Learning Efficient Binarized Object Detectors with Information Compression

Ziwei Wang, *Student Member, IEEE,* Jiwen Lu, *Senior Member, IEEE,* Ziyi Wu, *Student Member, IEEE,* and Jie Zhou, *Senior Member, IEEE*

Abstract—In this paper, we propose a binarized detection learning method (BiDet) for efficient object detection. Conventional network binarization methods directly quantize the weights and activations in one-stage or two-stage detectors with constrained representational capacity, so that the information redundancy in the networks causes numerous false positives and degrades the performance significantly. On the contrary, our BiDet fully utilizes the representational capacity of the binary neural networks by redundancy removal, through which the detection precision is enhanced with alleviated false positives. Specifically, we generalize the information bottleneck (IB) principle to object detection, where the amount of information in the high-level feature maps is constrained and the mutual information between the feature maps and object detection is maximized. Meanwhile, we learn sparse object priors so that the posteriors are concentrated on informative detection prediction with false positive elimination. Since BiDet employs a fixed IB trade-off to balance the total and relative information contained in the high-level feature maps, the information compression leads to ineffective utilization of the network capacity or insufficient redundancy removal for input in different complexity. To address this, we further present binary neural networks with automatic information compression (AutoBiDet) to automatically adjust the IB trade-off for each input according to the complexity. Moreover, we further propose the class-aware sparse object priors by assigning different sparsity to objects in various classes, so that the false positives are alleviated more effectively without recall decrease. Extensive experiments on the PASCAL VOC and COCO datasets show that our BiDet and AutoBiDet outperform the state-of-the-art binarized object detectors by a sizable margin.

Index Terms—Binary neural networks, object detection, information bottleneck, automatic information compression, sparse priors

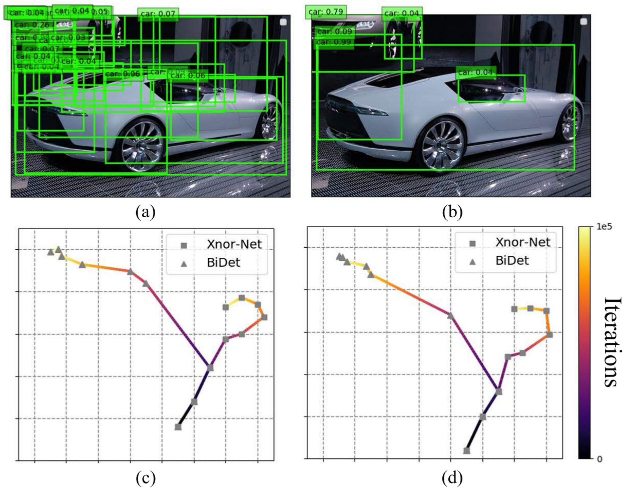
1 Introduction

Convolutional neural network (CNN) based object detectors [8], [13], [43], [31], [29], [65] have achieved state-of-the-art performance due to its strong discriminative power and generalization ability. However, the CNN based detection methods usually require massive computation and storage resources to achieve ideal performance, which limits their deployment on mobile devices. Therefore, it is desirable to develop object detectors with lightweight architectures and few parameters.

To reduce the complexity of deep neural networks, several deep model compression methods have been proposed including pruning [36], [63], [15], low-rank decomposition [26], [39], [21], quantization [54], [24], [10], knowledge distillation [56], [52], [4], efficient architecture design [45], [62], [40] and efficient architecture search [57], [48], [30]. Among these methods, network quantization reduces the bitwidth of the network parameters and activations for efficient inference. In the extreme cases, binarizing weights and activations of neural networks decreases the storage and computational cost by $32\times$ and $64\times$ respectively. However, directly deploying binary neural networks with constrained representational capacity in object detection causes numerous false positives due to the information redundancy in the networks.

In this paper, we present a binarized detector learning method (BiDet) that quantizes both the backbone part and the detection part for efficient object detection. Unlike existing methods which directly binarize the weights and activations in one-stage or two-stage detectors, our method fully utilizes the representational capacity of the binary neural networks for object detection via redundancy removal. Consequently, the detection precision is enhanced with false positive elimination. More specifically, we impose the information bottleneck (IB) principle on binarized object detector learning, where we simultaneously limit the amount of information in high-level feature maps and maximize the mutual information between object detection and the learned feature maps. Meanwhile, we employ sparse object priors (SOP) in IB, so that the posteriors are enforced to be concentrated on informative object prediction and the uninformative false positives are eliminated. Figure 1 (a) and (b) show an example of predicted positives obtained by Xnor-Net [41] and our BiDet respectively, where the false positives are significantly reduced in the latter. Figure 1 (c) and (d) depict the information plane dynamics for the training and test sets respectively. The horizontal axis means the mutual information between the high-level feature maps and input, and the vertical axis represents the mutual information between the object and the feature maps. Our BiDet enhances the precision by redundancy removal and full utilization of network capacity.

- *Ziwei Wang, Jiwen Lu and Ziyi Wu are with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100084, China. E-mail: wang-zw18@mails.tsinghua.edu.cn, lujiwen@tsinghua.edu.cn, wuzy17@mails.tsinghua.edu.cn.*
- *Jie Zhou is with the Beijing National Research Center for Information Science and Technology (BNRist), Department of Automation, Tsinghua University, Beijing 100084, China, and also with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. E-mail: jzhou@tsinghua.edu.cn*
- *Part of this work was presented in [55].*
- *Code: https://github.com/ZiweiWangTHU/BiDet.git.*



(a) (b)

(c) (d)

Figure 1. An example of the predicted objects with the binarized SSD detector on PASCAL VOC. (a) and (b) demonstrate the detection results via Xnor-Net and the proposed BiDet respectively, where the false positives are significantly reduced in our method. (c) and (d) reveal the information plane dynamics for the training and test sets respectively. The horizontal axis means the mutual information between the high-level feature map and input, and the vertical axis represents the mutual information between object detection and the feature map. Compared with Xnor-Net, our method removes the redundant information and fully utilizes the network capacity to achieve higher performance. (best viewed in color).

In fact, the optimal IB trade-off between the total and relative information of the high-level feature maps varies for different input samples. While BiDet applies a fixed IB trade-off for networks with constant capacity, the information compression leads to ineffective utilization of network capacity for input in low complexity and results in incomplete redundancy removal for input in high complexity. In order to address these limitations, we further present binary neural networks with automatic information compression (AutoBiDet) to automatically adjust the trade-off for each input according to the input complexity. Specifically, we employ the generative adversarial network (GAN) [11] based Complexity EStimator (CES) to evaluate the input complexity. The generator aims to reconstruct the input based on the high-level feature maps learned by the backbone. High discriminator probability that the reconstructed sample is real indicates low input complexity, so that the network capacity is sufficient to extract the image information for reconstruction. As a result, less compression should be adopted to fully utilize the network capacity. On the contrary, low discriminator probability requires more compression for the binarized object detectors in order to remove redundancy thoroughly. We train the binarized object detector with the dynamic IB trade-off that is adjusted according to the input complexity. Meanwhile, the sparse object priors (SOP) in BiDet equally limit the predicted positives for all classes. Because false positives are more likely to emerge in classes with more predicted positives, SOP results in degraded recall for classes with few predicted positives and leads to incomplete false positive elimination for classes with many predicted positives. We propose class-aware sparse object priors (C-SOP) that assign different sparsity to objects in various classes, so that the false positives are alleviated more effectively without recall decrease for all classes. Extensive experiments on the PASCAL VOC [7] and COCO [27] datasets show that our BiDet and AutoBiDet outperform the state-of-the-art binary neural networks in object detection across various architectures and detection frameworks. Moreover, the presented techniques in BiDet and AutoBiDet can be integrated to other compression methods including quantization, pruning and efficient architecture design to enhance the vanilla models.

This paper is an extended version of our conference paper [55], where we make the following new contributions:

1) We further propose a new AutoBiDet method based on BiDet in the conference version by automatically adjusting the IB trade-off with the input complexity, so that the network capacity is fully utilized and the redundancy is completely removed for all input.
2) We present class-aware sparse object priors (C-SOP) to assign different sparsity to objects in various classes, so that false positives are alleviated more effectively without recall decrease.
3) We conduct extensive experiments on PASCAL VOC and COCO to evaluate the proposed BiDet and AutoBiDet, and the results show the effectiveness and the efficiency of the presented methods. Moreover, we integrate the proposed techniques to other compression methods including quantization, pruning and efficient architecture design to enhance the vanilla models.

2 Related Work

In this section, we briefly review three related topics: 1) network quantization, 2) object detection and 3) information bottleneck.

Network Quantization: Network quantization has been widely studied in recent years due to its efficiency in storage and computation. Existing methods can be divided into two categories: neural networks with weights and activations in one bit or multiple bits. Binary neural networks reduce the model complexity significantly due to the extremely high compression ratio. Hubara *et al.* [19] and Rastegari *et al.* [41] binarized both weights and activations in neural networks and replaced the multiply-accumulation with xnor and bitcount operations, where straight-through estimators were applied to relax the non-differentiable sign function for back-propagation. Liu *et al.* [32] added extra shortcut between consecutive convolutional blocks to strengthen the representational capacity of the network. They also used custom gradients to optimize the non-differentiable networks. Yang *et al.* [60] leveraged the soft quantization strategy by approximating the rigid sign function with the sigmoid layer, where the discrepancy between the optimization objective and the gradient was minimized. Because binary neural networks perform poorly on difficult tasks such as object detection due to the low representational capacity, multi-bit quantization strategies have been proposed with wider bitwidth. Jacob *et al.* [20] presented an 8-bit quantized model for inference in object detection and their method could be integrated with efficient architectures. Wei *et al.* [56] applied the knowledge distillation to learn 8-bit neural networks in small size from large full-precision models. Li *et al.* [24] proposed fully quantized neural networks in four bits with hardware-friendly implementation.

