Learning 1-Bit Tiny Object Detector with Discriminative Feature Refinement

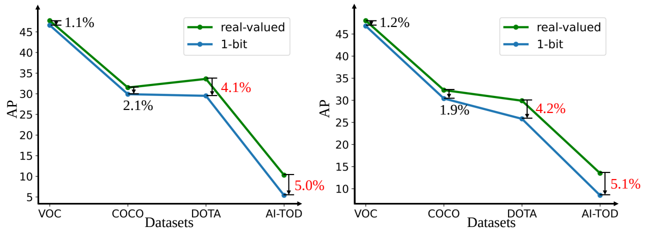
Sheng Xu [*1] **Mingze Wang** [*1] **Yanjing Li** [*1] **Mingbao Lin** [2] **Baochang Zhang** [1 3] **David Doermann** [4] **Xiao Sun** [5]

Abstract

1-bit detectors show impressive performance comparable to their real-valued counterparts when detecting commonly sized objects while exhibiting significant performance degradation on tiny objects. The challenge stems from the fact that high-level features extracted by 1-bit convolutions seem less compelling to reveal the discriminative foreground features. To address these issues, we introduce a **D**iscriminative **F**eature **R**efinement method for 1-bit **Det**ectors (DFR-Det), aiming to enhance the discriminative ability of foreground representation for tiny objects in aerial images. This is accomplished by refining the feature representation using an information bottleneck (IB) to achieve a distinctive representation of tiny objects. Specifically, we introduce a new decoder with a foreground mask, aiming to enhance the discriminative ability of high-level features for the target but suppress the background impact. Additionally, our decoder is simple but effective and can be easily mounted on existing detectors without extra burden added to the inference procedure. Extensive experiments on various tiny object detection (TOD) tasks demonstrate DFR-Det's superiority over state-of-the-art 1-bit detectors. For example, 1-bit FCOS achieved by DFR-Det achieves the 12.8% AP on AI-TOD dataset, approaching the performance of the real-valued counterpart.

1. Introduction

Recently, the tiny object detection (TOD) task (Wang et al., 2021; Ding et al., 2021) has significantly been promoted due to advances in deep neural networks (DNNs) (He et al., 2016), which is widely used in various real-world scenarios such as driving assistance, traffic management, and maritime rescue. However, such DNN-based detectors comprise many parameters and floating-point operations (FLOPs), restricting their deployment on embedded platforms in real-world scenarios. Techniques such as compact network design (Howard et al., 2017; Ma et al., 2018), network pruning (Li et al., 2016), low-rank decomposition (Denil et al., 2013), and quantization (Xu et al., 2022b; Zhao et al., 2022; Xu et al., 2023) have been developed to address these restrictions and accomplish an efficient inference on the detection task. Among these, 1-bit detectors have contributed to object detection by accelerating the CNN feature extracting for real-time bounding box localization and foreground classification (Wang et al., 2020; Xu et al., 2021; 2022c).



(a) AP of Faster-RCNN

(b) AP of FCOS

Figure 1. Average precision (AP) of (a) Faster-RCNN, and (b) FCOS using ResNet-18 backbone on various datasets with different sizes of objects. The AP of real-valued detectors and 1-bit counterparts are both shown.

When detecting commonly sized objects (Everingham et al., 2009; Lin et al., 2014), current 1-bit detectors (Wang et al., 2020; Xu et al., 2022c) show impressive performance, comparable to the real-valued counterparts. However, such 1-bit detectors severely deteriorate when detecting tiny objects featured with very few pixels. As shown in Fig. 1, the baseline 1-bit detector shows an increasing performance gap compared with real-valued counterparts with decreased object size. More specifically, on images possessing commonly sized objects, the performance gap between 1-bit detectors and their real-valued counterparts is about 1%∼2%, relatively small. However, such performance gaps magnify to about 4%∼5% on TOD datasets, such as DOTA (Ding et al., 2021) and AI-TOD (Wang et al., 2021).

The severe performance degradation mainly results from the poor capacity to refine the foreground information. This

[*]Equal contribution [1]Beihang University [2]Skywork AI [3]Zhongguancun Laboratory [4]University at Buffalo [5]Shanghai Artificial Intelligence Laboratory. Correspondence to: Baochang Zhang <bczhang@buaa.edu.cn>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

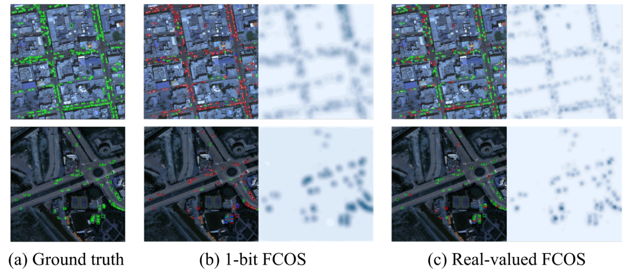


Figure 2. We visualize the (a) input image with ground truth label, prediction results (left), and saliency maps (right) of (b) vanilla 1-bit FCOS and (c) ral-valued FCOS. The **green**, **blue**, and **red** boxes denote true positive, false positive, and false negative, respectively. The saliency maps are plotted based on the ℓ_2-norm gradient in intermediate neck feature (Guo et al., 2021) at P_2. The darker area in the saliency map indicates the larger gradient.

can be observed by comparing to objects of ground-truth **green boxes** in Fig. 2(a) and massive false positives of **blue boxes** and false negatives of **red boxes** from the vanilla 1-bit FCOS in Fig. 2(b). In addition, the corresponding saliency maps (Guo et al., 2021) in the right part of Fig. 2(b) also manifest too much background noise to enable robust foreground information extraction. In comparison, the improved detection performance of the real-valued counterpart is attributed to the capacity to refine and discriminate salient foreground information from the noisy background information, as illustrated in the right part of Fig. 2(c).

