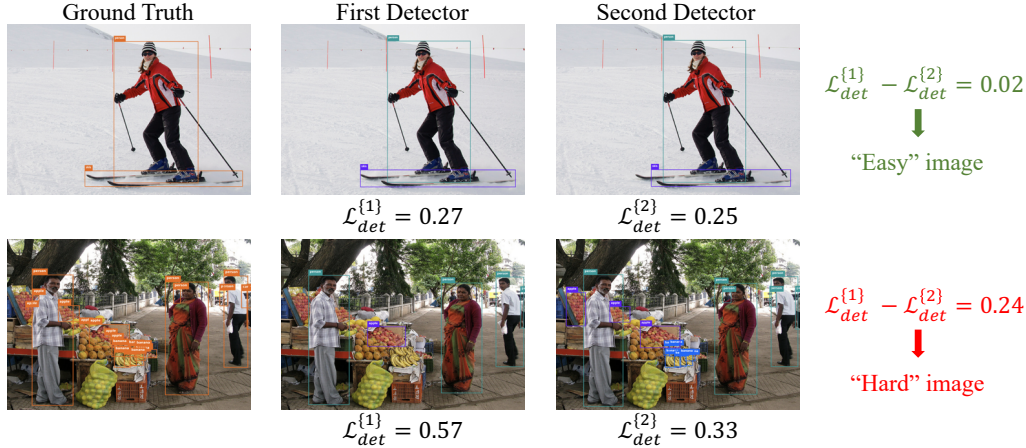


Figure 4. Illustration of the difficulty criterion based on the training loss difference between two cascaded detectors. For the top image, the loss difference between the first detector and the second detector is very small, so it should be classified as an “easy” image. On the contrary, the loss difference of the bottom image is large, so it should be classified as a “hard” image.

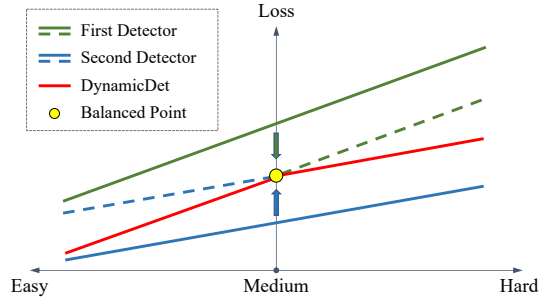


Figure 5. Qualitative analysis of the loss of two cascaded detectors on the images with different difficulties. With the proposed adaptive offset in our optimization strategy, the loss curves of two detectors intersect and reveal the optimal curve of DynamicDet.

second one. To achieve this, we introduce an adaptive offset to balance the losses of two detectors and optimize our router via gradient descent. In practice, we first calculate the median of the training loss difference Δ between the first and the second detector on the training set. Then, the training objective of our router can be formulated as

$$\min_{\Theta_{\mathcal{R}}} ((1-\phi)(\mathcal{L}_{det}^{\{1\}}(\mathbf{x}, \mathbf{y}|\Theta_1) - \Delta/2) + \phi(\mathcal{L}_{det}^{\{2\}}(\mathbf{x}, \mathbf{y}|\Theta_2) + \Delta/2)), \quad (12)$$

where $\Delta/2$ is used to reward the first detector and punish the second detector, respectively. As the qualitative analysis shown in Fig. 5, without this reward and penalty, the losses of the second detector are always smaller than the first detector. When the reward and penalty are conducted, their loss curves intersect and reveal the optimal curve.

Our training objective provides a means to optimize the adaptive router by introducing the following gradient through the difficulty score ϕ to all parameters $\Theta_{\mathcal{R}}$ of the

router as

$$\frac{\partial \mathcal{L}}{\partial \Theta_{\mathcal{R}}} = \frac{\partial \mathcal{L}}{\partial \phi} \frac{\partial \phi}{\partial \Theta_{\mathcal{R}}} = -\frac{\partial \phi}{\partial \Theta_{\mathcal{R}}} (\mathcal{L}_{det}^{\{1\}} - \mathcal{L}_{det}^{\{2\}} - \Delta). \quad (13)$$

To distinguish between “easy” and “hard” images better, we expect the optimization direction of the router to be related to the difficulty of the image, *i.e.*, the difference in loss between the two detectors. Obviously, the gradient at Eq. (13) enable such expectation.