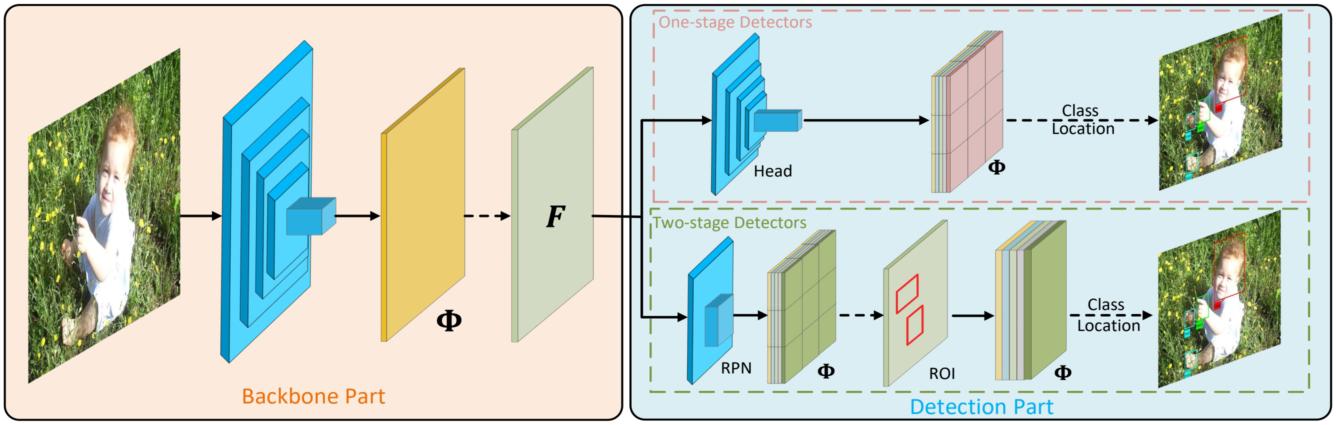


Figure 2. The pipeline of the information bottleneck based detectors, which consists of the backbone part and the detection part. The solid line represents the forward propagation in the network, while the dashed line means sampling from a parameterized distribution Φ. The high-level feature map F is sampled from the distribution parameterized by the backbone network. The one-stage and two-stage detector framework can be both employed in the detection part of our BiDet. For one-stage detectors, the head network parameterizes the posterior distribution of object classes and location. For two-stage detectors, Region Proposal Networks (RPN) parameterize the prior distribution of location and the posteriors are parameterized by the refining networks. (best viewed in color).

Meanwhile, the instabilities during training were overcome by the presented techniques. Nevertheless, multi-bit neural networks still suffer from heavy storage and computation cost. Directly applying binary neural networks with constrained representational power in object detection leads to numerous false positives and significantly degrades the performance due to the redundancy in the networks.

Object Detection: Object detection has aroused comprehensive interest in computer vision due to its wide application. Modern CNN based detectors are categorized into two-stage and one-stage detectors. In the former, R-CNN [9] was among the earliest CNN-based detectors with the pipeline of bounding box regression and classification. Progressive improvements were proposed for better efficiency and effectiveness. Fast R-CNN [8] presented the ROIpooling in the detection framework to achieve better accuracy and faster inference. Faster R-CNN [43] proposed the Region Proposal Networks to effectively generate region proposals instead of hand-crafted ones. FPN [28] introduced top-down architectures with lateral connections and the multi-scale features to integrate low-level and high-level features. In the latter regard, SSD [31] and YOLO [42] directly predicted the bounding box and the class without region proposal generation, so that real-time inference was achieved on GPUs with competitive accuracy. RetinaNet [29] proposed the focal loss to solve the problem of foreground-background class imbalance. However, CNN based detectors suffer from heavy storage and computational cost so that their deployment is limited. To address this, some efficient architectures have been designed for object detection. Light-head R-CNN [25] employed thin feature maps for ROI warping and R-CNN subnet to reduce the computational complexity. Pelee [53] utilized the variant of DenseNet [18] to improve the network connectivity and applied different scales of feature maps to extract information in various resolution. ThunderNet [40] added the context enhancement and spatial attention module to enhance the network capability, and modified the backbone architectures by changing the filter size to retain information. Nevertheless, these methods still require large memory, which is prohibited in portable devices.

Information Bottleneck: The information bottleneck (IB) principle was first proposed by [49] with the goal of extracting relevant information of the input regarding the task, and was widely applied in compression. The IB principle enforces the mutual information between the input and learned features to be minimized while simultaneously maximizing the mutual information between the features and groundtruth of the tasks. Louizos *et al.* [33] and Ullrich *et al.* [51] utilized the Minimal Description Length (MDL) principle [16] that was equivalent to IB to stochastically quantize deep neural networks. Moreover, they used the sparse horseshoe and Gaussian mixture priors for weight learning in order to reduce the quantization errors. Dai *et al.* [6] pruned individual neurons via variational IB so that redundancy between adjacent layers was minimized by aggregating useful information in a subset of neurons. Despite the network compression, IB is also utilized in compact feature learning. Amjad *et al.* [1] proposed stochastic deep neural networks where IB could be utilized to learn efficient representations for classification. Shen *et al.* [46] imposed IB on existing hash models to generate effective binary representations so that the data semantics were fully utilized. To quantitatively analyze the dynamics on the information plane, Wu *et al.* [59], [58] formulated the learnability of different IB trade-off by the second-order calculus of variations. In this paper, we extend the IB principle to squeeze the redundancy in binary detection networks, so that the false positives are alleviated and the detection precision is significantly enhanced.

3 Approach

We first present the efficient binarized object detectors called BiDet, and then detail the binary detection networks with automatic information compression named AutoBiDet.

3.1 BiDet

In this section, we first extend the IB principle to object detection for information redundancy removal. Then we present the details of learning sparse object priors for object detection, which concentrate posteriors on informative prediction with false positive elimination. Finally, we propose the efficient binarized object detectors.

3.1.1 Information Bottleneck for Object Detection

The information bottleneck (IB) principle directly relates to compression with the best hypothesis that the data misfit and the model complexity should simultaneously be minimized. The redundant information irrelevant to the task is exclusive in the compressed model and the capacity of the lightweight model is fully utilized. The task of object detection can be regarded as a Markov process with the following Markov chain:

$$X \rightarrow F \rightarrow L, C \tag{1}$$

where X means the input images and F stands for the high-level feature maps output by the backbone part. C and L represent the classes and location of the predicted objects respectively. According to the Markov chain, the objective of the IB principle is written as follows:

$$\min_{\phi_b, \phi_d} \quad I(X;F) - \beta I(F;C,L) \tag{2}$$

where ϕ_b and ϕ_d are the parameters of the backbone and the detection part respectively. $I(X;Y)$ means the mutual information between two random variables X and Y. Minimizing the mutual information between the images and the high-level feature maps constrains the amount of information that the detector extracts, and maximizing the mutual information between the high-level feature maps and object detection enforces the detector to preserve more information related to the task. As a result, the redundant information irrelevant to object detection is removed. Figure 2 shows the pipeline for information bottleneck based detectors, the IB principle can be imposed on the conventional one-stage and two-stage detectors. We rewrite the first term of (2) according to the definition of mutual information:

$$I(X;F) = \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \mathbb{E}_{\boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{x})} \log \frac{p(\boldsymbol{f}|\boldsymbol{x})}{p(\boldsymbol{f})} \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{f}$ are the input images and the corresponding high-level feature maps. $p(\boldsymbol{x})$ and $p(\boldsymbol{f})$ are the prior distribution of $\boldsymbol{x}$ and $\boldsymbol{f}$ respectively, and $\mathbb{E}$ represents the expectation. $p(\boldsymbol{f}|\boldsymbol{x})$ is the posterior distribution of the high-level feature map conditioned on the input. We parameterize $p(\boldsymbol{f}|\boldsymbol{x})$ by the backbone due to its intractability, where evidence-lower-bound (ELBO) minimization is applied for relaxation. To estimate $I(X;F)$, we sample the training set to obtain the image $\boldsymbol{x}$ and sample the distribution parameterized by the backbone to acquire the corresponding high-level feature map $\boldsymbol{f}$.

The location and classification of objects based on the high-level feature maps are predicted independently, so that the mutual information in the second term of (2) is factorized:

$$I(F;C,L) = I(F;C) + I(F;L) \tag{4}$$

Similar to (3), we rewrite the mutual information between the high-level feature maps and the classes as follows:

$$I(F;C) = \mathbb{E}_{\boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{x})} \mathbb{E}_{\boldsymbol{c} \sim p(\boldsymbol{c}|\boldsymbol{f})} \log \frac{p(\boldsymbol{c}|\boldsymbol{f})}{p(\boldsymbol{c})} \tag{5}$$

where $\boldsymbol{c}$ is the object class labels. $p(\boldsymbol{c})$ and $p(\boldsymbol{c}|\boldsymbol{f})$ are the prior class distribution and posterior class distribution when given the feature maps respectively. Same as the calculation of (3), we employ the classification networks in the detection part to parameterize the posterior distribution. Meanwhile, we divide the images into blocks for multiple object detection. For one-stage detectors such as SSD [31], we project the cells of the high-level feature maps to the raw image to obtain the block partition. For two-stage detectors such as Faster R-CNN [43], we scale the ROI to the original image for block split. $\boldsymbol{c} \in \mathbb{Z}^{1 \times b}$ represents the object class in b blocks of the image. We define c_i as the i_{th} element of $\boldsymbol{c}$, which demonstrates the class of the object whose center is in the i_{th} block of the image. The class of a block is assigned to background if the block does not contain the centers of any groundtruth objects.

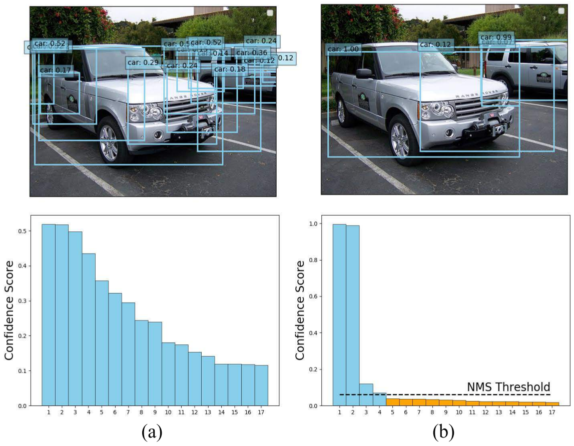


Figure 3. The detected objects and the corresponding confidence score (a) before and (b) after optimizing (6). The contrast of confidence score among different detected objects is significantly enlarged by minimizing alternate objective. As the NMS eliminates the positives with confidence score lower than the threshold, the sparse object priors are acquired and the posteriors are enforced to be concentrated on informative prediction. (best viewed in color).