In this paper, our **D**iscriminative **F**eature **R**efinement aims to improve the foreground discrimination of 1-bit **Det**ectors (DFR-Det) during the training process while bringing no extra burden to the inference procedure. Fig. 3 illustrates the proposed DFR-Det overview in cooperation with the 1-bit FCOS framework. Our intuition is analogous to how human beings recognize a small object in the cluttered background by possibly including attentional guidance to eliminate background regions that could be mistakenly regarded as the target (Wolfe et al., 2011; Choi et al., 2017). This is achieved by refining our detector upon information bottleneck (IB) principle (Shwartz-Ziv & Tishby, 2017), which involves the incorporation of a novel decoder that utilizes a foreground mask to boost the discriminative power of high-level features *w.r.t.* the foreground, while concurrently mitigating the negative impact of the background. In addition to benefiting a highly distinctive representation of tiny objects, our decoder is simple yet effective. It can be easily mounted on existing detectors without extra burden added to the inference procedure. Our major contributions in this paper are summarized as:

- We introduce a discriminative feature refinement (DFR) method to enhance the feature representation ability for 1-bit detectors on TOD, which is technically implemented by maximizing the mutual information between the intermediate feature and the input under the information bottleneck (IB) principle.
- We develop a new decoder network with a foreground mask to enhance the discriminative ability of high-level target features while simultaneously suppressing the negative impact from the clutter background.
- We compare our DFR-Det against state-of-the-art 1-bit detectors on various TOD datasets. Extensive results reveal that our DFR-Det outperforms state-of-the-art methods by a large margin. For instance, the FCOS detector using the ResNet-18 backbone obtained by DFR-Det achieves 12.8% AP on AI-TOD, achieving a new state-of-the-art.

2. Related Work

1-bit Detecters. BiDet (Wang et al., 2020) effectively leverages binarized convolutions by eliminating foreground redundancy, thus fully exploiting their representational capacity. This introduces an information bottleneck that restricts the data in high-level feature maps while maximizing mutual information between feature maps and object detection. LWS-Det (Xu et al., 2021) introduces a layer-wise searching approach, minimizing angular and amplitude errors for 1-bit detectors. Additionally, LWS-Det utilizes FGFI (Wang et al., 2019) to further distill the backbone feature map. The IDa-Det (Xu et al., 2022c) proposed an information discrepancy-aware distillation (IDa) method for 1-bit detectors. The IDa method localizes the distillation desired area with maximum information discrepancy, thus improving the effectiveness of distillation process.

Tiny Object Detection. Most current approaches for TOD can be grouped into four main categories: data augmentation, multi-scale feature learning, training strategy for tiny objects, and feature enhancement strategy.

In addition to conventional data augmentations, such as rotating, flipping images, and upsampling, Krisantal *et al.* (Kisantal et al., 2019) sought to enhance TOD performance by oversampling images that contain tiny objects and copy-pasting them. One conventional approach is to resize input images into different scales and train individual detectors at a particular scale range. To reduce the computation costs, some works (Liu et al., 2016; Lin et al., 2017a; Zhao et al., 2019) construct feature-level pyramids. For instance, SSD (Liu et al., 2016) detects objects from feature maps of different resolutions. Feature Pyramid Network (FPN) (Lin et al., 2017a) constructs a top-down structure with lateral connections to combine feature information of different scales for improving object detection performance. Subsequently, lots of techniques have endeavored to enhance FPN's performance, such as PANet (Liu et al.,

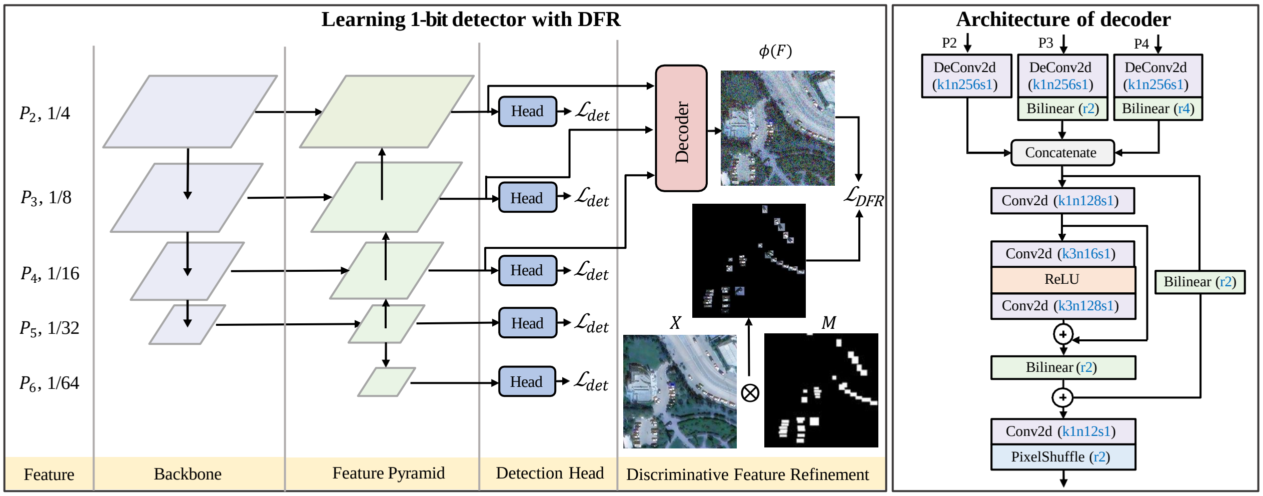


Figure 3. Overview of the proposed DFR-Det method with 1-bit FCOS framework. A sequence of feature maps with varying resolutions is generated by inputting the image to the backbone and Feature Pyramid Network (FPN). We feed the P_3, P_4, and P_5 layer of feature into the proposed decoder network and generate the reconstructed image $\phi(F)$. We then refine the foreground feature based on the ground truth mask. For better illustration, we brighten the image and select a patch in the figure.

2018a), BiFPN (Cai et al., 2018), and Recursive-FPN (Qiao et al., 2021). TridentNet (Li et al., 2019) builds multiple detection heads with diverse receptive fields to produce that are specific to certain scales.

Object detectors usually end with unsatisfactory performance on tiny objects and large objects simultaneously. Inspired by this fact, SNIP (Singh & Davis, 2018) and SNIPER (Singh et al., 2018) are designed to selectively train objects within a particular scale range. In addition, Kim *et al.* (Kim et al., 2018) introduced a scale-aware network that maps the features of different spaces onto a scale-invariant subspace, making detectors more robust to scale variation. Several techniques have been proposed to enhance the feature representation of small objects, such as super-solution and GAN. For example, PGAN (Li et al., 2017) introduces a GAN for small object detection. Bai *et al.* (Bai et al., 2018) proposed MT-GAN, which uses image-level super-resolution to improve small RoI features.

Unlike the methods above, we introduce a discriminative feature refinement strategy as a simple yet effective training method to improve the feature discriminative ability for 1-bit detectors from a new perspective. Also, our DFR-Det avoids any inference burden and can be easily combined with other enhancements for a better TOD task.