3.4. Variable-speed inference

We further propose a simple and effective method to determine the difficulty score thresholds to achieve variable-speed inference with only one dynamic detector. Specifically, our adaptive router will output a difficulty score and decide which detector to go through based on a certain threshold during inference. Therefore, we can set different thresholds to achieve different accuracy-speed trade-offs. Firstly, we count the difficulty scores S_{val} of the validation set. Then, based on the actual needs (*e.g.*, the target latency), we can obtain the corresponding threshold for our router. For example, assuming the latency of the first detector is lat_1 , the latency of the cascaded two detectors is lat_2 and the target latency is lat_t , we can calculate the maximum allowable proportion of the “hard” images k as

$$k = \frac{lat_t - lat_1}{lat_2 - lat_1}, \quad lat_1 \leq lat_t \leq lat_2, \quad (14)$$

and then the threshold τ_{val} will be

$$\tau_{val} = \text{percentile}(S_{val}, k), \quad (15)$$

where $\text{percentile}(\cdot, k)$ means to compute the k -th quantile of the data. It is worth noting that this threshold τ_{val} is robust in both validation set and test set because these two sets are independent and identically distributed (*i.e.*, i.i.d.).

Model	Size	FLOPs	FPS	AP
EAutoDet-X [48]	640	225.3G	41 [†]	49.2
YOLOX-L [9]	640	155.6G	69 [†]	50.1
YOLOX-X [9]	640	281.9G	58 [†]	51.5
YOLOv5-L (r6.2) [11]	640	109.1G	114	49.0
YOLOv5-X (r6.2) [11]	640	205.7G	100	50.9
YOLOv6-M [21]	640	82.2G	109	49.6
YOLOv6-L [21]	640	144.0G	76	52.4
PP-YOLOE+-M [53]	640	49.9G	123 [†]	50.0
PP-YOLOE+-L [53]	640	110.1G	78 [†]	53.3
PP-YOLOE+-X [53]	640	206.6G	45 [†]	54.9
YOLOv7 [45]	640	104.7G	114	51.4
Dy-YOLOv7 / 10	640	112.4G	110	52.1
Dy-YOLOv7 / 50	640	143.2G	96	53.3
Dy-YOLOv7 / 90	640	174.0G	85	53.8
Dy-YOLOv7 / 100	640	181.7G	83	53.9
YOLOv7-X [45]	640	189.9G	105	53.1
Dy-YOLOv7-X / 10	640	201.7G	98	53.3
Dy-YOLOv7-X / 50	640	248.9G	78	54.4
Dy-YOLOv7-X / 90	640	296.1G	65	55.0
Dy-YOLOv7-X / 100	640	307.9G	64	55.0
YOLOv5-M6 (r6.2) [11]	1280	200.0G	96	51.4
YOLOv5-L6 (r6.2) [11]	1280	445.6G	65	53.8
YOLOv5-X6 (r6.2) [11]	1280	839.2G	39	55.0
YOLOv7-W6 [45]	1280	360.0G	78	54.9
YOLOv7-E6 [45]	1280	515.2G	52	56.0
YOLOv7-D6 [45]	1280	806.8G	41	56.6
YOLOv7-E6E [45]	1280	843.2G	33	56.8
Dy-YOLOv7-W6 / 10	1280	384.2G	74	55.2
Dy-YOLOv7-W6 / 50	1280	480.8G	58	56.1
Dy-YOLOv7-W6 / 90	1280	577.4G	48	56.7
Dy-YOLOv7-W6 / 100	1280	601.6G	46	56.8

[†] The FPS marked with † are from the corresponding papers, and others are measured on the same machine with 1 NVIDIA V100 GPU.

Table 1. Comparison with the state-of-the-art real-time object detectors on COCO *test-dev*.

Based on the above strategy, one dynamic detector can directly cover the accuracy-speed trade-offs from the single to double detectors, avoiding redesigning and training multiple detectors under different hardware constraints.