As the localization contains shift parameters and scale parameters for anchors, we rewrite the mutual information between the object location and high-level feature maps:

$$I(F;L) = \mathbb{E}_{\boldsymbol{f} \sim p(\boldsymbol{f}|\boldsymbol{x})} \mathbb{E}_{\boldsymbol{l}_1 \sim p(\boldsymbol{l}_1|\boldsymbol{f})} \mathbb{E}_{\boldsymbol{l}_2 \sim p(\boldsymbol{l}_2|\boldsymbol{f})} \log \frac{p(\boldsymbol{l}_1|\boldsymbol{f}) p(\boldsymbol{l}_2|\boldsymbol{f})}{p(\boldsymbol{l}_1) p(\boldsymbol{l}_2)}$$

where $\boldsymbol{l}_1 \in \mathbb{R}^{2 \times b}$ represents the horizontal and vertical shift offset of the anchors in b blocks of the image, and $\boldsymbol{l}_2 \in \mathbb{R}^{2 \times b}$ means the height and width scale offset of the anchors. For the anchor whose center (x, y) is in the i_{th} block with height h and width w, the offset changes the bounding box in the following way: $(x, y) \rightarrow (x, y) + \boldsymbol{l}_{1,i}$ and $(h, w) \rightarrow (h, w) \cdot exp(\boldsymbol{l}_{2,i})$, where $\boldsymbol{l}_{1,i}$ and $\boldsymbol{l}_{2,i}$ represent the i_{th} column of $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$. The priors and the posteriors of shift offset conditioned on the feature maps are denoted as $p(\boldsymbol{l}_1)$ and $p(\boldsymbol{l}_1|\boldsymbol{f})$ respectively. Similarly, the scaling offset has the prior and the posteriors given feature maps $p(\boldsymbol{l}_2)$ and $p(\boldsymbol{l}_2|\boldsymbol{f})$. We leverage the localization branch in the detection part for posterior distribution parameterization.

3.1.2 Learning Sparse Object Priors

Since the feature maps are binarized in BiDet, we utilize the binomial distribution with equal probability as the priors for each element of the high-level feature map $\boldsymbol{f}$. We assign the priors for object localization in the following form:

$p(\boldsymbol{l}_{1,i}) = N(\boldsymbol{\mu}^0_{1,i}, \boldsymbol{\Sigma}^0_{1,i})$ and $p(\boldsymbol{l}_{2,i}) = N(\boldsymbol{\mu}^0_{2,i}, \boldsymbol{\Sigma}^0_{2,i})$, where $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ means the Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. For one-stage detectors, the object localization priors $p(\boldsymbol{l}_{1,i})$ and $p(\boldsymbol{l}_{2,i})$ are hypothesized to be the two-dimensional standard normal distribution. For two-stage detectors, Region Proposal Networks (RPN) output the parameters of the Gaussian priors.

As numerous false positives emerge in the detection prediction of binary networks, learning sparse object priors for detection part enforces the posteriors to be concentrated on informative detection prediction with false positive elimination. The priors for object classification is defined as follows:

$$p(c_i) = \mathbb{I}_{M_i} \cdot cat(\frac{1}{n+1} \cdot \mathbf{1}^{n+1}) + (1 - \mathbb{I}_{M_i}) \cdot cat([1, \mathbf{0}^n])$$

where $\mathbb{I}_x$ is the indicator function with $\mathbb{I}_1 = 1$ and $\mathbb{I}_0 = 0$, and M_i is the i_{th} element of the block mask $\boldsymbol{M} \in \{0,1\}^{1\times b}$. $cat(\boldsymbol{K})$ means the categorical distribution with the parameter $\boldsymbol{K}$. $\mathbf{1}^n$ and $\mathbf{0}^n$ are the all-one and zero vectors in n dimensions respectively, where n is the number of object classes. The multinomial distribution with equal probability is utilized for class priors in the i_{th} block if M_i equals to one. Otherwise, the categorical distribution with the probability one for background and zero probability for other object classes is leveraged for the prior class distribution. When M_i equals to zero, the class branch in the detection part definitely predicts the background for object in the i_{th} block according to (5). In order to obtain sparse priors for object classification with fewer predicted positives, we minimize the L_1 norm of the block mask $\boldsymbol{M}$. We propose an alternative way to optimize $\boldsymbol{M}$ due to the non-differentiability, where the objective is written as follows:

$$\min_{s_j} -\frac{1}{m}\sum_{j=1}^{m} s_j \log s_j \tag{6}$$

where m represents the number of detected foreground objects in the image. s_j is the normalized foreground confidence score of the j_{th} bounding box in the image, and the normalization is performed over the foreground confidence score of all m bounding boxes. We define the normalization as $s_j = \frac{p^o_j}{\sum_{k=1}^{m} p^o_k}$, where p^o_j is the original foreground confidence score of the j_{th} bounding box in the image. As shown in Figure 3, minimizing (6) increases the contrast of foreground confidence score among different predicted positives, and the predicted objects with low foreground confidence score are assigned to be negative by the non-maximum suppression (NMS) algorithms [37]. Therefore, the block mask becomes sparser with fewer predicted objects, and the posteriors are concentrated on informative prediction with false positive elimination.

3.1.3 Efficient Binarized Object Detectors

In this section, we first briefly introduce neural networks with binary weights and activations, and then detail the learning objectives of our BiDet. Let $\boldsymbol{W^l_r}$ be the real-valued weights and $\boldsymbol{A^l_r}$ be the full-precision activations of the l_{th} layer in a given L-layer detection model. During the forward propagation, the weights and activations are binarized via the sign function: $\boldsymbol{W^l_b} = sign(\boldsymbol{W^l_r})$ and $\boldsymbol{A^l_b} = sign(\boldsymbol{W^l_r} \odot \boldsymbol{A^l_b})$. $sign$ means the element-wise sign function which maps the number larger than zero to one and otherwise to minus one, and $\odot$ indicates the element-wise binary product consisting of xnor and bitcount operations. Due to the non-differentiability of the sign function, straight-through estimator (STE) is employed to calculate the approximate gradients and update the real-valued weights in the back-propagation stage. The learning objective for the proposed BiDet is written as follows:

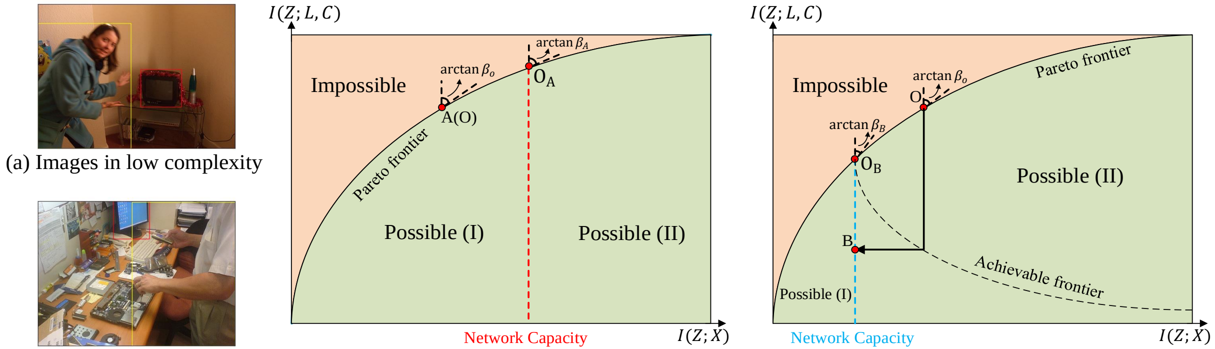
$$\begin{aligned} \min J &= J_1 + J_2 \\ &= (\sum_{t,s} \log \frac{p(f_{st}|\boldsymbol{x})}{p(f_{st})} - \beta \sum_{i=1}^{b} \log \frac{p(c_i|\boldsymbol{f})p(\boldsymbol{l}_{1,i}|\boldsymbol{f})p(\boldsymbol{l}_{2,i}|\boldsymbol{f})}{p(c_i)p(\boldsymbol{l}_{1,i})p(\boldsymbol{l}_{2,i})}) \\ &\quad - \gamma \cdot \frac{1}{m}\sum_{j=1}^{m} s_j \log s_j \end{aligned} \tag{7}$$

where γ is a hyperparameter that balances the importance of sparsity in object priors. The posterior distribution $p(c_i|\boldsymbol{f})$ is hypothesized to be the categorical distribution $cat(\boldsymbol{K}_i)$, where $\boldsymbol{K}_i \in \mathbb{R}^{1\times(n+1)}$ is the parameter and n is the number of classes. We assume the posterior distribution of the shift and scale offset follows the Gaussian distribution: $p(\boldsymbol{l}_{1,i}|\boldsymbol{f}) = N(\boldsymbol{\mu}_{1,i}, \boldsymbol{\Sigma}_{1,i})$ and $p(\boldsymbol{l}_{2,i}|\boldsymbol{f}) = N(\boldsymbol{\mu}_{2,i}, \boldsymbol{\Sigma}_{2,i})$. The posteriors of the element in the s_{th} row and t_{th} column of binary high-level feature maps $p(f_{st}|\boldsymbol{x})$ are assigned to binomial distribution $cat([p_{ts}, 1 - p_{ts}])$, where p_{ts} is the probability for f_{st} to be one. All the posterior distribution is parameterized by neural networks. J_1 represents for the information bottleneck employed in object detection, which aims to remove information redundancy and fully utilize the representational power of the binary neural networks. The goal of J_2 is to enforce the object priors to be sparse so that the posteriors are encouraged to be concentrated on informative prediction with false positive elimination.

In the learning objective, $p(f_{st})$ in the binomial distribution is a constant. Meanwhile, the sparse object class priors are imposed via J_2 so that $p(c_i)$ is also regarded as a constant. For one-stage detectors, constant $p(\boldsymbol{l}_{1,i})$ and $p(\boldsymbol{l}_{2,i})$ follow standard normal distribution. For two-stage detectors, $p(\boldsymbol{l}_{1,i})$ and $p(\boldsymbol{l}_{2,i})$ are parameterized by RPN which is learned by the objective function. Following [43], we iteratively train the RPN and the refining networks that predict priors and posteriors of bounding box location offset respectively, so that informative distribution of bounding box location can be obtained. The last layer of the backbone that outputs the parameters of the binary high-level feature maps is real-valued in training for Monte-Carlo sampling and is binarized with the sign function for deterministic forward propagation during inference. Meanwhile, the layers that output the parameters for object class and location distribution remain real-valued for accurate detection. During inference, we drop the network branch of covariance matrix for location offset, and assign all location prediction with the mean value of the Gaussian distribution to accelerate computation. Moreover, the prediction of object classes is set to that with the maximum probability to avoid time-consuming stochastic sampling in inference.