3. Methodology

In this section, we first overview the 1-bit detectors and design our objective under the IB principle. Then, we introduce a decoder network to improve the intermediate feature representation based on the input. Finally, we design a binary mask using the ground truth to decouple the foreground part, specifically designed to enhance the discriminative representation ability on tiny objects.

3.1. Problem Formulation

Preliminaries. In a specific convolutional layer, where $\mathbf{w}$ represents the weights, $\mathbf{a}_{in} \in \mathbb{R}^{C_{in}\times W_{in}\times H_{in}}$ denotes the input feature maps, and $\mathbf{a}_{out} \in \mathbb{R}^{C_{out}\times W_{out}\times H_{out}}$ is the output feature maps, where C_{in} and C_{out} denote the number of channels, and (H, W) represent the height and width of the feature maps, with K indicating the kernel size. We then have

$$\mathbf{a}_{out} = \mathbf{a}_{in} \otimes \mathbf{w}, \tag{1}$$

where $\otimes$ represents the convolution operation. For simplicity, we exclude batch normalization (BN) and activation layers. The objective of the 1-bit model is to quantize $\mathbf{w}$ and $\mathbf{a}_{in}$ into $\mathbf{b}^{\mathbf{w}} \in \{-1,+1\}^{C_{out}\times C_{in}\times K\times K}$ and $\mathbf{b}^{\mathbf{a}_{in}} \in \{-1,+1\}^{C_{in}\times H\times W}$ using efficient XNOR and Bit-count operations to replace full-precision operations. Following (Courbariaux et al., 2015; 2016), the forward process of the 1-bit detector is

$$\mathbf{a}_{out} = \boldsymbol{\alpha} \circ \mathbf{b}^{\mathbf{a}_{in}} \odot \mathbf{b}^{\mathbf{w}}, \tag{2}$$

where $\odot$ represents the XNOR and bit-count operations, and $\circ$ denotes channel-wise multiplication. $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_{C_{out}}] \in \mathbb{R}_+$ is the vector consisting of channel-wise scale factors. $\mathbf{b} = \text{sign}(\cdot)$ denotes the binarized variable using the sign function, which returns one if the input is greater than zero and -1 otherwise. It then goes through several non-linear layers, such as BN layer, non-linear activation layer, and max-pooling layer. We omit these for simplification. Afterward, the output $\mathbf{a}_{out}$ is binarized to $\mathbf{b}^{\mathbf{a}_{out}}$ via the sign function. The fundamental objective of BNNs is calculating $\mathbf{w}$. We aim for it to be as close as possible before and after binarization to minimize the binarization effect. Then, we define the reconstruction error as

$$L_R(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2}\|\mathbf{w} - \boldsymbol{\alpha} \circ \mathbf{b}^{\mathbf{w}}\|_2^2, \tag{3}$$

which is employed for all 1-bit convolution layers.

The 1-bit convolution shown in Eq. (2) possesses limited representation capacity compared with real-valued operations in Eq. (1), as revealed in multiple works (Liu et al., 2018b; Qin et al., 2020). Consequently, 1-bit detectors based on 1-bit convolutions show limited feature representation in the intermediate layers (Xu et al., 2021; 2022c). As shown in Fig. 2, such representation degradation leads to implicit foreground information, which needs to be refined for better detection results.

Objective Formulation. We start with a new perspective of the information bottleneck (IB) principle (Shwartz-Ziv & Tishby, 2017) to explore our framework. The IB advocates minimizing data misfitting and model complexity in concert such that the task-irrelevant information can be well diminished for better performance. A discriminative foreground representation is significant because the background information is often useless and disruptive and overwhelmingly takes up the feature space in tiny object detection.

Given an input three-channel image $X \in \mathbb{R}^{3\times W\times H}$, the conventional IB principle is written as:

$$\min_{\theta_{\boldsymbol{B}},\theta_{\boldsymbol{D}}} I(X;F) - \delta I(F;y^{GT}), \tag{4}$$

where $F \in \mathbb{R}^{C\times \frac{W}{s}\times\frac{H}{s}}$ is the high-level feature maps from the neck of the detector, s denotes the down-sampling stride, and y^{GT} denotes the ground-truth, $\theta_{\boldsymbol{B}}$ and $\theta_{\boldsymbol{D}}$ are the parameters of backbone and detection part in student respectively. Meanwhile, the δ is a Lagrange multiplier (Shwartz-Ziv & Tishby, 2017). The $I(\cdot)$ returns the mutual information between input variables. With $I(f;y^{GT})$ maximizing the mutual information between the features and ground truths, the first item $I(X;f)$ minimizes the mutual information between the input image and the high-level feature maps of the decoder to control the noise introduction. This part can be treated as the combination of original detection loss and the reconstruction error in Eq. (3) of the 1-bit detectors (Wang et al., 2020).

However, Eq. (4) shows its ineffectiveness for tiny object detection. The main reason lies in the unbalanced proportion of foreground and background features. Conventional CNNs fail to extract the discriminative foreground feature from the cluttered background without a unique design for unbalanced tiny objects. Consequently, it leads to less discriminative high-level features for the target and suffers from a cluttered background. As a result, the second item $I(F;y^{GT})$ is hard to maximize due to the less discriminative feature F. To address this issue, we reformulate the objective as:

$$\min_{\theta_{\boldsymbol{B}},\theta_{\boldsymbol{D}}} I(X;F) - \delta I(F;y^{GT}) - \lambda I(\phi(F);X_f), \tag{5}$$

where we add an additional item $I(F;X_f)$ to refine the foreground information. X_f denotes the foreground information of the input image and $\phi(\cdot)$ denotes a decoder network to keep the shape identical. More specifically, $I(\phi(F),X_f)$ is approximated to $I(\phi(F_f);X_f) + I(\phi(F_b);0)$, as the foreground and background feature are spatially mutually exclusive. The first item $I(F_f;X_f)$ denotes activating the foreground representation and reducing the foreground information degradation during feature extraction. The second item $I(\phi(F_b);0)$ is used to denoise the background feature and discriminate the foreground feature from targets. $I(\phi(F_b);0) = 0$ is equivalent with $\|\phi(F_b)\|_1 = 0$.