4. Experiments

In this section, we evaluate our DynamicDet through extensive experiments. In Sec. 4.1, we detail the experimental setups. In Sec. 4.2, we compare our DynamicDet with the state-of-the-art real-time detectors. In Sec. 4.3, we present the experimental results on two-stage detectors with CNN- and transformer-based backbones to demonstrate the generality of DynamicDet over different backbones and detectors. In Sec. 4.4, we ablate each component of DynamicDet

in detail. In Sec. 4.5, we visualize the “easy” and the “hard” images determined by the adaptive router.

4.1. Experimental setups

We conduct experiments on the COCO [28] benchmark. All the models presented are trained on the 118k training images, and tested on the 5k *minival* images and 20k *test-dev* images. We choose the YOLOv7 [45] series models as the real-time detector baseline, and the Faster R-CNN [39] (ResNet [15]) and the Mask R-CNN [14] (Swin Transformer [31]) as the two-stage detector baselines. All dynamic detectors are trained with the same hyper-parameters of their corresponding baselines. We use brief notation to indicate the easy-hard proportion for each dynamic detector: for instance, “Dy-YOLOv7-X/10” means the dynamic YOLOv7-X model with 10% images are classified as “hard” and the rest are classified as “easy”. The training of the adaptive router is conducted on a single GPU with batch-size 1 and two epochs, utilizing the AdamW [32] optimizer with a constant learning rate 1×10^{-5} and weight decay 5×10^{-3} . The reported FLOPs for dynamic detectors are the average FLOPs on the corresponding dataset. The speed performance is measured on a machine with 1 NVIDIA V100 GPU unless otherwise stated. The implementation of Dy-YOLOv7 is developed by the YOLOv7 [45] framework, with two identical detectors. The implementation of dynamic two-stage detectors is developed by the open-source CBNet [24] framework, with two identical backbones and a shared neck and head.

4.2. Comparison with the state-of-the-arts

As shown in Tab. 1, compared with the state-of-the-art high-performance real-time object detectors, our dynamic detectors obtain better results and achieve the new state-of-the-art accuracy-speed trade-offs. Specifically, Dy-YOLOv7-W6 / 50 achieves 56.1% AP with 58 FPS, which is 0.1% more accurate and 12% faster than YOLOv7-E6. Dy-YOLOv7-W6 / 100 achieves 56.8% AP with 46 FPS, which is 39% faster than YOLOv7-E6E with a similar accuracy. It is worth noting that *these trade-offs are obtained by only one dynamic detector instead of multiple independent models*.

4.3. Generality for two-stage detectors

We conduct experiments on two classic two-stage detectors (*i.e.*, Faster R-CNN [39], Mask R-CNN [14]) to show the generality of our DynamicDet. As shown in Tab. 2, our method is compatible with two-stage detectors and can also improve the accuracy-speed performance of baselines. For example, Dy-Faster R-CNN ResNet50 / 90 boosts the bbox AP by 1% with the comparable inference speed for Faster R-CNN ResNet101. Furthermore, DynamicDet is also compatible with transformer-based backbones (*e.g.*, Swin Transformer [31]). Dy-Mask R-CNN Swin-T / 90 improves the bbox AP to 49.9% with the comparable inference

Model	FLOPs	FPS	AP _{box}	AP _{mask}
Faster R-CNN ResNet50 [15, 39]	207.1G	23	37.4	-
Faster R-CNN ResNet101 [15, 39]	283.1G	18	39.4	-
Dy-Faster R-CNN ResNet50 / 50	245.4G	20	39.5	-
Dy-Faster R-CNN ResNet50 / 90	276.0G	17	40.4	-
Mask R-CNN Swin-T [14, 31]	263.8G	15	46.0	41.6
Mask R-CNN Swin-S [14, 31]	353.8G	12	48.2	43.2
Dy-Mask R-CNN Swin-T / 50	310.6G	12	48.7	43.6
Dy-Mask R-CNN Swin-T / 90	348.0G	11	49.9	44.2

Table 2. Comparison with two-stage detectors on COCO *minival*.

