Learning Calibrated-Guidance for Object Detection in Aerial Images

Dong Liang[1,†], Zongqi Wei[1,†], Dong Zhang[2,†], Qixiang Geng[1,†],
Liyan Zhang[1], Han Sun[1], Huiyu Zhou[3], Mingqiang Wei[1], Pan Gao[1]
[1]Nanjing University of Aeronautics and Astronautics.
[2]Nanjing University of Science and Technology. [3]University of Leicester.
{liangdong; weizongqi; gengqx; zhangliyan; sunhan; mqwei; pan.gao}@nuaa.edu.cn
dongzhang@njust.edu.cn; hz143@leicester.ac.uk

Abstract

*Recently, the study on object detection in aerial images has made tremendous progress in the community of computer vision. However, most state-of-the-art methods tend to develop elaborate attention mechanisms for the space-time feature calibrations with high computational complexity, while surprisingly ignoring the importance of feature calibrations in channels. In this work, we propose a simple yet effective Calibrated-Guidance (CG) scheme to enhance channel communications in a feature transformer fashion, which can adaptively determine the calibration weights for each channel based on the global feature affinity-pairs. Specifically, given a set of feature maps, CG first computes the feature similarity between each channel and the remaining channels as the intermediary calibration guidance. Then, re-representing each channel by aggregating all the channels weighted together via the guidance. Our CG can be plugged into any deep neural network, which is named as CG-Net. To demonstrate its effectiveness and efficiency, extensive experiments are carried out on both oriented and horizontal object detection tasks of aerial images. Results on two challenging benchmarks (*i.e.*, DOTA and HRSC2016) demonstrate that our CG-Net can achieve state-of-the-art performance in accuracy with a fair computational overhead. Code has been open sourced*[1]

1. Introduction

Object detection in aerial images is a fundamental yet challenging task, which aims to assign a bounding box with a unique semantic category label to each surficial object in the given aerial images [8, 27, 49, 50, 46]. Thanks to the recent promising development of deep Convolutional Neural Networks (CNNs) in image processing, object detection in aerial images has also made tremendous progress. The state-of-the-art methods are usually based on a one-stage detection (*e.g.*, RetinaNet [21] and YOLO [34]) or a two-stage detection (*e.g.*, Fast/Faster R-CNN [11, 35]) head-network with a CNN as the backbone.

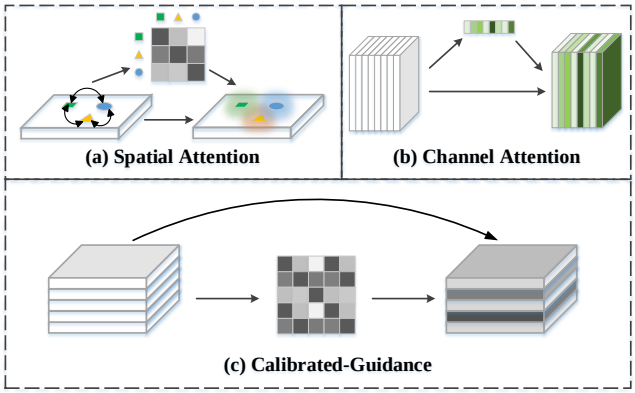

Figure 1. Illustrations of the attention-based feature calibration methods, (a) spatial attention, (b) channel attention, and (c) our proposed Calibrated-Guidance scheme.

Compared to objects in general images, objects in aerial images usually have a smaller size, a higher density, worse imaging quality, and more complex background [43, 18]. Therefore, it is difficult to directly achieve a satisfying outcome on aerial images using the existing natural scene object detectors. State-of-the-art methods focus on developing effective head networks [8], adaptive dense anchor [27], and labeling strategy [49, 46]. Besides, effective feature learning plays a crucial role. To this end, a large amount of feature calibration methods based on attention mechanisms have been proposed to improve the rough feature representations of CNNs [50, 12, 4, 39, 38]. Conceptually, these attention-based methods can be basically divided into two categories: I) the spatial-attention-based one, and II) the channel-attention-based one. For **the first category** (*e.g.*,

[1]https://github.com/WeiZongqi/CG-Net
†These authors contributed equally to this work.