3.2. The Decoder Framework

Considering the foreground and background features are hard to decouple during inference, we start with the vanilla intermediate feature F. Under the IB principle, we use mutual information between intermediate feature F and input image X, which is formulated as

$$I(\phi(F);X_f) = H(\phi(F)) - H(\phi(F)|X_f), \tag{6}$$

where $H(\cdot)$ returns the information entropy. Considering the self-information entropy $H(\phi(F))$ remains unchanged within each iteration, $H(\phi(F))$ can be regarded as a constant. However, evaluating such information items directly is challenging since the input data X and intermediate feature F are not identically shaped. Inspired by several prior works (Talebi & Milanfar, 2021; Cui et al., 2022), we then introduce a decoder network to decode the intermediate feature and quantify the mutual information. Fig. 3 illustrates how to implement our decoder on FCOS (Tian et al., 2019). Given a down-sampled feature map group $F \in \mathbb{R}^{C\times\frac{W}{s}\times\frac{H}{s}}$, we first input them into corresponding deconvolution and bilinear interpolation operations to up-sample them to the size of $C \times \frac{W}{4} \times \frac{H}{4}$. Then, we concatenate the intra-group features alongside the channel dimension. We introduce the sequence of residual blocks to extract the feature following (Talebi & Milanfar, 2021). Finally, we squeeze the channel dimension to 12 using a linear layer and feed the feature into a 2× pixel shuffle layer (Shi et al., 2016) to generate the reconstructed image. After the decoder network $\phi(\cdot)$ is accomplished, we thus get the reconstructed image $\phi(F) \in \mathbb{R}^{3\times W\times H}$, which is equally shaped as the input image X.

3.3. Foreground Information Refinement

First, we propose to refine the foreground information based on a binary mask. Given the reconstructed response of size $3 \times W \times H$, we first generate the binary mask M according to the ground truth box B as

$$M_{i,j} = \mathbf{1}[(i,j) \in B], \tag{7}$$

where $M \in \{0,1\}^{W \times H}$. The indicative function $\mathbf{1}[(i,j) \in B]$ denotes the value of location (i,j) is 1 if it belongs to an object and 0 otherwise. Then, we use the generated binary mask to reformulate the IB objective as

$$\begin{aligned} I(\phi(F); X_f) &= I(\phi(F); M \circledast X) \\ &= H(\phi(F)) - H(\phi(F)|M \circledast X), \end{aligned} \tag{8}$$

where $\circledast$ denotes Hadamard product with broadcasting. $H(\phi(F)|M \circledast X)$ is implicit since the conditional distribution is hardly computed. Alternatively, we minimize the ℓ_1 norm distance between $\phi(F)$ and $M \circledast X$ following the super-resolution tasks (Farsiu et al., 2004). Then, we use the generated binary mask to decouple the supervision for discriminative feature refinement as

$$\mathcal{L}_{DFR}(\phi(F), X; M) = \|\phi(F) - M \circledast X\|_1, \tag{9}$$

where $\mathcal{L}_{DFR}$ denotes the proposed discriminative feature refinement loss.

Here, we analyze our foreground refinement loss in detail. The input image (also the label for super-resolution) is decoupled by a generated mask M. Hence, the proposed foreground refinement loss can be decoupled as

$$\mathcal{L}_{DFR}(\phi(F), X; M) = \|\phi(F_f) - X_f\|_1 + \|\phi(F_b)\|_1, \tag{10}$$

where $\phi(F_f)$ and $\phi(F_b)$ are the reconstructed foreground and background feature, respectively. Accordingly, we optimize the detector based on the total loss as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{detection} + \gamma \mathcal{L}_R}_{I(X;F) - \delta I(F;y^{GT})} + \quad \lambda(\underbrace{\|\phi(F_f) - X_f\|_1}_{-I(\phi(F_f);X_f)} + \underbrace{\|\phi(F_b)\|_1}_{-I(\phi(F_b);0)}), \tag{11}$$

where γ and λ are balanced hyper-parameters and $\mathcal{L}_{detection}$ is the original detection loss.

4. Experiments

Our DFR-Det is evaluated first on AI-TOD (Wang et al., 2021) dataset for the tiny object detection task. Then, we evaluate DFR-Det on the TinyPerson (Yu et al., 2020) and DOTA-v2.0 (Ding et al., 2021). In this section, we first introduce the implementation details of DFR-Det. Then, we select the hyper-parameter and validate the effectiveness of the components in the ablation study. Finally, we compare our method with state-of-the-art 1-bit detectors on various datasets to demonstrate the superiority of our DFR-Det.

4.1. Datasets and Implementation Details

Datasets. We evaluate the proposed method on AI-TOD (Wang et al., 2021), DOTA-v2.0 (Xia et al., 2018) and TinyPerson (Yu et al., 2020). The main experiments were conducted using the AI-TOD dataset, which presented a significant challenge to tiny object detection. The dataset comprises 28,036 aerial images with 800×800 pixels, featuring 700,621 object instances distributed across eight categories. The average object size in AI-TOD is merely 12.8 pixels, significantly smaller than other object detection datasets such as PASCAL VOC (156.6 pixels) (Everingham et al., 2009), MS COCO (99.5 pixels) (Lin et al., 2014).

The DOTA-v2.0 (Xia et al., 2018) dataset significantly advances object detection and aerial imagery analysis. Designed to meet contemporary computer vision research's escalating demands, this dataset is an essential resource for scholars and practitioners alike. Comprising a diverse array of high-resolution aerial images, the DOTA-v2.0 dataset encapsulates real-world scenarios across various landscapes and terrains, reflecting the challenges intrinsic to object detection tasks in complex environments. The dataset encompasses a meticulously curated collection of annotated objects, encompassing a broad spectrum of categories, sizes, and orientations. Each annotation is meticulously refined, contributing to a meticulously tailored benchmark that transcends the boundaries of its predecessor.

Implementation Details. We conduct the experiments on a computer with 1 NVIDIA RTX 3090 GPU, utilizing PyTorch (Paszke et al., 2019) for code construction. The DFR-Det is built with three mainstream object detectors, *i.e.*, two-stage Faster-RCNN (Ren et al., 2015) and one-stage FCOS (Tian et al., 2019) and RetinaNet (Lin et al., 2017b). We utilize ResNet-18 (He et al., 2016) as the backbone, which is pre-trained on ImageNet (Krizhevsky et al., 2012). Following (Wang et al., 2020), we modify the network of ResNet-18 with an extra shortcut and PReLU (He et al., 2015). We binarize all convolution and FC layers except for the 1-st layer, shortcut layer, and last layers. The decoder network keeps real-valued during training and is not involved in the inference stage. All models are trained using the Stochastic Gradient Descent (SGD) optimizer for 12 epochs with 0.9 momentum, 0.0001 weight decay, and batch size of two. The initial learning rate is set to 0.005 and decays at the 8-th and 11-th epochs.

