EfficientDet: Scalable and Efficient Object Detection

Mingxing Tan  Ruoming Pang  Quoc V. Le
Google Research, Brain Team
{tanmingxing, rpang, qvl}@google.com

Abstract

Model efficiency has become increasingly important in computer vision. In this paper, we systematically study various neural network architecture design choices for object detection and propose several key optimizations to improve efficiency. First, we propose a weighted bi-directional feature pyramid network (BiFPN), which allows easy and fast multi-scale feature fusion; Second, we propose a compound scaling method that uniformly scales the resolution, depth, and width for all backbone, feature network, and box/class prediction networks at the same time. Based on these optimizations, we have developed a new family of object detectors, called EfficientDet, which consistently achieve an order-of-magnitude better efficiency than prior art across a wide spectrum of resource constraints. In particular, without bells and whistles, our EfficientDet-D7 achieves state-of-the-art 51.0% mAP on COCO dataset with 52M parameters and 326B FLOPS, being 4x smaller and using 9.3x fewer FLOPS yet still more accurate (+0.3% mAP) than the best previous detector.

1. Introduction

Tremendous progresses have been made in recent years towards more accurate object detection; meanwhile, state-of-the-art object detectors also become increasingly more expensive. For example, the latest AmoebaNet-based NAS-FPN detector [37] requires 167M parameters and 3045B FLOPS (30x more than RetinaNet [17]) to achieve state-of-the-art accuracy. The large model sizes and expensive computation costs deter their deployment in many real-world applications such as robotics and self-driving cars where model size and latency are highly constrained. Given these real-world resource constraints, model efficiency becomes increasingly important for object detection.

There have been many previous works aiming to develop more efficient detector architectures, such as one-stage [20, 25, 26, 17] and anchor-free detectors [14, 36, 32], or compress existing models [21, 22]. Although these methods tend to achieve better efficiency, they usually sacrifice accuracy. Moreover, most previous works only focus on a specific or a small range of resource requirements, but the variety of real-world applications, from mobile devices to datacenters, often demand different resource constraints.

A natural question is: Is it possible to build a scalable detection architecture with both higher accuracy and better efficiency across a wide spectrum of resource constraints (e.g., from 3B to 300B FLOPS)? This paper aims to tackle this problem by systematically studying various design choices of detector architectures. Based on the one-stage detector paradigm, we examine the design choices for backbone, feature fusion, and class/box network, and identify two main challenges:

Challenge 1: efficient multi-scale feature fusion – Since

Figure 1: Model FLOPS vs COCO accuracy – All numbers are for single-model single-scale. Our EfficientDet achieves much better accuracy with fewer computations than other detectors. In particular, EfficientDet-D7 achieves new state-of-the-art 51.0% COCO mAP with 4x fewer parameters and 9.3x fewer FLOPS. Details are in Table 2.
introduced in [16], FPN has been widely used for multi-scale feature fusion. Recently, PANet [19], NAS-FPN [5], and other studies [13, 12, 34] have developed more network structures for cross-scale feature fusion. While fusing different input features, most previous works simply sum them up without distinction; however, since these different input features are at different resolutions, we observe they usually contribute to the fused output feature unequally. To address this issue, we propose a simple yet highly effective weighted bi-directional feature pyramid network (BiFPN), which introduces learnable weights to learn the importance of different input features, while repeatedly applying top-down and bottom-up multi-scale feature fusion.

Challenge 2: model scaling – While previous works mainly rely on bigger backbone networks [17, 27, 26, 5] or larger input image sizes [8, 37] for higher accuracy, we observe that scaling up feature network and box/class prediction network is also critical when taking into account both accuracy and efficiency. Inspired by recent works [31], we propose a compound scaling method for object detectors, which jointly scales up the resolution/depth/width for all backbone, feature network, box/class prediction network.

Finally, we also observe that the recently introduced EfficientNets [31] achieve better efficiency than previous commonly used backbones (e.g., ResNets [9], ResNeXt [33], and AmoebaNet [24]). Combining EfficientNet backbones with our propose BiFPN and compound scaling, we have developed a new family of object detectors, named EfficientDet, which consistently achieve better accuracy with an order-of-magnitude fewer parameters and FLOPS than previous object detectors. Figure 1 and Figure 4 show the performance comparison on COCO dataset [18]. Under similar accuracy constraint, our EfficientDet uses 28x fewer FLOPS than YOLOv3 [26], 30x fewer FLOPS than RetinaNet [17], and 19x fewer FLOPS than the recent NAS-FPN [5]. In particular, with single-model and single test-time scale, our EfficientDet-D7 achieves state-of-the-art 51.0 mAP with 52M parameters and 326B FLOPS, being 4x smaller and using 9.3x fewer FLOPS yet still more accurate (+0.3% mAP) than the best previous models [37]. Our EfficientDet models are also up to 3.2x faster on GPU and 8.1x faster on CPU than previous detectors, as shown in Figure 4 and Table 2.