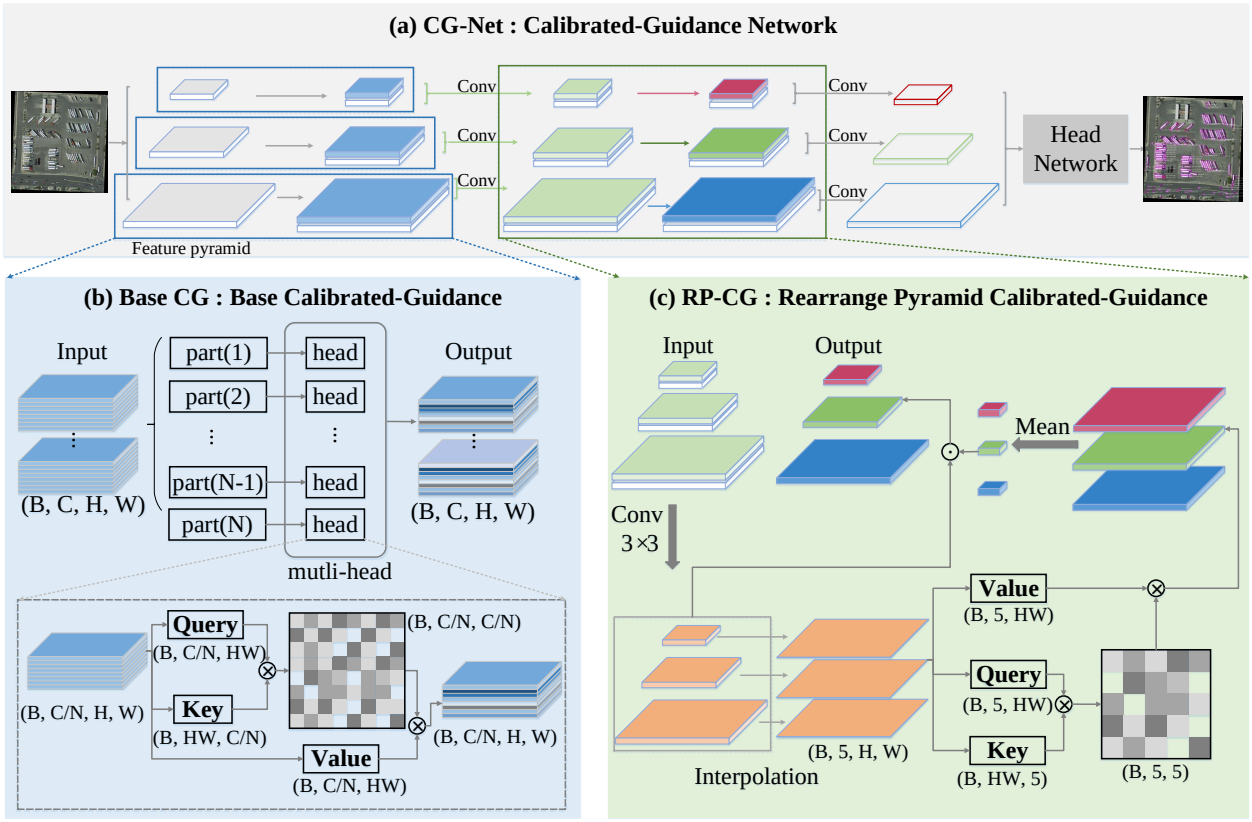

Figure 2. Overall architecture of our proposed Calibrated-Guidance Network (CG-Net), where CG is deployed on both the intra-layer feature maps and the feature pyramid (*i.e.*, feature pyramid network [20]). After that, we use a task-specific head network for dealing with both oriented and horizontal object detection tasks of aerial images. ResNet [14] is used as the backbone network.

spatial attention module [9, 4, 38], recurrent attention structure [39], self-attention mechanism [3], and non-local operation [40]), as shown in Figure 1 (a), a global context mapping for each position can be obtained by computing the similarities between the feature of each position and all the remaining features [52]. Through such an operation, each pixel can obtain the long-range dependence information. For **the second category** (*e.g.*, channel attention module [9], channel-wise attention [5, 4], and squeeze-and-excitation block [16, 12, 50]), as shown in Figure 1 (b), each channel can obtain a weight that reflects its own importance in detection, and then integrate the weight into the model by way of channel re-weighting.