In the inference stage, we use the preset score of 0.05 to filter out background bounding boxes, and NMS is applied with the IoU threshold of 0.5. We use the evaluation metric from the AI-TOD (Wang et al., 2021) dataset, which includes several metrics such as AP, AP$_{0.5}$, AP$_{0.75}$, AP$_{vt}$, AP$_t$, AP$_s$, and AP$_m$. The AP score is calculated by averaging mAP across different IoU thresholds, where IoU $\in \{0.5, 0.55, \cdots, 0.95\}$. Additionally, AP$_{0.5}$ and

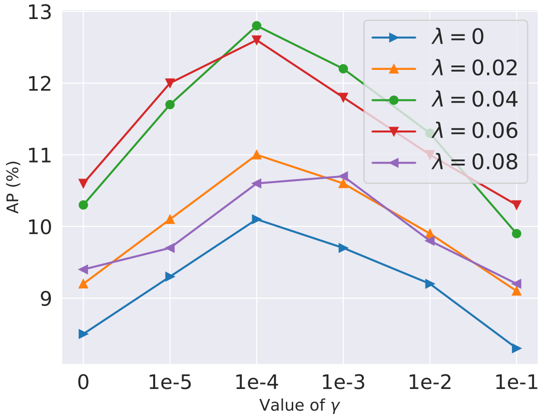


Figure 4. On AI-TOD, we select λ and γ for our DFR-Det method.

AP$_{0.75}$ correspond to the APs at IoU thresholds of 0.5 and 0.75, respectively. We also use AP$_{vt}$, AP$_t$, AP$_s$, and AP$_m$ metrics to evaluate objects of very tiny (2-8 pixels), tiny (8-16 pixels), small (16-32 pixels), and medium (32-64 pixels) scales. Such criterion is also applied to experiments on DOTA-v2.0 (Xia et al., 2018).

4.2. Ablation Study

Hyper-Parameter Selection. As mentioned, we select hyper-parameters γ and λ in this part using the FCOS (Tian et al., 2019) with ResNet-18 backbone. We show the model performance (AP) with different setups of hyper-parameters λ and γ in Fig. 4, where we conduct ablative experiments on the baseline detector. The performances increase first and then decrease with the increase of γ from left to right. Since γ controls the proportion of the applied reconstruction error for the binarization process, we find the 1-bit FCOS achieves the best performance when γ is set as $1e$-4. Thus, we set γ as $1e$-4 for an extended ablation study. For the λ controlling discriminative feature refinement loss, we show that the vanilla baseline ($\lambda = 0$) performs worse than any versions with discriminative feature refinement loss ($\lambda > 0$), showing the proposed loss is necessary. With the varying value of λ, we find $\lambda = 0.04$ boosts the performance of our DFR, achieving 12.8% mAP on AI-TOD using the FCOS detector. Based on the ablative study above, we set hyper-parameter λ as 0.04 for the experiments in this paper.

Effectiveness of Components. We show quantitative improvements of components in the DFR method in Tab. 1. We first reveal the impact of levels of decoded features. On 1-bit FCOS using the $P_2 \sim P_6$ features for detection, we first separately select the single-level features from P_2 to P_6. As seen in the first seven rows, singly decoding the features from P_2 to P_4 positively affects the performance. However, decoding the features of P_5 and P_6 defects the baseline.

Table 1. The effects of different components in DFR-Det with 1-bit FCOS framework on AI-TOD dataset. We evaluate different decoded features and formulation of foreground refinement loss function with $\{\gamma, \lambda\}$ set as $\{1e-4, 0.05\}$.

Method	Decoded feature	Formulation of $\mathcal{L}_{DFR}$	AP
Baseline	✗	✗	10.1
Baseline + DFR	P_2	Eq. (10)	10.9
	P_3		10.6
	P_4		10.2
	P_5		9.8
	P_6		9.1
Baseline + DFR	$\{P_2, P_3\}$	Eq. (10)	12.0
	$\{P_2, P_3, P_5\}$		11.7
	$\{P_2, P_3, P_6\}$		11.1
Baseline + DFR	$\{P_2, P_3, P_4\}$	$\Vert\phi(F) - X\Vert_1$	10.6
	$\{P_2, P_3, P_4\}$	$\Vert\phi(F_f) - X_f\Vert_1$	11.2
	$\{P_2, P_3, P_4\}$	$\Vert\phi(F_b)\Vert_1$	11.7
Baseline + DFR (DFR-Det)	$\boldsymbol{\{P_2, P_3, P_4\}}$	**Eq. (10)**	**12.8**

Then, we evaluate the combination of different levels of features. As can be seen in the following lines, decoding feature group $\{P_2, P_3, P_4\}$ boosts the baseline most, up to 2.7%. This enlightens us on deploying the decoder network from the largest-scaled feature to the P_4 level.

After finding the optimal feature group for discriminative refinement, we further validate the rationality of the proposed discriminative feature refinement loss ($\mathcal{L}_{dfr}$ in Eq. (10)). Intuitively, a straightforward solution is to fully consider the entire image as $\Vert\phi(F) - X\Vert_1$ (equivalent as $\Vert\phi(F_f) - X_f\Vert_1 + \Vert\phi(F_b) - X_b\Vert_1$), which shows less effectiveness with only 0.5% performance improvement gained. This inspires us that an attended region refinement is desired. We further show the situation when only foreground is considered, *i.e.*, $\Vert\phi(F_f) - X_f\Vert_1$. As seen in the last two lines, this equation brings 0.5% performance improvement, defeating the complete imitation. This reveals that foreground reconstruction boosts the network, whereas the background information defects it. Then, we aim to regularize the background information based on $\Vert\phi(F_b)\Vert_1$, which significantly brings a 1.6% performance gain. Inspired by the above phenomenons, we find the proposed discriminative feature refinement loss (Eq. (10)) achieves an outstanding 2.7% performance improvement, which subtly validates our theory and derivations.

4.3. Results on AI-TOD

We first compared our method with other state-of-the-art BNN ReActNet (Liu et al., 2020), as well as the 1-bit detectors BiDet (Wang et al., 2020), LWS-Det (Xu et al., 2021),

Table 2. Main results with various frameworks ResNet-18 backbone on AI-TOD. Note that models are trained on the AI-TOD `trainval` and validated on the AI-TOD `test`. We report APs (%) with different IoU thresholds and APs (%) for objects of various sizes based on the AI-TOD criterion. The **bold** denotes the best result.