3.2 AutoBiDet

We first propose the binarized object detectors with automatic information compression, and then present the class-



(a) Images in low complexity

(b) Images in high complexity

(c) Information plane for image (a)

(d) Information plane for image (b)

Figure 4. The motivation of the proposed AutoBiDet. (a) and (b) illustrate the images in low and high complexity respectively. (c) and (d) depict the information plane for the representations of (a) and (b). The information plane is divided into the Impossible region and Possible region by the Pareto frontier, where the learned representations can only be in the Possible region. The mutual information between the high-level feature maps and the input is limited by the network capacity so that the high-level feature maps are restricted in the region of Possible (I). When the constant network capacity is fully utilized, the high level feature maps of images in high complexity carry less information regarding to the input due to the more significant information loss and vice versa. O_A and O_B mean the optimal representations for image (a) and (b) in the information plane, and the optimal hyperparameter β in (2) to control the IB trade-off is denoted as β_A and β_B respectively. As we apply the fixed IB trade-off in BiDet which is represented by O with the hyperparameter β_O in (2), the resulted representations for image (a) and (b) denoted as A and B are far from O_A and O_B. For images in low complexity, the fixed IB trade-off causes excessive compression and network capacity is not fully utilized. For images in high complexity, the insufficient redundancy removal obstacles the acquisition of representations in the Pareto frontier. When the network capacity is occupied by redundancy, the achievable frontier for representations is degraded significantly compared with the Pareto frontier.

aware sparse object priors to effectively alleviate false positives without recall decrease.

3.2.1 Binarized Object Detectors with Automatic Information Compression

The optimal IB trade-off varies with different input samples. Limiting the total information of the representation is more important for images in high complexity to completely remove the redundancy, while increasing the relevant information of learned representations is more significant for samples in low complexity to fully utilize the network capacity. As BiDet applies the fixed IB trade-off, the information compression leads to insufficient utilization of network capacity for images in low complexity and results in incomplete redundancy removal for images in high complexity. In order to address these limitations, we further propose binary neural networks with automatic information compression, where the optimal IB trade-off is dynamically selected according to the input samples.

Figure 4 demonstrates the information plane for the high-level feature maps of images in low and high complexity. According to the Pareto frontier in the information plane, the relevant information is increased when the high-level feature maps carry more information. As the network capacity of binary detectors is constrained, the learned high-level feature maps are limited in the region of Possible (I). When the constant network capacity is fully utilized, the high level feature maps of images in high complexity carry less information regarding to the input due to the more significant information loss during compression. The optimal IB trade-off is obtained when fully utilizing the network capacity and removing the redundancy to achieve the Pareto frontier. O_A and O_B mean the optimal representations of images in low and high complexity in the information plane, and the optimal hyperparameter β in (2) to control the IB trade-off is denoted as β_A and β_B respectively. The fixed IB trade-off denoted as O with the hyperparameter β_O results in representations in A and B for images in low and high complexity respectively, which are far from O_A and O_B respectively. The fixed IB trade-off leads to excessive compression for images in low complexity, where the network capacity is not fully utilized. For images in high complexity, the insufficient redundancy removal fails to learn representations in the Pareto frontier. When the network capacity is occupied by redundancy, the achievable frontier for representations is degraded significantly compared with the Pareto frontier. The proposed AutoBiDet automatically adjusts the IB trade-off according to the input complexity, so that the network capacity is fully utilized with complete redundancy removal for all samples.

Following [50], [38], we employ the complexity definition $\mathcal{C}(\boldsymbol{x})$ that is widely adopted:

$$\mathcal{C}(\boldsymbol{x}) = \inf_{\boldsymbol{w}} L(p(\boldsymbol{l}, \boldsymbol{c}|\boldsymbol{f})) \tag{8}$$

where $\mathcal{C}(\boldsymbol{x})$ means the complexity of $\boldsymbol{x}$, and $\boldsymbol{w}$ represents the network weights. $L(p(\boldsymbol{l}, \boldsymbol{c}|\boldsymbol{f}))$ is the discriminative loss for the prediction distribution $p(\boldsymbol{l}, \boldsymbol{c}|\boldsymbol{f})$. The samples that result in higher discriminative loss for the optimal neural networks are more complex. In our method, we employ the log likelihood for the discriminative loss, which is represented as $L(p(\boldsymbol{l}, \boldsymbol{c}|\boldsymbol{f})) = -\log p(\boldsymbol{l} = \boldsymbol{l_x}, \boldsymbol{c} = \boldsymbol{c_x}|\boldsymbol{f}) = -\log p(\boldsymbol{l} = \boldsymbol{l_x}|\boldsymbol{f})p(\boldsymbol{c} = \boldsymbol{c_x}|\boldsymbol{f})$, where $\boldsymbol{l_x}$ and $\boldsymbol{c_x}$ are the groundtruth location and class of the image $\boldsymbol{x}$. The complexity cannot be directly calculated during training, since the lower bound of $L(p(\boldsymbol{l}, \boldsymbol{c}|\boldsymbol{f}))$ for each image can only be obtained after acquiring the well-trained models. In order to leveraging the optimal IB trade-off during the training process, we propose the Complexity EStimator (CES) based on GAN to evaluate the sample complexity in training.

The CES consists of a generator and a discriminator. The generator aims to recover the input according to the high-level feature maps learned via the backbone, and the discriminator outputs the probability that the reconstructed

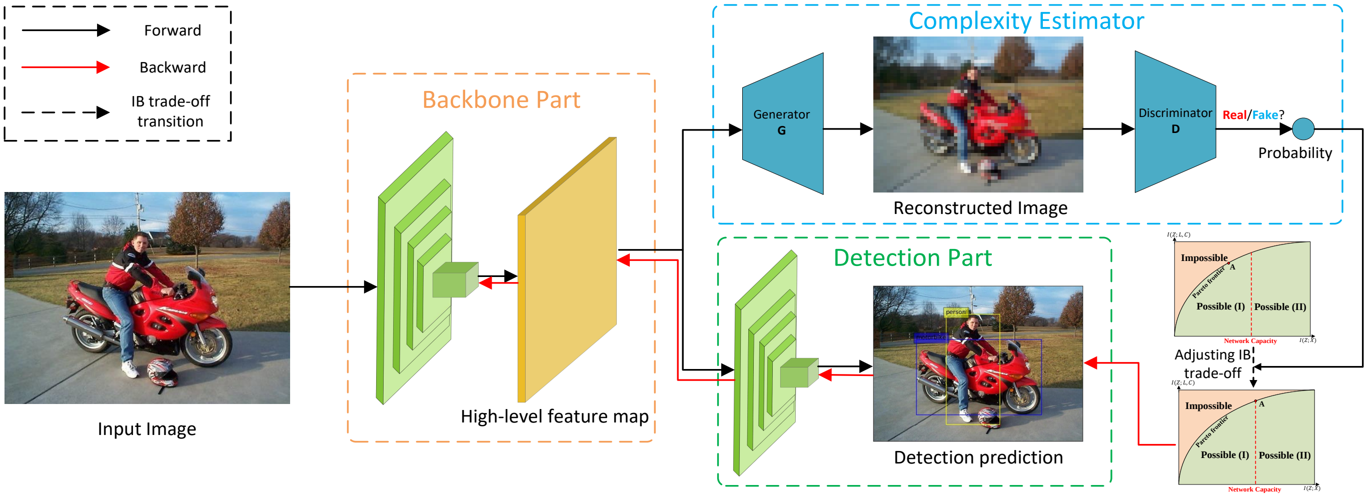


Figure 5. The pipeline of AutoBiDet. The black solid line, the red solid line and the black dashed line represent the forward propagation, the backward propagation and the IB trade-off transition in AutoBiDet respectively. The backbone part learns the high-level feature maps to represent the input image, and the detection part predicts the class and location of the objects in the input image according to the high-level feature map. The complexity estimator reconstructs the input image via the generator, and the discriminator outputs the probability that the recovered image comes from the true sample set. The probability from the discriminator is used to acquire the optimal IB trade-off via the transformation functions, and the dynamic IB trade-off is employed to train the backbone and detection networks. As a result, the network capacity is fully utilized and the redundancy is effectively removed for all input.

image is the real one. Figure 5 shows the pipeline of the presented AutoBiDet. High probability from the discriminator indicates that the network capacity is sufficient to extract the contained information in images for reconstruction, which indicates low input complexity. In contrast, low probability of the discriminator shows images in high complexity, so that the network capacity is deficient for precise image representation. Since less compression should be adopted for images in low complexity and vice versa, the discriminator probability is leveraged to automatically adjust the IB trade-off. We theoretically and empirically prove the strong correlation between the discriminator probability and image complexity in Appendix C. The CES and the binarized object detectors are iteratively optimized in each round. The GAN loss is applied for training the CES:

$$\min_{\mathcal{G}} \max_{\mathcal{D}} \log \mathcal{D}(\boldsymbol{x}) + \log(1 - \mathcal{D}(\mathcal{G}(\boldsymbol{f}))) \tag{9}$$

where $\mathcal{G}$ and $\mathcal{D}$ mean the generator and the discriminator respectively. Then we train the binarized object detectors with the objectives (7) that replaces fixed β with $\mathcal{T}(\mathcal{D}(\mathcal{G}(\boldsymbol{f})))$. $\mathcal{T}$ is the transformation function that maps the discriminator probability to the hyperparameter β. Since high discriminator probability depicts large weights for the mutual information between the learned high-level feature maps and object detection, we employ the linear, exponential and sine transformation functions with different convexity as follows:

$$\begin{aligned} \mathcal{T}(\mathcal{D}(\mathcal{G}(\boldsymbol{f}))) &= \delta_1 + \theta_1 \mathcal{D}(\mathcal{G}(\boldsymbol{f})) \\ \mathcal{T}(\mathcal{D}(\mathcal{G}(\boldsymbol{f}))) &= \delta_2 \exp(\theta_2 \mathcal{D}(\mathcal{G}(\boldsymbol{f}))) \\ \mathcal{T}(\mathcal{D}(\mathcal{G}(\boldsymbol{f}))) &= \delta_3 + \theta_3 \cdot \sin((\pi/2) \cdot \mathcal{D}(\mathcal{G}(\boldsymbol{f}))) \end{aligned} \tag{10}$$

where $\delta_1 \sim \delta_3$ and $\theta_1 \sim \theta_3$ are hyperparameters.