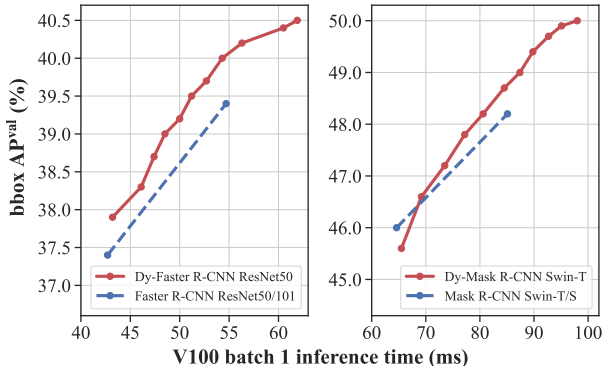


Figure 6. Bounding box mAP v.s. inference speed for two-stage detectors on COCO *minival*.

speed of Mask R-CNN Swin-S. Notably, our two-stage dynamic detector can also perform variable-speed inference as illustrated in Fig. 6.

4.4. Ablation study

4.4.1 Lightweight adaptive router

The FLOPs ratios for the adaptive router in different models are presented in Tab. 3. We can find that this ratio is less than 0.002% in all models, demonstrating that the computational burden of the adaptive router can be negligible. This lightweight router avoids slowing down the detection process and ensures the fast decision-making for dynamic inference.

4.4.2 Effective training strategy for adaptive router

We ablate the effectiveness of the proposed training and optimization strategy for adaptive router. We first train a Mask R-CNN [14] with cascaded Swin-T [31] as our baseline detector. Then, we apply three strategies to achieve the decision-making for router: random, AP-based (*i.e.*, dividing “easy” and “hard” images based on the validation accuracy and using them to train the router, similar to Adaptive Feeding [58]), and our proposed strategy. As shown in Fig. 7, we compare the bbox AP on the *test-dev* set of different training strategies. It is shown that our optimization strategy outperforms another two strategies under all latency constraints. Taking the detector with 84.5 ms la-

Model	Router	Total	Ratio
Dy-YOLOv7	2.1M	104.7G	0.0020%
Dy-YOLOv7-W6	1.9M	360.0G	0.0005%
Dy-Faster R-CNN ResNet50	3.7M	283.7G	0.0013%
Dy-Mask R-CNN Swin-T	0.5M	357.4G	0.0001%

Table 3. Comparison of the adaptive router’s FLOPs and the total FLOPs in different dynamic models.

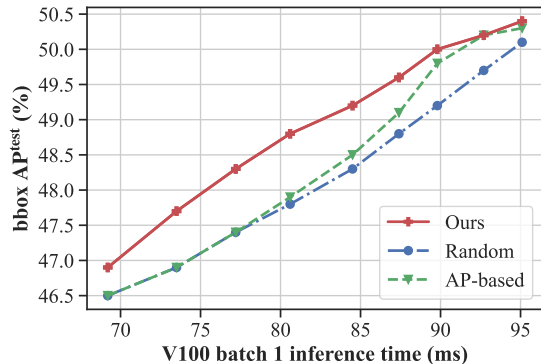


Figure 7. Comparison of the proposed strategy and other two decision-making strategies.

tency (*i.e.*, 50% easy and 50% hard) as an example, our strategy exceeds random selection 0.9% AP and AP-based strategy 0.7% AP. This proves that our optimization strategy effectively improves the discrimination accuracy of the router and outperforms AP-based strategy [58].

4.4.3 Robust variable-speed inference strategy

To achieve variable-speed inference for a dynamic detector, we count the difficulty scores on the validation set and directly adopt the corresponding thresholds for the test set. This strategy requires the validation set to be large enough. However, with custom datasets, this is not always sufficient. To demonstrate the robustness of our variable-speed inference strategy, we analyze the impact of the validation set size on the threshold consistency between the validation set and test set. Taking the Dy-Mask R-CNN Swin-T on COCO [28] dataset as an example, its threshold for 50% quantile on the test set is 0.62. Then, we count the thresholds for 50% quantile on the validation set of different sizes (*i.e.*, 0.5k, 1k, ..., 5k). As shown in Fig. 9, the threshold obtained from 5k validation images is consistent with the threshold of the test set, which confirms our assumption in Sec. 3.4. Later, as the data size decreases, the thresholds of the validation set change within a small range. However, when the data size is less than 1.5k, the threshold of the validation set and the test set will occur a large deviation (*i.e.*, 0.11 at 1k). Overall, our variable-speed inference strategy is stable when the validation set size is relatively sufficient (*e.g.*, about 2k validation images for the 20k test set on COCO [28]).