Framework	Method	#Bits	Size$_{(\mathrm{MB})}$	OPs$_{(\mathrm{G})}$	AP	AP$_{0.5}$	AP$_{0.75}$	AP$_{vt}$	AP$_{t}$	AP$_{s}$	AP$_{m}$
FCOS	Real-valued	32-32	76.44	318.25	13.5	31.3	9.2	5.0	15.5	16.9	17.2
	LSQ	4-4	21.37	55.88	12.7	31.5	7.9	4.3	14.0	16.2	17.3
	ReActNet	1-1	15.47	23.08	7.9	19.5	4.0	1.3	8.6	9.9	10.1
	BiDet				8.2	20.7	4.6	1.8	9.2	10.9	10.7
	LWS-Det				9.0	23.1	5.0	2.2	10.4	11.6	11.1
	IDa-Det				10.7	27.8	6.2	3.4	11.9	14.0	14.2
	DFR-Det				**12.8**	**32.5**	**7.7**	**4.8**	**13.7**	**16.4**	**17.7**
RetinaNet	Real-valued	32-32	79.73	321.16	11.3	30.1	7.6	3.4	12.9	16.5	20.6
	LSQ	4-4	21.78	56.25	10.4	29.9	7.2	3.2	12.3	15.8	20.4
	ReActNet	1-1	15.57	23.13	6.1	18.9	3.9	1.3	8.0	9.7	11.8
	BiDet				7.1	20.7	4.6	1.7	8.9	10.7	14.9
	LWS-Det				8.4	23.7	5.3	2.1	9.9	12.1	16.3
	IDa-Det				9.1	27.4	6.1	2.7	10.6	14.7	18.8
	DFR-Det				**10.6**	**29.6**	**7.1**	**3.2**	**12.6**	**16.0**	**20.1**
Faster-RCNN	Real-valued	32-32	113.26	102.83	10.2	24.7	6.8	0.0	7.8	20.1	28.5
	LSQ	4-4	22.19	28.96	9.1	21.7	5.7	0.0	7.2	18.7	25.9
	ReActNet	1-1	16.78	19.72	4.8	14.5	3.2	0.0	4.5	10.1	17.9
	BiDet				5.9	15.2	3.6	0.0	4.9	11.2	18.8
	LWS-Det				6.9	17.4	3.9	0.0	5.2	13.7	20.1
	IDa-Det				7.8	19.6	4.6	0.0	5.6	16.0	23.4
	DFR-Det				**8.6**	**20.7**	**5.3**	**0.4**	**6.4**	**17.1**	**24.8**

and IDa-Det (Xu et al., 2022c), using the AI-TOD benchmark (Xu et al., 2022a). We also report the detection result of the 4-bit quantization method LSQ (Esser et al., 2019) for reference. We evaluate the proposed DFR-Det method with various detectors using ResNet-18 (He et al., 2016) backbone and the P2∼P6 feature of FPN (Lin et al., 2017a) for detection. Tab. 2 compares several quantization approaches and detection frameworks regarding computing complexity, storage cost, and the mAP. Our DFR-Det significantly accelerates computation and reduces storage requirements for various detectors. We follow BiDet (Wang et al., 2020) to calculate memory usage, which is calculated by adding 32× the number of full-precision kernels and 1× the number of binary kernels in the networks. The number of float operations (FLOPs) is calculated in the same way as Bi-Real-Net (Liu et al., 2018b). The current CPUs can handle both bit-wise XNOR and bit-count operations in parallel. The number of real-valued FLOPs plus $\frac{1}{64}$ of the number of 1-bit multiplications equals OPs.

FCOS: Compared with the state-of-the-art 1-bit methods, our DFR-Det outperforms other methods by significant margins. As shown in the 2-nd∼9-th rows of Tab. 2, with the FCOS (Tian et al., 2019) framework, our DFR-Det improves the AP by 4.6%, 3.8%, and 2.1% compared with state-of-the-art BiDet, LWS-Det, and IDa-Det, respectively. Compared with the vanilla real-valued counterpart, the proposed DFR-Det achieves comparable performance as real-valued FCOS with apparent computation acceleration and storage savings by 13.79× and 4.94×. The above results are of great significance in the real-time inference of object detection.

RetinaNet: We further validate the effectiveness of DFR-Det on the RetinaNet (Lin et al., 2017b) framework. In the 10-th to 17-th rows of Tab. 2, it is evident that utilizing the RetinaNet framework, our DFR-Det demonstrates improvements in AP by 3.5%, 2.2%, and 1.5% when compared to the state-of-the-art methods BiDet, LWS-Det, and IDa-Det, respectively. Compared to the standard real-valued model, DFR-Det performs on par with the real-valued RetinaNet while significantly accelerating computation and saving storage by 13.88× and 5.12×, respectively.

Faster-RCNN: Finally, we extend DFR-Det to the two-stage Faster-RCNN (Lin et al., 2017b) for effectiveness assessment. In the bottom rows of Tab. 2, it is evident that when integrated into the Faster-RCNN framework, DFR-Det showcases improvements in AP by 2.7%, 1.7%, and 0.8%, surpassing the state-of-the-art methods BiDet, LWS-Det, and IDa-Det, respectively. Compared to the standard real-valued model, DFR-Det achieves comparable performance to real-valued RetinaNet, all while significantly accelerating computation and saving storage by 5.21× and 6.75×.

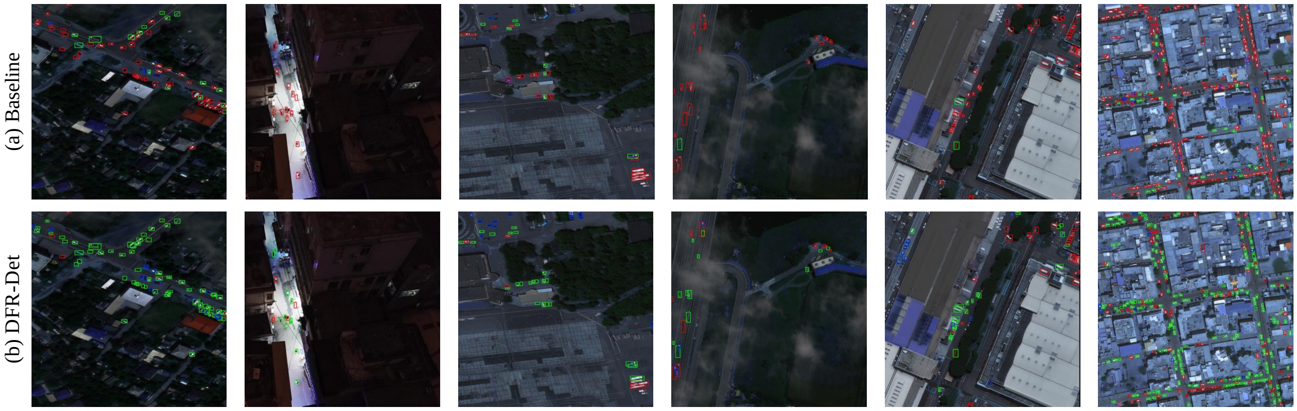


Figure 5. We visualize the detection result of (a) Baseline and (b) DFR-Det. The **green**, **blue**, and **red** boxes denote true positive, false positive, and false negative, respectively.