3.2.2 Class-aware Sparse Object Priors

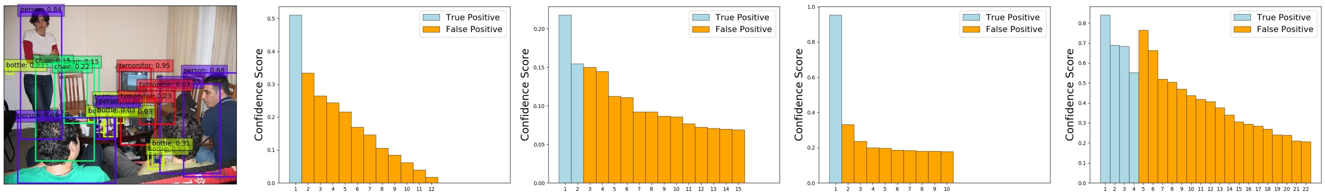
Since the sparse object priors (SOP) presented in BiDet constrain the number of predicted positives in each class equally, the recall is decreased for classes with few predicted positives and the false positive alleviation is insufficient for classes with many predicted positives. Figure 6 demonstrates the prediction for an example. It is observed that false positives are more likely to emerge in classes with many predicted positives. As a result, it is beneficial to impose strict object sparsity constraint on the classes with many predicted positives and less sparse object priors on classes with few predicted positives. In order to eliminate false positives effectively without recall degradation, we propose class-aware sparse object priors (C-SOP) to apply different sparsity on objects in various classes. The C-SOP is defined as follows:

$$p(c_i) = \mathbb{I}_{M_i} \cdot cat(\boldsymbol{\lambda}(\boldsymbol{x})) + (1 - \mathbb{I}_{M_i}) \cdot cat([1, \boldsymbol{0}^n])$$

where $\boldsymbol{\lambda}(\boldsymbol{x}) \in [0,1]^{n+1}$ illustrates the parameters of the categorical prior distribution for object classes in $\boldsymbol{x}$. Let us denote the t_{th} element of $\boldsymbol{\lambda}(\boldsymbol{x})$ as $\lambda_t(\boldsymbol{x})$. For the given input image $\boldsymbol{x}$, $\lambda_t(\boldsymbol{x})$ is expected to be small if the number of predicted positives in the t_{th} class is large, so that numerous false positives can be alleviated effectively. On the contrary, we require $\lambda_t(\boldsymbol{x})$ to be large when there are only few predicted positives in the t_{th} class, so that the recall is not degraded. Similar to BiDet, minimizing the L1 norm of $\boldsymbol{M}$ to obtain the C-SOP is non-differentiable. We also propose an alternative objective to solve the optimization difficulty and avoid the hand-crafted design for $\boldsymbol{\lambda}(\boldsymbol{x})$ in the following:

$$\min_{u_j^c} \sum_c -\frac{1}{n_c} \sum_{j=1}^{m_c} u_j^c \log u_j^c \tag{11}$$

where n_c is the number of groundtruth objects in the c_{th} class. u_j^c means the normalized foreground confidence score of the j_{th} bounding box in the c_{th} class, which is normalized across the foreground confidence score of all m_c bounding boxes in the c_{th} class. The definition of u_j^c is $u_j^c = \frac{p_{j,c}^o}{\sum_{k=1}^{m_c} p_{k,c}^o}$, where $p_{k,c}^o$ is the original foreground confidence score of the j_{th} bounding box of the c_{th} class. u_j^c should be encouraged



(a) Image (b) Bottle (c) Chair (d) TV monitor (e) Person

Figure 6. (a) The detection prediction of a sampled image, where boxes in different colors illustrate the object predictions for various classes. (b), (c), (d) and (e) demonstrate the foreground confidence score of all predicted positives for the classes of bottles, chairs, TV monitors and persons respectively, where blue and orange bars mean the true and false positives respectively. It is obvious that the predicted positives for the TV monitor are the fewest among the four classes, which leads to the minimum false positives among all classes. On the contrary, the class of persons obtain the most predicted positives and the false positives of persons are also the maximum.

to have large contrast among different predicted positives so that the posteriors are concentrated on the informative predicted positives with high foreground confidence score. Similar to BiDet, the false positives with low foreground confidence score is eliminated by NMS. Since more groundtruth positives also increase predicted positives, we add the weight $\frac{1}{n_c}$ in order to offset the impact brought by the number of groundtruth positives.

4 Experiments

In this section, we conducted comprehensive experiments to evaluate our proposed method on two datasets for object detection: PASCAL VOC [7] and COCO [27]. We first describe the implementation details of our BiDet and AutoBiDet. Secondly, we validate the effectiveness of IB and SOP, and investigate influence of automatic information compression and C-SOP for binarized object detectors by the ablation study. Thirdly, we compare our BiDet and AutoBiDet with state-of-the-art binary neural networks in the task of object detection to demonstrate the superiority. Finally, we generalize the presented techniques in BiDet and AutoBiDet to quantization, pruning and efficient architecture design to verify the benefit for other network compression methods.

4.1 Datasets and Implementation Details

We first introduce the datasets that we carried out experiments on and data preprocessing techniques:

PASCAL VOC: The PASCAL VOC dataset contains natural images from 20 different classes. We trained our model on the VOC 2007 and VOC 2012 trainval sets which consist of around 16k images, and we evaluated our method on VOC 2007 test set including about 5k images. Following [7], we used the mean average precision (mAP) as the evaluation criterion.

COCO: The COCO dataset consists of images from 80 different categories. We conducted experiments on the 2014 COCO object detection track. We trained our model with the combination of 80k images from the training set and 35k images sampled from validation set (trainval35k [2]), and tested our method on the remaining 5k images in the validation set (minival [2]). Following the standard COCO evaluation metric [27], we report the average precision (AP) for IoU $\in [0.5 : 0.05 : 0.95]$ denoted as mAP@[.5, .95]. We also report average precision with IOU threshold 50% and 75% (AP$_{50}$ and AP$_{75}$) as well as average precision of small, medium and large objects (AP$_s$, AP$_m$ and AP$_l$).

We trained our BiDet and AutoBiDet with the SSD300 [31] and Faster R-CNN [43] detection framework whose backbone were VGG16 [47] and ResNet-18 [12] respectively. Following the implementation of binary neural networks in [19], we remained the first and last layer in the detection networks real-valued. We used the data augmentation techniques in [31] and [43] when utilizing the SSD300 and Faster R-CNN detection frameworks respectively. In most cases, the backbone network was pre-trained on ImageNet [44] in the task of image classification, whose implementation details are shown in Appendix B.1. For training binarized object detectors in both BiDet and AutoBiDet, we jointly finetuned the backbone part and trained the detection part for the object detection task. The batch size was assigned to be 32 and the Adam optimizer [22] was applied. The learning rate was set initially as $1e-3$ with decay to $1e-4$ and $1e-5$ at the 40_{th} and 60_{th} epoch out of 80 epochs for PASCAL VOC, and started from 0.001 and multiplied 0.1 at the 6_{th} and 10_{th} epoch out of 12 epochs for COCO.

For AutoBiDet, we iteratively trained the binarized detector and the CES in each round. For the generator and the discriminator in CES, we employed the similar architecture as used in SN-GAN [35], which utilized the spectral normalization to stabilize the training of GAN. The difference was that we applied a convolutional layer as the input layer to feed forward the high-level feature maps learned by the backbone. Moreover, we increased the number of channels and the batch size to further stabilize the GAN training as suggested in [61] and [3]. Due to the limited GPU memory, the generator reconstructed images whose side length were $\frac{1}{2}$ and $\frac{1}{4}$ of the original size for SSD300 and Faster R-CNN respectively. Following [3], Adam optimizer [22] with a constant learning rate of $2e-4$ and $5e-5$ was adopted for the discriminator and generator respectively. Additionally, we updated the discriminator for two steps when optimizing the generator for one step during the training of GAN. For the GAN training in each round, we kept optimizing GANs until the loss of the generator converged.

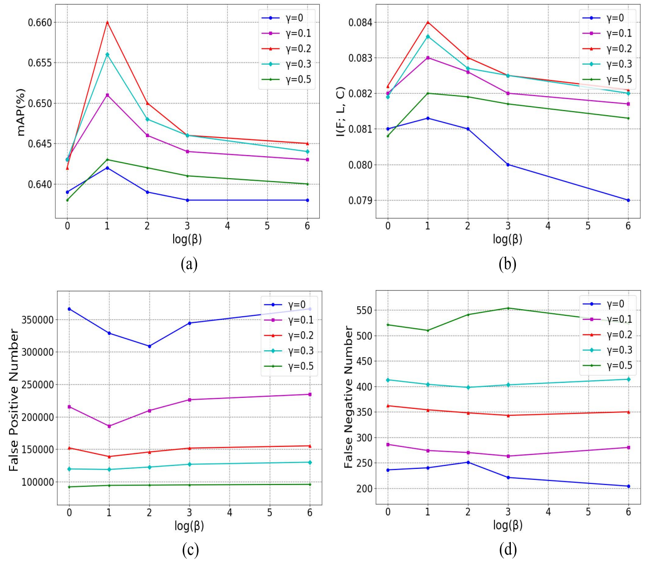
Hyperparamters β and γ in BiDet were set as 10 and 0.2 respectively. In AutoBiDet, we discovered that automatic information compression with the sine transformation function yielded the best result in all cases, where we assigned $\delta_3 = 10$, $\theta_3 = 5$ and $\gamma = 0.2$ in most experiments.

4.2 Ablation Study

In this section, we analyze the effect of the information bottleneck principle and the sparse object priors in BiDet at first, and then show the effectiveness of the proposed automatic information compression and the class-aware sparse object priors in AutoBiDet by the ablation study. For the experiments in this section, we adopted the SSD

Table 2
Comparison of the parameter size, FLOPs and mAP (%) with the state-of-the-art binary neural networks in both one-stage and two-stage detection frameworks on PASCAL VOC. BiDet (SC) and AutoBiDet (SC) mean the proposed methods with extra shortcut in the network architectures.