Figure 8. Visualization of the “easy” and the “hard” images. The horizontal direction corresponds to the difficulty scores predicted by our adaptive router in Dy-Mask R-CNN Swin-T.

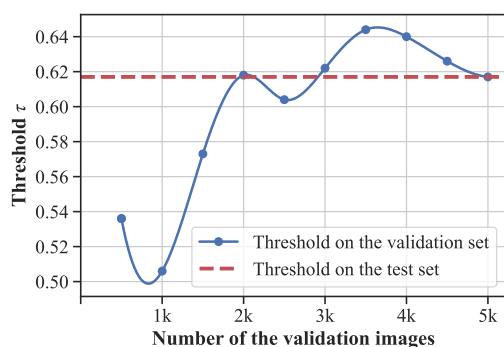


Figure 9. Comparison of the thresholds obtained from the test set and the validation sets of different sizes.

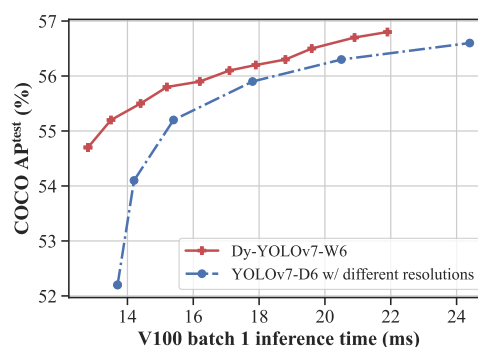


Figure 10. Comparison of the trade-offs obtained from our Dy-YOLOv7-W6 and YOLOv7-D6 with different resolutions.

4.4.4 Comparison with the trade-offs obtained by adjusting the input resolution

For a well-trained detector, changing its input resolution can also quickly obtain a series of accuracy-speed trade-offs. Here we compare this method with our dynamic detector. As shown in Fig. 10, we compare our Dy-YOLOv7-W6 and the YOLOv7-D6 with different input resolutions (*i.e.*, 640~1280), and we observe that our dynamic detector achieves better accuracy-speed trade-offs. For example, our Dy-YOLOv7-W6 achieves 55.2% AP at 74 FPS (13.5 ms), while YOLOv7-D6 with 640 input resolution only achieves 52.2% AP at an even slower inference speed.

4.5. Visualization of images with different difficulty scores

We depict the images with different predicted difficulty scores in Fig. 8, ascending from left to right. That is, the images on the left are considered as the “easy” images, while those on the right are considered as the “hard” images. We can observe that the “easy” images usually contain fewer objects, with the usual camera viewpoint and the clean background. In contrast, the “hard” images usually have more complex scenes with severe occlusion and much more small objects.

5. Conclusion

In this paper, we present a unified dynamic architecture for object detection, DynamicDet. We first design a dynamic architecture to support dynamic inference on mainstream detectors. Then, we propose an adaptive router to predict the difficulty score of each image and determine the inference route. With the above architecture and router, we then propose a hyperparameter-free optimization strategy with an adaptive offset to training our dynamic detectors. Last, we present a variable-speed inference strategy. With the settable threshold for dynamic inference, we can achieve a wide range of accuracy-speed trade-offs with only one dynamic detector. Extensive experimental results demonstrate the superiority of the proposed DynamicDet in accuracy and efficiency, and new state-of-the-art accuracy-speed trade-offs are achieved.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grant 62176007. This work was also a research achievement of Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology).