Our contributions can be summarized as:

- We proposed BiFPN, a weighted bidirectional feature network for easy and fast multi-scale feature fusion.
- We proposed a new compound scaling method, which jointly scales up backbone, feature network, box/class network, and resolution, in a principled way.
- Based on BiFPN and compound scaling, we developed EfficientDet, a new family of detectors with significantly better accuracy and efficiency across a wide spectrum of resource constraints.

2. Related Work

One-Stage Detectors: Existing object detectors are mostly categorized by whether they have a region-of-interest proposal step (two-stage [6, 27, 3, 8]) or not (one-stage [28, 20, 25, 17]). While two-stage detectors tend to be more flexible and more accurate, one-stage detectors are often considered to be simpler and more efficient by leveraging predefined anchors [11]. Recently, one-stage detectors have attracted substantial attention due to their efficiency and simplicity [14, 34, 36]. In this paper, we mainly follow the one-stage detector design, and we show it is possible to achieve both better efficiency and higher accuracy with optimized network architectures.

Multi-Scale Feature Representations: One of the main difficulties in object detection is to effectively represent and process multi-scale features. Earlier detectors often directly perform predictions based on the pyramidal feature hierarchy extracted from backbone networks [2, 20, 28]. As one of the pioneering works, feature pyramid network (FPN) [16] proposes a top-down pathway to combine multi-scale features. Following this idea, PANet [19] adds an extra bottom-up path aggregation network on top of FPN; STDL [35] proposes a scale-transfer module to exploit cross-scale features; M2det [34] proposes a U-shape module to fuse multi-scale features, and G-FRNet [1] introduces gate units for controlling information flow across features. More recently, NAS-FPN [5] leverages neural architecture search to automatically design feature network topology. Although it achieves better performance, NAS-FPN requires thousands of GPU hours during search, and the resulting feature network is irregular and thus difficult to interpret. In this paper, we aim to optimize multi-scale feature fusion with a more intuitive and principled way.

Model Scaling: In order to obtain better accuracy, it is common to scale up a baseline detector by employing bigger backbone networks (e.g., from mobile-size models [30, 10] and ResNet [9], to ResNeXt [33] and AmoebaNet [24]), or increasing input image size (e.g., from 512x512 [17] to 1536x1536 [37]). Some recent works [5, 37] show that increasing the channel size and repeating feature network can also lead to higher accuracy. These scaling methods mostly focus on single or limited scaling dimensions. Recently, [31] demonstrates remarkable model efficiency for image classification by jointly scaling up network width, depth, and resolution. Our proposed compound scaling method for object detection is mostly inspired by [31].

3. BiFPN

In this section, we first formulate the multi-scale feature fusion problem, and then introduce the two main ideas for
3.1. Problem Formulation

Multi-scale feature fusion aims to aggregate features at different resolutions. Formally, given a list of multi-scale features $\tilde{P}^{\text{in}} = (P_{i_1}^{\text{in}}, P_{i_2}^{\text{in}}, \ldots)$, where $P_{i}^{\text{in}}$ represents the feature at level $i$, our goal is to find a transformation $f$ that can effectively aggregate different features and output a list of new features: $\tilde{P}^{\text{out}} = f(\tilde{P}^{\text{in}})$. As a concrete example, Figure 2(a) shows the conventional top-down FPN [16]. It takes level 3-7 input features $\tilde{P}^{\text{in}} = (P_{i_3}^{\text{in}}, \ldots P_{i_7}^{\text{in}})$, where $P_{i}^{\text{in}}$ represents a feature level with resolution of $1/2^i$ of the input images. For instance, if input resolution is 640x640, then $P_{i_3}^{\text{in}}$ represents feature level 3 ($640/2^3 = 80$) with resolution 80x80, while $P_{i_7}^{\text{in}}$ represents feature level 7 with resolution 5x5. The conventional FPN aggregates multi-scale features in a top-down manner:

\[
P_{7}^{\text{out}} = \text{Conv}(P_{7}^{\text{in}})
\]
\[
P_{6}^{\text{out}} = \text{Conv}(P_{6}^{\text{in}} + \text{Resize}(P_{7}^{\text{out}}))
\]
\[
\vdots
\]
\[
P_{3}^{\text{out}} = \text{Conv}(P_{3}^{\text{in}} + \text{Resize}(P_{4}^{\text{out}}))
\]

where $\text{Resize}$ is usually a upsampling or downsampling op for resolution matching, and $\text{Conv}$ is usually a convolutional op for feature processing.