Framework	Input	Backbone	Quantization	W/A (bit)	#Params	MFLOPs	mAP
SSD300	300×300	VGG16	-	32/32	100.28MB	$31,750$	72.4
		MobileNetV1			30.07MB	$1,556$	68.0
		VGG16	TWN	2/32	24.54MB	$8,531$	67.8
			DoReFa-Net	4/4	29.58MB	$4,661$	69.2
			BNN	1/1	22.06MB	$1,275$	42.0
			Xnor-Net		22.16MB	$1,279$	50.2
			BiDet		22.06MB	$1,275$	52.4
			AutoBiDet		22.06MB	$1,275$	**53.5**
			Bi-Real-Net	1/1	21.88MB	$3,215$	63.8
			BiDet (SC)		21.88MB	$3,215$	66.0
			AutoBiDet (SC)		21.88MB	$3,215$	**67.5**
		MobileNetV1	Xnor-Net	1/1	22.48MB	836	48.9
			BiDet		22.48MB	836	51.2
			AutoBiDet		22.48MB	836	**52.0**
Faster R-CNN	600×1000	ResNet-18	-	32/32	47.35MB	$36,013$	74.5
			TWN	2/32	3.83MB	$9,196$	69.9
			DoReFa-Net	4/4	6.73MB	$4,694$	71.0
			BNN	1/1	2.38MB	779	35.6
			Xnor-Net		2.48MB	783	48.4
			BiDet		2.38MB	779	50.0
			AutoBiDet		2.38MB	779	**50.7**
			Bi-Real-Net	1/1	2.39MB	781	58.2
			BiDet (SC)		2.39MB	781	59.5
			AutoBiDet (SC)		2.39MB	781	**60.4**



(a)

(b)

(c)

(d)

Figure 7. Ablation study w.r.t. hyperparameters β and γ, where the variety of (a) mAP, (b) the mutual information between high-level feature maps and the object detection $I(F;L,C)$, (c) the number of false positives and (d) the number of false negatives are demonstrated. (best viewed in color).

detection framework with VGG16 backbone and evaluated the presented method on the PASCAL VOC dataset.

Table 1
mAP (%) on PASCAL VOC w.r.t. different transformation functions in automatic information compression (AIC) and different object priors. W/o and w/ mean without and with respectively. C-SOP w/o $\frac{1}{n_c^k}$ demonstrates C-SOP that fails to offset the impact brought by the number of groundtruth objects on predicted positives.

	w/o AIC	w/ AIC		
	-	Linear	Exponential	Sine
SOP	66.0	66.5	66.2	66.8
C-SOP w/o $\frac{1}{n_c^k}$	66.4	67.1	66.6	67.2
C-SOP	66.6	67.3	66.9	**67.5**

4.2.1 Ablation Study for BiDet

Since the IB principle removes the redundant information in binarized object detectors and the learned sparse object priors concentrate the posteriors on informative prediction with false positive alleviation, the detection accuracy is enhanced significantly. To verify the effectiveness of the IB principle and the learned sparse priors, we conducted the ablation study to evaluate our BiDet w.r.t. the hyperparameter β and γ in the objective function (7). We report the mAP, the mutual information between high-level feature maps and the object detection $I(F;L,C)$, the number of false positives and the number of false negatives with varing β and γ in Figure 7 (a), (b), (c) and (d) respectively.

By observing Figure 7 (a) and (b), we conclude that mAP and $I(F;L,C)$ are positively correlated as they demonstrate the detection performance and the amount of related information respectively. Medium β provides the optimal trade-off between the amount of total information and the related information in the high-level feature maps, so that the representational capacity of the binary object detectors is fully utilized with redundancy removal. Small β fails to fully leverage the representational power of the networks as the amount of extracted information is limited excessively, while large β enforces the networks to learn redundant information which leads to significant performance degradation. Meanwhile, medium γ offers optimal sparse object priors that enforces the posteriors to concentrate on infor-

Table 3
Comparison with the state-of-the-art binarized object detectors on COCO, where mAP@$[.5, .95]$ (%), AP with different IOU threshold and AP for objects in various sizes are demonstrated. BiDet (SC) and AutoBiDet (SC) mean the proposed methods with extra shortcut in the network architectures.

Framework	Input	Backbone	Quantization	mAP@$[.5, .95]$	AP$_{50}$	AP$_{75}$ (%)	AP$_s$	AP$_m$	AP$_l$
SSD300	300 × 300	VGG16	-	23.2	41.2	23.4	5.3	23.2	39.6
			TWN	16.9	33.0	15.8	5.0	16.9	27.2
			DoReFa-Net	19.5	35.0	19.6	5.1	20.5	32.8
			BNN	6.2	15.9	3.8	2.4	10.0	9.9
			Xnor-Net	8.1	19.5	5.6	2.6	8.3	13.3
			BiDet	9.8	22.5	7.2	3.1	10.8	16.1
			AutoBiDet	**10.6**	**23.0**	**7.9**	**3.1**	**11.3**	**17.1**
			Bi-Real-Net	11.2	26.0	8.3	3.1	12.0	18.3
			BiDet (SC)	13.2	28.3	10.5	5.1	14.3	20.5
			AutoBiDet (SC)	**14.3**	**30.3**	**12.2**	**5.6**	**16.1**	**21.9**
Faster R-CNN	600 × 1000	ResNet-18	-	26.0	44.8	27.2	10.0	28.9	39.7
			TWN	19.7	35.3	19.7	5.1	20.7	33.3
			DoReFa-Net	22.9	38.6	23.7	8.0	24.9	36.3
			BNN	5.6	14.3	2.6	2.0	8.5	9.3
			Xnor-Net	10.4	21.6	8.8	2.7	11.8	15.9
			BiDet	12.1	24.8	10.1	4.1	13.5	17.7
			AutoBiDet	**12.6**	**25.9**	**10.7**	**4.4**	**14.0**	**18.3**
			Bi-Real-Net	14.4	29.0	13.4	3.7	15.4	24.1
			BiDet (SC)	15.7	31.0	14.4	4.9	16.7	25.4
			AutoBiDet (SC)	**16.0**	**31.5**	**14.6**	**5.0**	**17.2**	**25.9**

mative prediction. Small γ is not capable of sparsifying the predicted objects, and large γ encourages the posteriors to predict objects with excessive sparsity.

By comparing the variety of false positives and false negatives w.r.t. β and γ, we know that medium β decreases false positives most significantly and changing β does not varies the number of false negatives notably. The redundancy removal only alleviates the uninformative false positives while remains the informative true positives unaffected. Meanwhile, small γ fails to constrain the false positives and large γ clearly increases the false negatives, which both degrade the performance significantly.

β and γ both affect the mAP on PASCAL VOC by 1-2% so that their influence on performance is similar. However, γ influences the number of false positives and false negatives more obviously as the sparse object priors directly control the number of detected objects via (5).

4.2.2 Ablation Study for AutoBiDet

Since the automatic information compression adopts less compression for images in low complexity and vice versa, the network capacity is fully utilized and the redundancy is completely removed for object detection. To investigate the influence of transformation functions on the detection performance, we implemented AutoBiDet with the linear, exponential and sine transformation function, where AutoBiDet without automatic information compression was also evaluated for reference. Meanwhile, C-SOP assigns sparser object priors for classes with more predicted positives and vice versa, the false positives can be eliminated effectively without recall decrease. To verify the effectiveness of different components in C-SOP, we experimented the AutoBiDet via SOP, C-SOP without the offset weight $\frac{1}{n_c^k}$ and C-SOP respectively. Table 1 shows the results.

Observing results across different columns, we know that automatic information compression can enhance the performance of the binarized object detectors since the network capacity is fully utilized and the redundancy is completely removed for images in various complexity. The sine transformation function obtains the best performance among all transformation functions in automatic information compression, as the IB trade-off in the objective changes obviously for images in high complexity and only varies slightly for samples in low complexity. When the network capacity is more sufficient, the IB-trade off needs to be more sensitive to keep the optimal because the tangent in the Pareto frontier shown in Figure 4 changes more significantly.

Comparing results in various rows, we conclude that C-SOP improves the mAP as the object priors are required to be sparser for classes with more predicted positives and vice versa. The false positives are significantly removed without recall degradation. We also implemented our AutoBiDet via C-SOP without $\frac{1}{n_c^k}$. Although directly applying different sparsity to objects in various classes based on the number of predicted positives can alleviated false positives, it fails to offset the impact brought by the number of groundtruth objects on the number of predicted positives. By adding the offset weight $\frac{1}{n_c^k}$ to the object priors of different classes, we obtain effective C-SOP and the mAP is further increased.

Other factors that influence the performance of the proposed AutoBiDet include the hyperparameters in the transformation function shown in (10). As demonstrated in Table 1, the sine transformation function leads to the best performance, so that we conducted the ablation study by grid search for δ_3 and θ_3. The experimental results are shown in Appendix B.3.

4.3 Comparison with the State-of-the-art Methods

In this section, we compare the proposed BiDet with the state-of-the-art binary neural networks including BNN [5], Xnor-Net [41] and Bi-Real-Net [32] in the task of object detection on the PASCAL VOC and COCO datasets. For