References

- [1] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016. 3
- [2] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1, 2
- [3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, pages 6154–6162, 2018. 2
- [4] Yukang Chen, Tong Yang, Xiangyu Zhang, Gaofeng Meng, Xinyu Xiao, and Jian Sun. Detnas: Backbone search for object detection. *NeurIPS*, 32, 2019. 2
- [5] Jack Choquette, Wishwesh Gandhi, Olivier Giroux, Nick Stam, and Ronny Krashinsky. Nvidia a100 tensor core gpu: Performance and innovation. *IEEE Micro*, 41(2):29–35, 2021. 3
- [6] Xianzhi Du, Tsung-Yi Lin, Pengchong Jin, Golnaz Ghiasi, Mingxing Tan, Yin Cui, Quoc V Le, and Xiaodan Song. Spinenet: Learning scale-permuted backbone for recognition and localization. In *CVPR*, pages 11592–11601, 2020. 2
- [7] Michael Figurnov, Maxwell D Collins, Yukun Zhu, Li Zhang, Jonathan Huang, Dmitry Vetrov, and Ruslan Salakhutdinov. Spatially adaptive computation time for residual networks. In *CVPR*, pages 1039–1048, 2017. 3
- [8] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(2):652–662, 2019. 2
- [9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 1, 2, 6
- [10] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *CVPR*, pages 7036–7045, 2019. 1
- [11] Jocher Glenn. Yolov5 release v6.2, 2022. <https://github.com/ultralytics/yolov5/releases/tag/v6.2>. 1, 2, 6
- [12] Yizeng Han, Gao Huang, Shiji Song, Le Yang, Honghui Wang, and Yulin Wang. Dynamic neural networks: A survey. *IEEE TPAMI*, 2021. 1, 2, 3
- [13] Yizeng Han, Zhihang Yuan, Yifan Pu, Chenhao Xue, Shiji Song, Guangyu Sun, and Gao Huang. Latency-aware spatial-wise dynamic networks. In *NeurIPS*, 2022. 3
- [14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 2, 6, 7
- [15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2, 6, 7
- [16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, pages 7132–7141, 2018. 4
- [17] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *ICLR*, 2018. 2, 3
- [18] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat’s visual cortex. *The Journal of physiology*, 160(1):106, 1962. 1
- [19] Chenhan Jiang, Hang Xu, Wei Zhang, Xiaodan Liang, and Zhenguo Li. Sp-nas: Serial-to-parallel backbone search for object detection. In *CVPR*, pages 11863–11872, 2020. 2
- [20] Zequn Jie, Peng Sun, Xin Li, Jiashi Feng, and Wei Liu. Any-time recognition with routing convolutional networks. *IEEE TPAMI*, 43(6):1875–1886, 2019. 2, 3
- [21] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: a single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. 1, 2, 6
- [22] Changlin Li, Guangrun Wang, Bing Wang, Xiaodan Liang, Zhihui Li, and Xiaojun Chang. Dynamic slimmable network. In *CVPR*, pages 8607–8617, 2021. 3
- [23] Hao Li, Hong Zhang, Xiaojuan Qi, Ruigang Yang, and Gao Huang. Improved techniques for training adaptive deep networks. In *ICCV*, pages 1891–1900, 2019. 3
- [24] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cbnet: A composite backbone network architecture for object detection. *IEEE TIP*, 2022. 2, 3, 4, 6
- [25] Tingting Liang, Yongtao Wang, Zhi Tang, Guosheng Hu, and Haibin Ling. Opanas: One-shot path aggregation network architecture search for object detection. In *CVPR*, pages 10195–10203, 2021. 1
- [26] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. In *NeurIPS*, 2022. 1
- [27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 2
- [28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 6, 7
- [29] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, pages 21–37. Springer, 2016. 3
- [30] Yudong Liu, Yongtao Wang, Siwei Wang, Tingting Liang, Qijie Zhao, Zhi Tang, and Haibin Ling. Cbnet: A novel composite backbone network architecture for object detection. In *AAAI*, volume 34, pages 11653–11660, 2020. 2, 3, 4
- [31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 2, 6, 7

- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. [6](#)
- [33] Mason McGill and Pietro Perona. Deciding how to decide: Dynamic routing in artificial neural networks. In *International Conference on Machine Learning*, pages 2363–2372. PMLR, 2017. [2](#)
- [34] Akira Murata, Vittorio Gallese, Giuseppe Luppino, Masakazu Kaseda, and Hideo Sakata. Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip. *Journal of neurophysiology*, 83(5):2580–2601, 2000. [1](#)
- [35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. [3](#)
- [36] Linrun Qiu, Dongbo Zhang, Yuan Tian, and Najla Al-Nabhan. Deep learning-based algorithm for vehicle detection in intelligent transportation systems. *The Journal of Supercomputing*, 77(10):11083–11098, 2021. [1](#)
- [37] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *NeurIPS*, 34:13937–13949, 2021. [3](#)
- [38] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. [3](#)
- [39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. [2](#), [6](#), [7](#)
- [40] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pages 10529–10538, 2020. [1](#)
- [41] Zhenhong Sun, Ming Lin, Xiuyu Sun, Zhiyu Tan, Hao Li, and Rong Jin. Mae-det: Revisiting maximum entropy principle in zero-shot nas for efficient object detection. In *International Conference on Machine Learning*, pages 20810–20826. PMLR, 2022. [2](#)
- [42] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *CVPR*, pages 10781–10790, 2020. [2](#)
- [43] Surat Teerapittayanon, Bradley McDanel, and Hsiang-Tsung Kung. Branchynet: Fast inference via early exiting from deep neural networks. In *ICPR*, pages 2464–2469. IEEE, 2016. [2](#), [3](#)
- [44] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Scaled-yolov4: Scaling cross stage partial network. In *CVPR*, pages 13029–13038, 2021. [1](#), [2](#)
- [45] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv preprint arXiv:2207.02696*, 2022. [1](#), [2](#), [6](#)
- [46] Chien-Yao Wang, Hong-Yuan Mark Liao, Yueh-Hua Wu, Ping-Yang Chen, Jun-Wei Hsieh, and I-Hau Yeh. Cspnet: A new backbone that can enhance learning capability of cnn. In *CVPRW*, pages 390–391, 2020. [2](#)
- [47] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 568–578, 2021. [2](#)
- [48] Xiaoxing Wang, Jiale Lin, Juanping Zhao, Xiaokang Yang, and Junchi Yan. Eautodet: efficient architecture search for object detection. In *ECCV*, pages 668–684. Springer, 2022. [1](#), [3](#), [6](#)
- [49] Xin Wang, Fisher Yu, Zi-Yi Dou, Trevor Darrell, and Joseph E Gonzalez. Skipnet: Learning dynamic routing in convolutional networks. In *ECCV*, pages 409–424, 2018. [3](#)
- [50] Yulin Wang, Rui Huang, Shiji Song, Zeyi Huang, and Gao Huang. Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. *NeurIPS*, 34:11960–11973, 2021. [2](#), [3](#)
- [51] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 1492–1500, 2017. [2](#)
- [52] Jiarui Xu, Yue Cao, Zheng Zhang, and Han Hu. Spatial-temporal relation networks for multi-object tracking. In *ICCV*, pages 3988–3998, 2019. [1](#)
- [53] Shangliang Xu, Xinxin Wang, Wenyu Lv, Qinyao Chang, Cheng Cui, Kaipeng Deng, Guanzhong Wang, Qingqing Dang, Shengyu Wei, Yuning Du, et al. Pp-yoloe: An evolved version of yolo. *arXiv preprint arXiv:2203.16250*, 2022. [6](#)
- [54] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *CVPR*, pages 2369–2378, 2020. [2](#), [3](#)
- [55] Shuo Yang, Huimin Lu, and Jianru Li. Multifeature fusion-based object detection for intelligent transportation systems. *IEEE Transactions on Intelligent Transportation Systems*, 2022. [1](#)
- [56] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *CVPR*, pages 11040–11048, 2020. [1](#)
- [57] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *ECCV*, pages 1–21. Springer, 2022. [1](#)
- [58] Hong-Yu Zhou, Bin-Bin Gao, and Jianxin Wu. Adaptive feeding: Achieving fast and accurate detections by adaptively combining object detectors. In *CVPR*, pages 3505–3513, 2017. [2](#), [3](#), [7](#)
- [59] Mingjian Zhu, Kai Han, Changbin Yu, and Yunhe Wang. Dynamic feature pyramid networks for object detection. *arXiv preprint arXiv:2012.00779*, 2020. [4](#)