Table 3. Results with various frameworks using ResNet-18 backbone on DOTA-v2.0. Models are trained on the DOTA-v2.0 `train` and validated on the DOTA-v2.0 `val`. We report APs (%) with different IoU thresholds and APs (%) for objects in various sizes following the DOTA-v2.0 benchmark. The **bold** denotes the best result.

Framework	Method	#Bits	Size$_{(\mathrm{MB})}$	OPs$_{(\mathrm{G})}$	AP	AP$_{0.5}$	AP$_{0.75}$	AP$_{vt}$	AP$_t$	AP$_s$	AP$_m$
FCOS	Real-valued	32-32	76.44	521.34	31.6	55.1	31.4	0.3	4.2	19.6	38.2
	BiDet	1-1	15.47	34.65	23.1	49.1	27.6	0.2	2.8	17.9	33.5
	LWS-Det				25.9	50.3	28.1	0.3	3.4	16.9	35.0
	IDa-Det				28.7	52.1	29.5	0.3	4.1	18.1	36.4
	DFR-Det				**30.9**	**54.8**	**31.0**	**0.6**	**4.6**	**19.4**	**37.6**
Faster-RCNN	Real-valued	32-32	113.26	157.13	33.7	57.4	55.1	0.0	6.4	26.3	41.3
	BiDet	1-1	16.78	32.13	24.6	49.2	26.5	0.0	3.6	20.1	34.3
	LWS-Det				26.7	50.4	27.7	0.0	4.5	21.3	35.9
	IDa-Det				29.1	52.7	29.8	0.0	5.2	22.9	37.5
	DFR-Det				**31.1**	**54.5**	**31.7**	**0.3**	**5.9**	**24.7**	**39.2**

Moreover, visualization results on the AI-TOD dataset are shown in Fig. 5. When applying DFR-Det into 1-bit FCOS, false negative predictions can be significantly eliminated.

4.4. Results on DOTA-V2.0

We further evaluate the proposed DFR-Det on the DOTA-v2.0 (Ding et al., 2021) dataset containing both a large-scale variance and a significant number of tiny objects. We compared our method with other state-of-the-art 1-bit detectors BiDet (Wang et al., 2020), LWS-Det (Xu et al., 2021), and IDa-Det (Xu et al., 2022c). As shown in Tab. 3, on FCOS (Tian et al., 2019) framework, our DFR-Det method surpasses IDa-Det by 2.2% AP point, a significant improvement in TOD task. Also, on Faster-RCNN (Ren et al., 2015), the DFR-Det surpasses IDa-Det by 2.0% AP point. Compared to the real-valued models, DFR-Det exhibits a relatively modest performance gap, with only 0.7% and 2.6% differences on FCOS and Faster-RCNN. Notably, DFR-Det respectively accelerates the detectors by 15.04× and 4.89×, which is significant and meaningful for real-time TOD.

In summary, our approach demonstrates superior performance over state-of-the-art 1-bit detectors in both AP across various IoU thresholds and AP for objects of different sizes on the DOTA-v2.0 dataset. These results highlight the superiority and generalization of DFR-Det across a range of application scenarios.

5. Conclusion

This paper introduces a discriminative feature refinement for 1-bit detectors (DFR-Det) to enhance the discriminative ability of foreground representation for tiny objects in aerial images. Our approach obtains a more distinctive representation of tiny objects within the information bottleneck (IB) principle by mimicking how human beings filter background regions and focus on the differences between the target and the background when perceiving tiny objects. To achieve this, we design a new decoder with a foreground-aware mask to enhance the discriminative ability of high-level features for the target while suppressing the background impact. The proposed method's effectiveness is demonstrated through extensive experiments on three datasets, which show that our DFR-Det achieves new state-of-the-art results.

Acknowledgement

The work was supported by the National Key Research and Development Program of China (Grant No. 2023YFC3300029) and National Natural Science Foundation of China under Grant 623B2016. This research was also supported by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F020007, Beijing Natural Science Foundation L223024, National Natural Science Foundation of China under Grant 62076016, "One Thousand Plan" projects in Jiangxi Province Jxsg2023102268, Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park Grant No.Z231100005923035. Taiyuan City "Double hundred Research action" 2024TYJB0127.

Impact Statement

The binary neural network community may benefit from our research. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Bai, Y., Zhang, Y., Ding, M., and Ghanem, B. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Eur. Conf. Comput. Vis.*, pp. 206–221, 2018.

Cai, H., Chen, T., Zhang, W., Yu, Y., and Wang, J. Efficient architecture search by network transformation. In *AAAI Conf. Art. Intell.*, pp. 2787–2794, 2018.

Choi, J., Jin Chang, H., Yun, S., Fischer, T., Demiris, Y., and Young Choi, J. Attentional correlation filter network for adaptive visual tracking. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4807–4816, 2017.

Courbariaux, M., Bengio, Y., and David, J.-P. Binaryconnect: Training deep neural networks with binary weights during propagations. In *Adv. Neural Inform. Process. Syst.*, pp. 3123–3131, 2015.

Courbariaux, M., Hubara, I., Soudry, D., El-Yaniv, R., and Bengio, Y. Binarized neural networks: Training deep neural networks with weights and activations constrained to+ 1 or-1. *arXiv preprint arXiv:1602.02830*, 2016.

Cui, Z., Zhu, Y., Gu, L., Qi, G.-J., Li, X., Zhang, R., Zhang, Z., and Harada, T. Exploring resolution and degradation clues as self-supervised signal for low quality object detection. In *Eur. Conf. Comput. Vis.*, pp. 473–491, 2022.

Denil, M., Shakibi, B., Dinh, L., Ranzato, M., and De Freitas, N. Predicting parameters in deep learning. In *Adv. Neural Inform. Process. Syst.*, pp. 1–9, 2013.