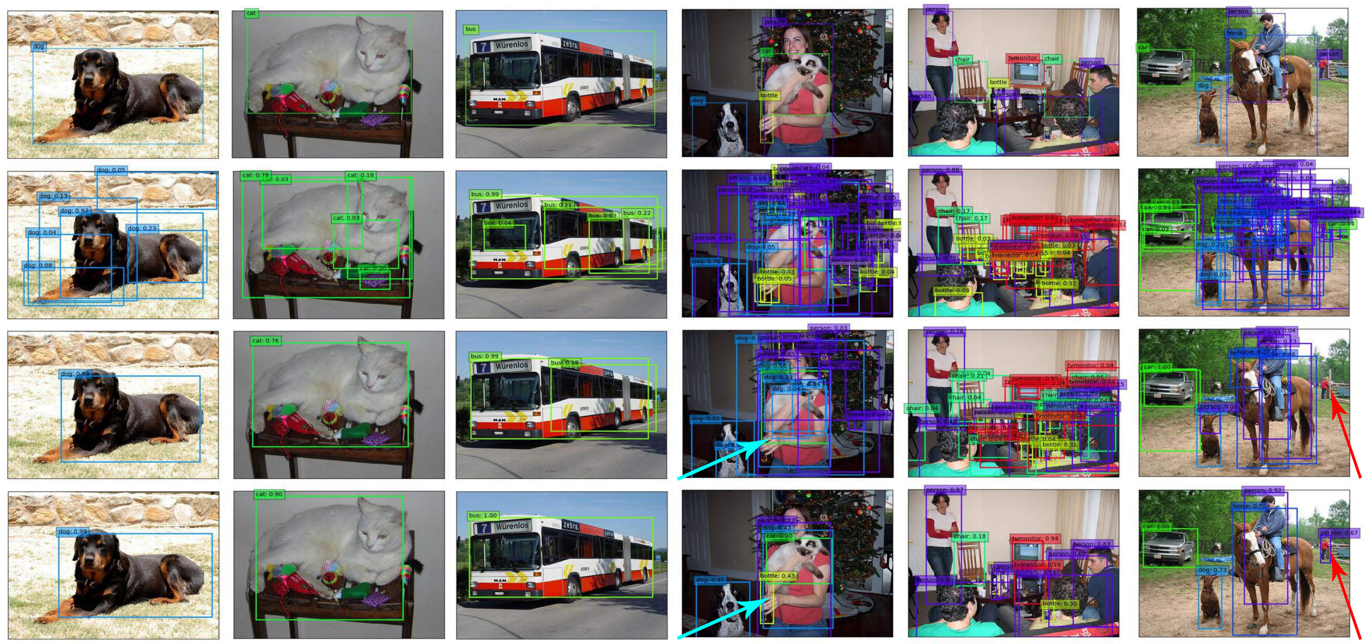


Figure 8. Qualitative results on PASCAL VOC. Images from the top to the bottom row show the groundtruth objects, the objects predicted by Xnor-Net, BiDet and AutoBiDet respectively. The proposed BiDet removes the false positives significantly compared with Xnor-Net. Moreover, our AutoBiDet eliminates the false positives more thoroughly for all classes and enhances the recall especially for small objects. The arrows in the figures represent objects missed by BiDet while detected by AutoBiDet (best viewed in color).

Table 4
Extension of techniques in BiDet and AutoBiDet to different model compression methods. The mAP@$[.5, .95]$ (%) in both one-stage and two-stage detection frameworks on COCO is reported for comparison. *Tech.* in BiDet means the proposed techniques in BiDet including IB and SOP, and *Tech.* in AutoBiDet represents the presented techniques in AutoBiDet containing automatic information compression and C-SOP.

Framework	Input	Backbone	Compression	mAP@$[.5, .95]$
SSD300	300×300	VGG16	-	23.2
			DoReFa-Net	19.5
			DoReFa-Net+*Tech.* in BiDet	20.0
			DoReFa-Net+*Tech.* in AutoBiDet	**20.4**
			SFP	18.2
			SFP+*Tech.* in BiDet	19.1
			SFP+*Tech.* in AutoBiDet	**19.7**
Faster R-CNN	600×1000	ResNet-18	-	26.0
			DoReFa-Net	22.9
			DoReFa-Net+*Tech.* in BiDet	23.4
			DoReFa-Net+*Tech.* in AutoBiDet	**23.6**
			SFP	22.9
			SFP+*Tech.* in BiDet	23.9
			SFP+*Tech.* in AutoBiDet	**24.4**
Light-Head R-CNN	800×1200	ShuffleNet-V2 x0.5	Light-Head R-CNN	22.5
			Light-Head R-CNN+*Tech.* in BiDet	22.7
			Light-Head R-CNN+*Tech.* in AutoBiDet	**23.0**

reference, we report the detection performance of the multi-bit quantized networks containing DoReFa-Net [64] and TWN [23] and the lightweight networks MobileNetV1 [17].

Results on PASCAL VOC: Table 2 illustrates the comparison of computation complexity, storage cost and the mAP across different quantization methods and detection frameworks. Our BiDet significantly accelerates the computation and saves the storage by $24.90\times$ and $4.55\times$ with the SSD300 detector in VGG16 and $46.23\times$ and $19.81\times$ with the Faster R-CNN detector in ResNet-18, and AutoBiDet shares the same computational and storage cost with BiDet. The efficiency is enhanced more notably in Faster R-CNN, as there are multiple real-valued output layers of the head networks in SSD300 for multi-scale feature extraction.

Compared with the state-of-the-art binary neural networks, the proposed BiDet improves the mAP of Xnor-Net by 2.2% and 1.6% with SSD300 and Faster R-CNN frameworks respectively while requiring fewer FLOPs and parameters than Xnor-Net. AutoBiDet further enhances the corresponding mAP by 1.1% and 0.7%. As demonstrated in [32], adding extra shortcut between consecutive convolutional layers can further enhance the representational power of the binary neural networks, we also employed architecture with additional skip connection to evaluate our BiDet and AutoBiDet in networks with stronger capacity. Due to the information redundancy, the performance of Bi-Real-Net with constrained network capacity is degraded significantly compared with their full-precision counterparts in both one-stage and two-stage detection frameworks. On the contrary, our BiDet imposes the IB principle on learning binary neural networks for object detection and fully utilizes the network capacity with redundancy removal. As a result, the proposed BiDet increases the mAP of Bi-Real-Net by 2.2% and 1.3% in SSD300 and Faster R-CNN detectors respectively without additional computational and storage cost. However, the sub-optimal IB trade-off in BiDet leads to ineffective utilization of network capacity and insufficient redundancy removal for input samples in various complexity, and SOP in BiDet causes degraded recall and incomplete false positive elimination due to the equal sparsity imposed on objects in different classes. The proposed AutoBiDet automatically learns the optimal IB trade-off with the class-aware sparsity in the object priors, where the mAP is further raised by 1.5% and 0.9% in SSD300 and Faster R-CNN detectors respectively with additional shortcut.

Due to the different pipelines in one-stage and two-stage detectors, the mAP gained from the proposed BiDet and AutoBiDet with Faster R-CNN is less than SSD300. As analyzed in [29], one-stage detectors face the severe positive-negative class imbalance problem which two-stage detectors are free of, so that one-stage detectors are usually more vulnerable to false positives. Therefore, one-stage object detection framework obtains more benefits from the proposed BiDet and AutoBiDet, which learns the sparse object priors to concentrate the posteriors on informative prediction with false positive elimination.

Moreover, our BiDet and AutoBiDet can be integrated with other efficient networks in object detection for further computation speedup and storage saving. We employed our BiDet and AutoBiDet as a plug-and-play module in SSD detector with the MobileNetV1 backbone, and saves the computational and storage cost by $1.47\times$ and $1.38\times$ respectively. Compared with the detectors that directly binarize weights and activations in MobileNetV1 with Xnor-Net, BiDet and AutoBiDet improve the mAP by a sizable margin, which depicts the effectiveness of redundancy removal for networks with extremely low capacity.

Results on COCO: The COCO dataset is much more challenging than PASCAL VOC due to the high diversity and large scale. Table 3 demonstrates mAP, AP with different IOU threshold and AP of objects in various sizes for different methods. Compared with the state-of-the-art binary neural networks Xnor-Net, our BiDet improves the mAP by 1.7% and 1.7% in SSD300 and Faster R-CNN detection framework respectively due to the information redundancy removal. Moreover, the proposed BiDet also enhances the binary one-stage and two-stage detectors with extra shortcut by 2.0% and 1.3% on mAP. AutoBiDet further enhances BiDet sizably across different architectures with various detection frameworks. Comparing with the baseline methods of network quantization, our method achieves better performance in the AP with different IOU threshold and AP for objects in different sizes, which demonstrates the universality in different application settings.

Figure 8 shows the qualitative results of the groundtruth, Xnor-Net, BiDet and AutoBiDet in the SSD300 detection framework with VGG16. Compared with Xnor-Net, our BiDet significantly alleviates false positives. Moreover, the proposed AutoBiDet removes the false positives more thoroughly for all classes while enhances the recall especially for small objects. To show the intuitive logic and technical soundness of our method, we also provide the detection results and the predicted foreground confidence score of different images in Appendix A.

4.4 Extension on Other Compressed Object Detectors

Compressing detection models decreases the network capacity, where the information redundancy causes many false positives especially for highly compressed object detectors. In order to improve the performance of other compressed models for object detection, the proposed techniques including IB and SOP in BiDet and automatic information compression and C-SOP in AutoBiDet can be utilized as the off-the-shelf module to remove the redundancy for models with different compression methods. We combined our techniques with other compressed neural networks including the quantized model DoReFa-Net, the pruned model SFP [14] and the efficiently designed model Light-Head R-CNN [25], and evaluated the detection performance on COCO. For quantization and pruning methods, we applied the SSD300 detection framework with the VGG16 architecture and the Fast R-CNN detectors with the ResNet-18 architecture respectively. The ShuffleNet-V2 [34] was employed as the backbone for the efficiently designed model Light-Head R-CNN. The implementation details are demonstrated in Appendix B.2.1.

Since the distribution of pixel values in high-level feature maps are different in detectors with other compression techniques, we assumed various prior distribution and parameterize the posteriors differently. For quantized models, we assumed the priors to be the $2n$-class categorical distribution with equal probability for each class, where n was the bitwidth of high-level feature maps. The posteriors were also designed as the $2n$-class categorical distribution that was parameterized by the backbone. For pruned and efficiently designed models, the priors and posteriors were both assigned to be the Gaussian distribution. The mean and the variance of each pixel were assumed to be zero and one respectively for the prior distribution, and those for the posteriors were parameterized by the backbone.

Table 4 shows the results, and we provide more evaluation of our techniques on different compressed detection models in Appendix B.2.2. The proposed techniques in BiDet and AutoBiDet both enhance the performance of the vanilla compression methods, which demonstrates the effectiveness of the information redundancy removal with false positive elimination for compressed object detectors. Compared with other model compression methods, the performance increase in binarized object detectors is more sizable due to the extremely low network capacity, which benefits more from the redundancy removal.

5 Conclusion

In this paper, we have proposed a binarized neural network learning method called BiDet for efficient object detection. The presented BiDet removes the redundant information via information bottleneck principle to fully utilize the network capacity, and enforces the posteriors to be concentrated on informative prediction to eliminate false positives via sparse object priors. We have also presented AutoBiDet that automatically learns the optimal IB trade-off for input samples in different complexity, and designed the class-aware sparse object priors to completely eliminate the false positives without recall degradation. Extensive experiments have depicted the superiority of BiDet and AutoBiDet in object detection compared with the state-of-the-art binary neural networks, and the presented techniques of our method have been generalized to other compression techniques to further enhance the vanilla model.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFA0700802, in part by the National Natural Science Foundation of China under Grant 61822603, Grant U1813218, Grant U1713214, and Grant 61672306, in part by Beijing Academy of Artificial Intelligence (BAAI), and in part by the Shenzhen Fundamental Research Fund (Subject Arrangement) under Grant JCYJ20170412170602564.