3.2. Cross-Scale Connections

Conventional top-down FPN is inherently limited by the one-way information flow. To address this issue, PANet [19] adds an extra bottom-up path aggregation network, as shown in Figure 2(b). Cross-scale connections are further studied in [13, 12, 34]. Recently, NAS-FPN [5] employs neural architecture search to search for better cross-scale feature network topology, but it requires thousands of GPU
hours during search and the found network is irregular and difficult to interpret or modify, as shown in Figure 2(c).

By studying the performance and efficiency of these three networks (Table 4), we observe that PANet achieves better accuracy than FPN and NAS-FPN, but with the cost of more parameters and computations. To improve model efficiency, this paper proposes several optimizations for cross-scale connections: First, we remove those nodes that only have one input edge. Our intuition is simple: if a node has only one input edge with no feature fusion, then it will have less contribution to feature network that aims at fusing different features. This leads to a simplified PANet as shown in Figure 2(e); Second, we add an extra edge from the original input to output node if they are at the same level, in order to fuse more features without adding much cost, as shown in Figure 2(f); Third, unlike PANet [19] that only has one top-down and one bottom-up path, we treat each bidirectional (top-down & bottom-up) path as one feature network layer, and repeat the same layer multiple times to enable more high-level feature fusion. Section 4.2 will discuss how to determine the number of layers for different resource constraints using a compound scaling method. With these optimizations, we name the new feature network as bidirectional feature pyramid network (BiFPN), as shown in Figure 2(f) and 3.

3.3. Weighted Feature Fusion

When fusing multiple input features with different resolutions, a common way is to first resize them to the same resolution and then sum them up. Pyramid attention network [15] introduces global self-attention upsampling to recover pixel localization, which is further studied in [5].

Previous feature fusion methods treat all input features equally without distinction. However, we observe that since different input features are at different resolutions, they usually contribute to the output feature unequally. To address this issue, we propose to add an additional weight for each input during feature fusion, and let the network to learn the importance of each input feature. Based on this idea, we consider three weighted fusion approaches:

**Unbounded fusion:** \( O = \sum_i w_i \cdot I_i \), where \( w_i \) is a learnable weight that can be a scalar (per-feature), a vector (per-channel), or a multi-dimensional tensor (per-pixel). We find a scale can achieve comparable accuracy to other approaches with minimal computational costs. However, since the scalar weight is unbounded, it could potentially cause training instability. Therefore, we resort to weight normalization to bound the value range of each weight.

**Softmax-based fusion:** \( O = \sum_i \frac{w_i}{\sum_j w_j} \cdot I_i \). An intuitive idea is to apply softmax to each weight, such that all weights are normalized to be a probability with value range from 0 to 1, representing the importance of each input. However, as shown in our ablation study in section 6.3, the extra softmax leads to significant slowdown on GPU hardware. To minimize the extra latency cost, we further propose a fast fusion approach.

**Fast normalized fusion:** \( O = \sum_i \frac{w_i}{\sum_j w_j} \cdot I_i \), where \( w_i \geq 0 \) is ensured by applying a Relu after each \( w_i \), and \( \epsilon = 0.0001 \) is a small value to avoid numerical instability. Similarly, the value of each normalized weight also falls between 0 and 1, but since there is no softmax operation here, it is much more efficient. Our ablation study shows this fast fusion approach has very similar learning behavior and accuracy as the softmax-based fusion, but runs up to 30% faster on GPUs (Table 5).