Ding, J., Xue, N., Xia, G.-S., Bai, X., Yang, W., Yang, M. Y., Belongie, S., Luo, J., Datcu, M., Pelillo, M., et al. Object detection in aerial images: A large-scale benchmark and challenges. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44 (11):7778–7796, 2021.

Esser, S. K., McKinstry, J. L., Bablani, D., Appuswamy, R., and Modha, D. S. Learned step size quantization. In *Int. Conf. Learn. Represent.*, pp. 1–12, 2019.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes (voc) challenge. *Int. Conf. Comput. Vis.*, 88:303–308, 2009.

Farsiu, S., Robinson, M. D., Elad, M., and Milanfar, P. Fast and robust multiframe super resolution. *IEEE Trans. Image Process.*, 13(10):1327–1344, 2004.

Guo, J., Han, K., Wang, Y., Wu, H., Chen, X., Xu, C., and Xu, C. Distilling object detectors via decoupled features. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2154–2164, 2021.

He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Int. Conf. Comput. Vis.*, pp. 1026–1034, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 770–778, 2016.

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Kim, Y., Kang, B.-N., and Kim, D. San: Learning relationship between convolutional features for multi-scale object detection. In *Eur. Conf. Comput. Vis.*, pp. 316–331, 2018.

Kisantal, M., Wojna, Z., Murawski, J., Naruniec, J., and Cho, K. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296*, 2019.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, pp. 1097–1105, 2012.

Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. Pruning filters for efficient convnets. In *Int. Conf. Learn. Represent.*, 2016.

Li, J., Liang, X., Wei, Y., Xu, T., Feng, J., and Yan, S. Perceptual generative adversarial networks for small object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1222–1230, 2017.

Li, Y., Chen, Y., Wang, N., and Zhang, Z. Scale-aware trident networks for object detection. In *Int. Conf. Comput. Vis.*, pp. 6054–6063, 2019.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pp. 740–755, 2014.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2117–2125, 2017a.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Int. Conf. Comput. Vis.*, pp. 2980–2988, 2017b.

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. Path aggregation network for instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 8759–8768, 2018a.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. Ssd: Single shot multibox detector. In *Eur. Conf. Comput. Vis.*, pp. 21–37, 2016.

Liu, Z., Wu, B., Luo, W., Yang, X., Liu, W., and Cheng, K.-T. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Eur. Conf. Comput. Vis.*, pp. 747–763, 2018b.

Liu, Z., Shen, Z., Savvides, M., and Cheng, K.-T. Reactnet: Towards precise binary neural network with generalized activation functions. In *Eur. Conf. Comput. Vis.*, pp. 143–159, 2020.

Ma, N., Zhang, X., Zheng, H.-T., and Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Eur. Conf. Comput. Vis.*, pp. 116–131, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inform. Process. Syst. Wsp.*, pp. 1–12, 2019.

Qiao, S., Chen, L.-C., and Yuille, A. Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 10213–10224, 2021.

Qin, H., Gong, R., Liu, X., Shen, M., Wei, Z., Yu, F., and Song, J. Forward and backward information retention for accurate binary neural networks. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2250–2259, 2020.

Ren, S., He, K., Girshick, R., and Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Adv. Neural Inform. Process. Syst.*, pp. 1–12, 2015.

Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 1874–1883, 2016.

Shwartz-Ziv, R. and Tishby, N. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Singh, B. and Davis, L. S. An analysis of scale invariance in object detection snip. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3578–3587, 2018.

Singh, B., Najibi, M., and Davis, L. S. Sniper: Efficient multi-scale training. *Adv. Neural Inform. Process. Syst.*, pp. 1–11, 2018.

Talebi, H. and Milanfar, P. Learning to resize images for computer vision tasks. In *Int. Conf. Comput. Vis.*, pp. 497–506, 2021.

Tian, Z., Shen, C., Chen, H., and He, T. Fcos: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, pp. 9627–9636, 2019.

Wang, J., Yang, W., Guo, H., Zhang, R., and Xia, G.-S. Tiny object detection in aerial images. In *Int. Conf. Pattern Recog.*, pp. 3791–3798, 2021.

Wang, T., Yuan, L., Zhang, X., and Feng, J. Distilling object detectors with fine-grained feature imitation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4933–4942, 2019.

Wang, Z., Wu, Z., Lu, J., and Zhou, J. Bidet: An efficient binarized object detector. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2049–2058, 2020.

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., and Sherman, A. M. Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, 73:1650–1671, 2011.

Xia, G.-S., Bai, X., Ding, J., Zhu, Z., Belongie, S., Luo, J., Datcu, M., Pelillo, M., and Zhang, L. Dota: A large-scale dataset for object detection in aerial images. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3974–3983, 2018.

Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., and Xia, G.-S. Detecting tiny objects in aerial images: A normalized wasserstein distance and a new benchmark. *ISPRS J. Photo. and Rem. Sen.*, 190:79–93, 2022a.

Xu, S., Zhao, J., Lu, J., Zhang, B., Han, S., and Doermann, D. Layer-wise searching for 1-bit detectors. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5682–5691, 2021.

Xu, S., Li, Y., Wang, T., Ma, T., Zhang, B., Gao, P., Qiao, Y., Lü, J., and Guo, G. Recurrent bilinear optimization for binary neural networks. In *Eur. Conf. Comput. Vis.*, pp. 19–35, 2022b.

Xu, S., Li, Y., Zeng, B., Ma, T., Zhang, B., Cao, X., Gao, P., and Lü, J. Ida-det: An information discrepancy-aware distillation for 1-bit detectors. In *Eur. Conf. Comput. Vis.*, pp. 346–361, 2022c.

Xu, S., Li, Y., Ma, T., Lin, M., Dong, H., Zhang, B., Gao, P., and Lu, J. Resilient binary neural network. In *AAAI Conf. Art. Intell.*, pp. 10620–10628, 2023.

Yu, X., Gong, Y., Jiang, N., Ye, Q., and Han, Z. Scale match for tiny person detection. In *Wint. Conf. Apps. Comput. Vis.*, pp. 1257–1265, 2020.

Zhao, J., Xu, S., Zhang, B., Gu, J., Doermann, D., and Guo, G. Towards compact 1-bit cnns via bayesian learning. *Int. Conf. Comput. Vis.*, 130(2):201–225, 2022.

Zhao, Q., Sheng, T., Wang, Y., Tang, Z., Chen, Y., Cai, L., and Ling, H. M2det: A single-shot object detector based on multi-level feature pyramid network. In *AAAI Conf. Art. Intell.*, pp. 9259–9266, 2019.