References

[1] Rana Ali Amjad and Bernhard Claus Geiger. Learning representations for neural network-based classification using the information bottleneck principle. *TPAMI*, 2019.

[2] Sean Bell, C Lawrence Zitnick, Kavita Bala, and Ross Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, pages 2874–2883, 2016.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NIPS*, pages 742–751, 2017.

[5] Matthieu Courbariaux, Yoshua Bengio, and Jean-Pierre David. Binaryconnect: Training deep neural networks with binary weights during propagations. In *NIPS*, pages 3123–3131, 2015.

[6] Bin Dai, Chen Zhu, and David Wipf. Compressing neural networks using the variational information bottleneck. *arXiv preprint arXiv:1802.10399*, 2018.

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 88(2):303–338, 2010.

[8] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015.

[9] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014.

[10] Ruihao Gong, Xianglong Liu, Shenghu Jiang, Tianxiang Li, Peng Hu, Jiazhen Lin, Fengwei Yu, and Junjie Yan. Differentiable soft quantization: Bridging full-precision and low-bit neural networks. *arXiv preprint arXiv:1908.05033*, 2019.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[14] Yang He, Guoliang Kang, Xuanyi Dong, Yanwei Fu, and Yi Yang. Soft filter pruning for accelerating deep convolutional neural networks. *arXiv preprint arXiv:1808.06866*, 2018.

[15] Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *ICCV*, pages 1389–1397, 2017.

[16] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *NIPS*, pages 3–10, 1994.

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[19] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Binarized neural networks. In *NIPS*, pages 4107–4115, 2016.

[20] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, pages 2704–2713, 2018.

[21] Hyeji Kim, Muhammad Umar Karim Khan, and Chong-Min Kyung. Efficient neural network compression. In *CVPR*, pages 12569–12577, 2019.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Fengfu Li, Bo Zhang, and Bin Liu. Ternary weight networks. *arXiv preprint arXiv:1605.04711*, 2016.

[24] Rundong Li, Yan Wang, Feng Liang, Hongwei Qin, Junjie Yan, and Rui Fan. Fully quantized network for object detection. In *CVPR*, pages 2810–2819, 2019.

[25] Zeming Li, Chao Peng, Gang Yu, Xiangyu Zhang, Yangdong Deng, and Jian Sun. Light-head r-cnn: In defense of two-stage object detector. *arXiv preprint arXiv:1711.07264*, 2017.

[26] Shaohui Lin, Rongrong Ji, Chao Chen, Dacheng Tao, and Jiebo Luo. Holistic cnn compression via low-rank decomposition with knowledge transfer. *TPAMI*, 2018.

[27] 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, 2014.

[28] 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.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017.

[30] Hanxiao Liu, Karen Simonyan, and Yiming Yang. Darts: Differentiable architecture search. *arXiv preprint arXiv:1806.09055*, 2018.

[31] 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, 2016.

[32] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *ECCV*, pages 722–737, 2018.

[33] Christos Louizos, Karen Ullrich, and Max Welling. Bayesian compression for deep learning. In *NIPS*, pages 3288–3298, 2017.

[34] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *ECCV*, pages 116–131, 2018.

[35] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

[36] Pavlo Molchanov, Arun Mallya, Stephen Tyree, Iuri Frosio, and Jan Kautz. Importance estimation for neural network pruning. In *CVPR*, pages 11264–11272, 2019.

[37] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *ICPR*, volume 3, pages 850–855, 2006.

[38] Cuong V Nguyen, Tal Hassner, Cedric Archambeau, and Matthias Seeger. Leep: A new measure to evaluate transferability of learned representations. *arXiv preprint arXiv:2002.12462*, 2020.

[39] Bo Peng, Wenming Tan, Zheyang Li, Shun Zhang, Di Xie, and Shiliang Pu. Extreme network compression via filter group approximation. In *ECCV*, pages 300–316, 2018.

[40] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection. *arXiv preprint arXiv:1903.11752*, 2019.

[41] Mohammad Rastegari, Vicente Ordonez, Joseph Redmon, and Ali

Farhadi. Xnor-net: Imagenet classification using binary convolutional neural networks. In *ECCV*, pages 525–542, 2016.

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016.

[43] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, *et al.* Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015.

[45] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *CVPR*, pages 4510–4520, 2018.

[46] Yuming Shen, Jie Qin, Jiaxin Chen, Li Liu, Fan Zhu, and Ziyi Shen. Embarrassingly simple binary representation learning. In *ICCVW*, pages 0–0, 2019.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, and Quoc V Le. Mnasnet: Platform-aware neural architecture search for mobile. In *CVPR*, pages 2820–2828, 2019.

[49] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[50] Anh T Tran, Cuong V Nguyen, and Tal Hassner. Transferability and hardness of supervised classification tasks. In *ICCV*, pages 1395–1405, 2019.

[51] Karen Ullrich, Edward Meeds, and Max Welling. Soft weight-sharing for neural network compression. *arXiv preprint arXiv:1702.04008*, 2017.

[52] Ji Wang, Weidong Bao, Lichao Sun, Xiaomin Zhu, Bokai Cao, and S Yu Philip. Private model compression via knowledge distillation. In *AAAI*, volume 33, pages 1190–1197, 2019.

[53] Robert J Wang, Xiang Li, and Charles X Ling. Pelee: A real-time object detection system on mobile devices. In *NIPS*, pages 1963–1972, 2018.

[54] Ziwei Wang, Jiwen Lu, Chenxin Tao, Jie Zhou, and Qi Tian. Learning channel-wise interactions for binary convolutional neural networks. In *CVPR*, pages 568–577, 2019.

[55] Ziwei Wang, Ziyi Wu, Jiwen Lu, and Jie Zhou. Bidet: An efficient binarized object detector. *arXiv preprint arXiv:2003.03961*, 2020.

[56] Yi Wei, Xinyu Pan, Hongwei Qin, Wanli Ouyang, and Junjie Yan. Quantization mimic: Towards very tiny cnn for object detection. In *ECCV*, pages 267–283, 2018.

[57] Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Yanghan Wang, Fei Sun, Yiming Wu, Yuandong Tian, Peter Vajda, Yangqing Jia, and Kurt Keutzer. Fbnet: Hardware-aware efficient convnet design via differentiable neural architecture search. In *CVPR*, pages 10734–10742, 2019.

[58] Tailin Wu and Ian Fischer. Phase transitions for the information bottleneck in representation learning. *arXiv preprint arXiv:2001.01878*, 2020.

[59] Tailin Wu, Ian Fischer, Isaac L Chuang, and Max Tegmark. Learnability for the information bottleneck. *Entropy*, 21(10):924, 2019.

[60] Jiwei Yang, Xu Shen, Jun Xing, Xinmei Tian, Houqiang Li, Bing Deng, Jianqiang Huang, and Xian-sheng Hua. Quantization networks. In *CVPR*, pages 7308–7316, 2019.

[61] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[62] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, pages 6848–6856, 2018.

[63] Chenglong Zhao, Bingbing Ni, Jian Zhang, Qiwei Zhao, Wenjun Zhang, and Qi Tian. Variational convolutional neural network pruning. In *CVPR*, pages 2780–2789, 2019.

[64] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. Dorefa-net: Training low bitwidth convolutional neural networks with low bitwidth gradients. *arXiv preprint arXiv:1606.06160*, 2016.

[65] Chenchen Zhu, Yihui He, and Marios Savvides. Feature selective anchor-free module for single-shot object detection. In *CVPR*, pages 840–849, 2019.

Ziwei Wang received the BS degree from the Department of Physics, Tsinghua University, China, in 2018. He is currently working toward the PhD degree in the Department of Automation, Tsinghua University, China. His research interests include network compression and binary representation. He has published 7 scientific papers in the IEEE Transactions on Pattern Analysis and Machine Intelligence, CVPR and ECCV. He serves as a regular reviewer member for the IEEE Transactions on Image Processing, Pattern Recognition Letters, CVPR, NeurIPS, ICPR and ICIP.

Jiwen Lu (M'11-SM'15) received the B.Eng. degree in mechanical engineering and the M.Eng. degree in electrical engineering from the Xi'an University of Technology, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in electrical engineering from Nanyang Technological University, Singapore, in 2012. He is currently an Associate Professor with the Department of Automation, Tsinghua University, Beijing, China. His current research interests include computer vision, pattern recognition, and intelligent robotics, where he has authored/co-authored over 250 scientific papers in these areas. He was/is a member of the Multimedia Signal Processing Technical Committee and the Information Forensics and Security Technical Committee of the IEEE Signal Processing Society, and a member of the Multimedia Systems and Applications Technical Committee and the Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society. He serves as the Co-Editor-of-Chief for Pattern Recognition Letters, an Associate Editor for the IEEE Transactions on Image Processing, the IEEE Transactions on Circuits and Systems for Video Technology, the IEEE Transactions on Biometrics, Behavior, and Identity Science, and Pattern Recognition. He also serves as the Program Co-Chair of IEEE FG'2023, IEEE VCIP'2022, IEEE AVSS'2021 and IEEE ICME'2020.

Ziyi Wu is currently an undergraduate student with the Department of Automation, Tsinghua University, China. His research interests include computer vision, efficient inference and point cloud analysis. He has obtained the National Scholarship of Tsinghua in 2018.

Jie Zhou (M'01-SM'04) received the BS and MS degrees both from the Department of Mathematics, Nankai University, Tianjin, China, in 1990 and 1992, respectively, and the PhD degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology (HUST), Wuhan, China, in 1995. From then to 1997, he served as a postdoctoral fellow in the Department of Automation, Tsinghua University, Beijing, China. Since 2003, he has been a full professor in the Department of Automation, Tsinghua University. His research interests include computer vision, pattern recognition, and image processing. In recent years, he has authored more than 100 papers in peer-reviewed journals and conferences. Among them, more than 30 papers have been published in top journals and conferences such as the IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE Transactions on Image Processing, and CVPR. He is an associate editor for the IEEE Transactions on Pattern Analysis and Machine Intelligence and two other journals. He received the National Outstanding Youth Foundation of China Award. He is a senior member of the IEEE.