Our final BiFPN integrates both the bidirectional cross-scale connections and the fast normalized fusion. As a concrete example, here we describe the two fused features at level 6 for BiFPN shown in Figure 2(f):

\[
P_{6}^{td} = \text{Conv} \left( \frac{w_1 \cdot P_{6}^{in} + w_2 \cdot \text{Resize}(P_{7}^{in})}{w_1 + w_2 + \epsilon} \right)
\]

\[
P_{6}^{out} = \text{Conv} \left( \frac{w_1' \cdot P_{6}^{in} + w_2' \cdot P_{6}^{td} + w_3' \cdot \text{Resize}(P_{5}^{out})}{w_1' + w_2' + w_3' + \epsilon} \right)
\]

where \( P_{6}^{td} \) is the intermediate feature at level 6 on the top-down pathway, and \( P_{6}^{out} \) is the output feature at level 6 on the bottom-up pathway. All other features are constructed in a similar manner. Notably, to further improve the efficiency, we use depthwise separable convolution [4, 29] for feature fusion, and add batch normalization and activation after each convolution.

4. EfficientDet

Based on our BiFPN, we have developed a new family of detection models named EfficientDet. In this section, we will discuss the network architecture and a new compound scaling method for EfficientDet.

4.1. EfficientDet Architecture

Figure 3 shows the overall architecture of EfficientDet, which largely follows the one-stage detectors paradigm [20, 25, 16, 17]. We employ ImageNet-pretrained EfficientNets as the backbone network. Our proposed BiFPN serves as the feature network, which takes level 3-7 features \( \{P_{3}, P_{4}, P_{5}, P_{6}, P_{7}\} \) from the backbone network and repeatedly applies top-down and bottom-up bidirectional feature fusion. These fused features are fed to a class and box network to produce object class and bounding box predictions respectively. Similar to [17], the class and box network weights are shared across all levels of features.
Figure 3: EfficientDet architecture – It employs EfficientNet [31] as the backbone network, BiFPN as the feature network, and shared class/box prediction network. Both BiFPN layers and class/box net layers are repeated multiple times based on different resource constraints as shown in Table 1.

4.2. Compound Scaling

Aiming at optimizing both accuracy and efficiency, we would like to develop a family of models that can meet a wide spectrum of resource constraints. A key challenge here is how to scale up a baseline EfficientDet model.

Previous works mostly scale up a baseline detector by employing bigger backbone networks (e.g., ResNeXt [33] or AmoebaNet [24]), using larger input images, or stacking more FPN layers [5]. These methods are usually ineffective since they only focus on a single or limited scaling dimensions. Recent work [31] shows remarkable performance on image classification by jointly scaling up all dimensions of network width, depth, and input resolution. Inspired by these works [5, 31], we propose a new compound scaling method for object detection, which uses a simple compound coefficient $\phi$ to jointly scale up all dimensions of backbone network, BiFPN network, class/box network, and resolution. Unlike [31], object detectors have much more scaling dimensions than image classification models, so grid search for all dimensions is prohibitive expensive. Therefore, we use a heuristic-based scaling approach, but still follow the main idea of jointly scaling up all dimensions.

**Backbone network** – we reuse the same width/depth scaling coefficients of EfficientNet-B0 to B6 [31] such that we can easily reuse their ImageNet-pretrained checkpoints.

**BiFPN network** – we exponentially grow BiFPN width $W_{bifpn}$ (#channels) as similar to [31], but linearly increase depth $D_{bifpn}$ (#layers) since depth needs to be rounded to small integers. Formally, we use the following equation:

$$W_{bifpn} = 64 \cdot (1.35^\phi), \quad D_{bifpn} = 2 + \phi \quad (1)$$

| Input size $R_{input}$ | Backbone Network $R_{bifpn}$ | BiFPN $W_{bifpn}$ | #layers $D_{bifpn}$ | Box/class $D_{class}$ |
|------------------------|-------------------------------|-------------------|-----------------|---------------------|
| D0 ($\phi = 0$)        | 512 B0                        | 64                | 2               | 3                   |
| D1 ($\phi = 1$)        | 640 B1                        | 88                | 3               | 3                   |
| D2 ($\phi = 2$)        | 768 B2                        | 112               | 4               | 3                   |
| D3 ($\phi = 3$)        | 896 B3                        | 160               | 5               | 4                   |
| D4 ($\phi = 4$)        | 1024 B4                       | 224               | 6               | 4                   |
| D5 ($\phi = 5$)        | 1280 B5                       | 288               | 7               | 4                   |
| D6 ($\phi = 6$)        | 1408 B6                       | 384               | 8               | 5                   |
| D7                     | 1536 B6                       | 384               | 8               | 5                   |

Table 1: Scaling configs for EfficientDet D0-D7 – $\phi$ is the compound coefficient that controls all other scaling dimensions; BiFPN, box/class net, and input size are scaled up using equation 1, 2, 3 respectively. D7 has the same settings as D6 except using larger input size.

**Box/class prediction network** – we fix their width to be always the same as BiFPN (i.e., $W_{Pred} = W_{bifpn}$), but linearly increase the depth (#layers) using equation:

$$D_{box} = D_{class} = 3 + \lfloor \phi/3 \rfloor \quad (2)$$

**Input image resolution** – Since feature level 3-7 are used in BiFPN, the input resolution must be divisible by $2^7 = 128$, so we linearly increase resolutions using equation:

$$R_{input} = 512 + \phi \cdot 128 \quad (3)$$

Following Equations 1,2,3 with different $\phi$, we have developed EfficientDet-D0 ($\phi = 0$) to D6 ($\phi = 6$) as shown in Table 1. Notably, models scaled up with $\phi \geq 7$ could not fit memory unless changing batch size or other settings; therefore, we simply expand D6 to D7 by only enlarging input size while keeping all other dimensions the same, such that we can use the same training settings for all models.
| Model                      | mAP   | #Params | Ratio | #FLOPS | Ratio | GPU LAT(ms) | Speedup | CPU LAT(s) | Speedup |
|----------------------------|-------|---------|-------|--------|-------|-------------|---------|------------|---------|
| EfficientDet-D0            | 32.4  | 3.9M    | 1x    | 2.5B   | 1x    | 16 ±1.6     | 1x      | 0.32 ±0.002 | 1x      |
| YOLOv3 [26]                | 33.0  | -       | -     | -      | -     | 511         | -       | -          | -       |
| EfficientDet-D1            | 38.3  | 6.6M    | 1x    | 6B     | 1x    | 20 ±1.1     | 1x      | 0.74 ±0.003 | 1x      |
| MaskRCNN [8]               | 37.9  | 4.4M    | 6.7x  | 149B   | 25x   | 921         | -       | -          | -       |
| RetinaNet-R50 (640) [17]   | 37.0  | 34.0M   | 6.7x  | 97B    | 16x   | 27 ±1.1     | 1.4x    | 2.8 ±0.007 | 3.8x    |
| RetinaNet-R101 (640) [17]  | 37.9  | 53.0M   | 8x    | 127B   | 21x   | 34 ±0.5     | 1.7x    | 3.6 ±0.002 | 4.9x    |
| EfficientDet-D2            | 41.1  | 8.1M    | 1x    | 11B    | 1x    | 24 ±0.5     | 1x      | 1.2 ±0.003 | 1x      |
| RetinaNet-R50 (1024) [17]  | 40.1  | 34.0M   | 4.3x  | 248B   | 23x   | 51 ±0.9     | 2.0x    | 7.5 ±0.006 | 6.3x    |
| RetinaNet-R101 (1024) [17] | 41.1  | 53.0M   | 6.6x  | 326B   | 30x   | 65 ±0.4     | 2.7x    | 9.7 ±0.008 | 8.1x    |
| NAS-FPN R-50 (640) [5]     | 39.9  | 60.3M   | 7.5x  | 141B   | 13x   | 41 ±0.6     | 1.7x    | 4.1 ±0.027 | 3.4x    |
| EfficientDet-D3            | 44.3  | 12.0M   | 1x    | 25B    | 1x    | 42 ±0.8     | 1x      | 2.5 ±0.002 | 1x      |
| NAS-FPN R-50 (1024) [5]    | 44.2  | 60.3M   | 5.1x  | 360B   | 15x   | 79 ±0.3     | 1.9x    | 11 ±0.003  | 4.6x    |
| NAS-FPN R-50 (1280) [5]    | 44.8  | 60.3M   | 5.1x  | 563B   | 25x   | 119 ±0.9    | 2.8x    | 17 ±0.150  | 6.8x    |
| EfficientDet-D4            | 46.6  | 20.7M   | 1x    | 55B    | 1x    | 74 ±0.5     | 1x      | 4.8 ±0.003 | 1x      |
| EfficientDet-D5 + AA       | 49.8  | 33.7M   | 1x    | 136B   | 1x    | 141 ±2.1    | 1x      | 11 ±0.002  | 1x      |
| AmoebaNet+ NAS-FPN + AA(1280) [37] | 48.6  | 185M    | 5.5x  | 1317B  | 9.7x  | 259 ±1.2    | 1.8x    | 38 ±0.004  | 3.5x    |
| EfficientDet-D6 + AA       | 50.6  | 51.9M   | 1x    | 227B   | 1x    | 190 ±1.1    | 1x      | 16 ±0.003  | 1x      |
| AmoebaNet+ NAS-FPN + AA(1536) [37] | 50.7  | 209M    | 4.0x  | 3045B  | 13x   | 608 ±1.4    | 3.2x    | 83 ±0.092  | 5.2x    |
| EfficientDet-D7 + AA       | 51.0  | 51.9M   | 1x    | 326B   | 1x    | 262 ±2.2    | 1x      | 24 ±0.003  | 1x      |

1 We omit ensemble and test-time multi-scale results [23, 7].
2 Latency numbers marked with 1 are from papers, and all others are measured on the same machine.

Table 2: EfficientDet performance on COCO [18] – Results are for single-model single-scale. #Params and #FLOPS denote the number of parameters and multiply-adds. LAT denotes inference latency with batch size 1. AA denotes auto-augmentation [37]. We group models together if they have similar accuracy, and compare the ratio or speedup between EfficientDet and other detectors in each group.

5. Experiments

We evaluate EfficientDet on COCO 2017 detection datasets [18]. Each model is trained using SGD optimizer with momentum 0.9 and weight decay 4e-5. Learning rate is first linearly increased from 0 to 0.08 in the initial 5% warm-up training steps and then annealed down using cosine decay rule. Batch normalization is added after every convolution with batch norm decay 0.997 and epsilon 1e-4. We use exponential moving average with decay 0.9998. We also employ commonly-used focal loss [17] with $\alpha = 0.25$ and $\gamma = 1.5$, and aspect ratio \{1/2, 1, 2\}. Our models are trained with batch size 128 on 32 TPUv3 chips. We use RetinaNet [17] preprocessing for EfficientDet-D0/D1/D3/D4, but for fair comparison, we use the same auto-augmentation for EfficientDet-D5/D6/D7 when comparing with the prior art of AmoebaNet-based NAS-FPN detectors [37].

Table 2 compares EfficientDet with other object detectors, under the single-model single-scale settings with no test-time augmentation. Our EfficientDet achieves better accuracy and efficiency than previous detectors across a wide range of accuracy or resource constraints. On relatively low-accuracy regime, our EfficientDet-D0 achieves similar accuracy as YOLOv3 with 28x fewer FLOPS. Compared to RetinaNet [17] and Mask-RCNN [8], our EfficientDet-D1 achieves similar accuracy with up to 8x fewer parameters and 25x fewer FLOPS. On high-accuracy regime, our EfficientDet also consistently outperforms recent NAS-FPN [5] and its enhanced versions in [37] with an order-of-magnitude fewer parameters and FLOPS. In particular, our EfficientDet-D7 achieves a new state-of-the-art 51.0 mAP for single-model single-scale, while still being 4x smaller and using 9.3x fewer FLOPS than previous best results in [37]. Notably, unlike the large AmoebaNet-based NAS-FPN models [37] that require special settings (e.g., change anchors from 3x3 to 9x9 and train with model parallelism), all EfficientDet models use the same 3x3 anchors and trained without model parallelism.

In addition to parameter size and FLOPS, we have also compared the real-world latency on Titan-V GPU and single-thread Xeon CPU. We run each model 10 times with batch size 1 and report the mean and standard deviation. Figure 4 illustrates the comparison on model size, GPU latency, and single-thread CPU latency. For fair comparison, these figures only include results that are measured on the same machine. Compared to previous detectors, EfficientDet models are up to 3.2x faster on GPU and 8.1x faster on CPU, suggesting they are efficient on real-world hardware.
6. Ablation Study

In this section, we ablate various design choices for our proposed EfficientDet.

6.1. Disentangling Backbone and BiFPN

Since EfficientDet uses both a powerful backbone and a new BiFPN, we want to understand how much each of them contributes to the accuracy and efficiency improvements. Table 3 compares the impact of backbone and BiFPN. Starting from a RetinaNet detector [17] with ResNet-50 [9] backbone and top-down FPN [16], we first replace the backbone with EfficientNet-B3, which improves accuracy by about 3 mAP with slightly less parameters and FLOPS. By further replacing FPN with our proposed BiFPN, we achieve additional 4 mAP gain with much fewer parameters and FLOPS. These results suggest that EfficientNet backbones and BiFPN are both crucial for our final models.

| Model       | mAP | Parameters (M) | FLOPS |
|-------------|-----|----------------|-------|
| ResNet50 + FPN | 37.0 | 34M            | 97B   |
| EfficientNet-B3 + FPN | 40.3 | 21M            | 75B   |
| EfficientNet-B3 + BiFPN | 44.4 | 12M            | 24B   |

Table 3: Disentangling backbone and BiFPN – Starting from the standard RetinaNet (ResNet50+FPN), we first replace the backbone with EfficientNet-B3, and then replace the baseline FPN with our proposed BiFPN.

6.2. BiFPN Cross-Scale Connections

Table 4 shows the accuracy and model complexity for feature networks with different cross-scale connections listed in Figure 2. Notably, the original FPN [16] and PANet [19] only have one top-down or bottom-up flow, but for fair comparison, here we repeat each of them 5 times (same as BiFPN). We use the same backbone and class/box prediction network, and the same training settings for all experiments. As we can see, the conventional top-down FPN is inherently limited by the one-way information flow and thus has the lowest accuracy. While repeated PANet achieves slightly better accuracy than NAS-FPN [5], it also requires more parameters and FLOPS. Our BiFPN achieves similar accuracy as repeated PANet, but uses much less parameters and FLOPS. With the additional weighted feature fusion, our BiFPN further achieves the best accuracy with fewer parameters and FLOPS.

| Feature Network | mAP | #Params Ratio | #FLOPS Ratio |
|----------------|-----|---------------|--------------|
| Top-Down FPN   | 42.29 | 1.0x          | 1.0x         |
| Repeated PANet | 44.08 | 1.0x          | 1.0x         |
| NAS-FPN [5]    | 43.16 | 0.71x         | 0.72x        |
| Fully-Connected FPN | 43.06 | 1.24x         | 1.21x        |
| BiFPN (w/o weighted) | 43.94 | 0.88x         | 0.67x        |
| BiFPN (w/ weighted) | 44.39 | 0.88x         | 0.68x        |

Table 4: Comparison of different feature networks – Our weighted BiFPN achieves the best accuracy with fewer parameters and FLOPS.

6.3. Softmax vs Fast Normalized Fusion

As discussed in Section 3.3, we propose a fast normalized feature fusion approach to get rid of the expensive softmax while retaining the benefits of normalized weights. Table 5 compares the softmax and fast normalized fusion approaches in three detectors with different model sizes. As shown in the results, our fast normalized fusion approach achieves similar accuracy as the softmax-based fusion, but runs 1.26x - 1.31x faster on GPUs.
In order to further understand the behavior of softmax-based and fast normalized fusion, Figure 5 illustrates the learned weights for three feature fusion nodes randomly selected from the BiFPN layers in EfficientDet-D3. Notably, the normalized weights (e.g., $e_i / \sum_j e_j$ for softmax-based fusion, and $w_i / (\epsilon + \sum_j w_j)$ for fast normalized fusion) always sum up to 1 for all inputs. Interestingly, the normalized weights change rapidly during training, suggesting different features contribute to the feature fusion unequally. Despite the rapid weight change, our fast normalized fusion approach always shows very similar learning behavior to the softmax-based fusion during training for all three nodes.

### 6.4. Compound Scaling

As discussed in section 4.2, we employ a compound scaling method to jointly scale up all dimensions of depth/width/resolution for backbone, BiFPN, and box/class prediction networks. Figure 6 compares our compound scaling with other alternative methods that scale up a single dimension of resolution/depth/width. Although starting from the same baseline detector, our compound scaling method achieves better efficiency than other methods, suggesting the benefits of jointly scaling by better balancing different architecture dimensions.

### 7. Conclusion

In this paper, we systematically study various network architecture design choices for efficient object detection, and propose a weighted bidirectional feature network and a customized compound scaling method, in order to improve accuracy and efficiency. Based on these optimizations, we have developed a new family of detectors, named EfficientDet, which consistently achieve better accuracy and efficiency than the prior art across a wide spectrum of resource constraints. In particular, our EfficientDet-D7 achieves state-of-the-art accuracy with an order-of-magnitude fewer parameters and FLOPS than the best existing detector. Our EfficientDet is also up to 3.2x faster on GPUs and 8.1x faster on CPUs. Code will be made public.