Despite the success of the existing attention-based methods in calibrating features for object detection, we argue that most of these methods on feature calibrations in channels are not enough. That is to say, they cannot introduce channel communications to capture the dependencies between channel feature maps, which has empirically shown their benefits to a wide range of visual recognition tasks [56, 42, 15, 13, 33, 51]. Although the existing channel-attention-based methods can enable different channels to obtain different weights, modules (*e.g.*, global average/max pooling) based on their channel feature maps cannot guarantee all the channels have sufficient communications. Therefore, from the perspective of channel communications, these methods are still local-based.

To address the above problem, in this paper, we propose a simple yet effective Calibrated-Guidance (CG) scheme to enhance channel communications in a feature transformer fashion, which can adaptively determine the calibration weights for each channel based on the global feature affinity-pairs. CG is an active feature communication mechanism, as illustrated in Figure 1 (c), which can explicitly introduces feature dependencies in channels. Specifically, for a given set of aerial image feature maps, CG consists of two steps: first, feature similarities (via the dot product operation) between each channel and the remaining channels are computed as the intermediary calibration guidance. Then, we represent each channel by aggregating all the channels weighted together via the guidance. Particularly, the standard CG has low computational complexity, $O(NC^2)$ both in time and space (while the spatial-attention-based ones are of $O(NH^2W^2)$), where H, W, and C denotes the height, width, and channel dimension of the input feature maps; N is the size of the multi-head structure.

CG is a general unit that can be plugged into any deep neural network, named as CG-Net. The overall architec-

ture is shown in Figure 2. To demonstrate its effectiveness and efficiency, we conduct extensive experiments on both oriented and horizontal object detection tasks. Experimental results on the challenging benchmarks DOTA [43] and HRSC2016 [25] for oriented object detection show that our proposed CG-Net boosts substantial improvements compared to the baseline methods and achieves the state-of-the-art performance in accuracy (*i.e.*, 77.89% and 90.58% mAP, respectively) with a fair computational overhead. Besides, experimental results on DOTA [43] for horizontal object detection also validate the flexibility and effectiveness of the proposed CG-Net, which also achieves state-of-the-art performance with the accuracy by 78.26% mAP.

In summary, our main contributions are two-fold: 1) a simple yet effective CG scheme is proposed to enhance channel communications in a feature transformer fashion; 2) we propose a CG-Net, which can achieve the state-of-the-art oriented and horizontal object detection performance on two challenging benchmarks for aerial images, including DOTA and HRSC2016.

2. Related Work

Object detection in aerial images. Object detection in aerial images can be divided into horizontal and oriented object detection tasks. Horizontal object detection aims to detect objects with horizontal bounding boxes [35, 6, 21, 34]. Being observed from an overhead perspective, the objects in aerial images present more diversified orientations. Oriented object detection [23, 8, 45, 44, 31] is an extension of horizontal object detection to accurately outline the objects, especially those with large aspect ratios. In this work, we focus on both oriented and horizontal object detection tasks for aerial images.

Feature calibration over images. Currently, most of the state-of-the-art methods are designed from the perspective of feature calibration to deal with the challenges of complex background and noise in object detection [9, 40, 16, 4, 38]. Among those methods, attention-based ones are proposed to calibrate features from two aspects, including spatial-attention and channel-attention-based.

Spatial-attention-based mechanisms capture object positions in the spatial dimension. Position attention module [9] / Non-local operation [40] build rich contextual relationships on local features by using a self-attention mechanism. Transformer [37] is the first sequence transduction model combined with multi-headed self-attention. DETR [3] is proposed to explore the relationship between objects in the global context, which is of precision similar to those of the two-stage detectors, but has a weakness on detecting large objects with high computational overheads. In aerial image analysis, ARCNet [39] utilizes a recurrent attention structure to squeeze high-level semantic features for learning to reduce parameters.

Channel-attention-based mechanisms allocate resources for channels referring to their importance. SENet [16] utilizes a squeeze-and-excitation block to implement dynamic channel-wise feature re-calibration. For obtaining better feature representations, DANet [9] utilizes a channel attention module to capture contextual relationships based on the self-attention mechanism. In aerial image, a residual-based network combining channel attention [12] is used to learn the most relevant high-frequency features.

There are also some works that combine spatial attention with channel-wise attention together, *e.g.*, SCA-CNN [5], and DONet [55]. Despite the success of the existing attention-based methods, they are not sufficient for feature calibration in channels. In this work, we propose a simple yet effective CG scheme to enhance channel communications in a feature transformer fashion, which can adaptively determine the calibration weights for each channel based on the global feature affinity-pairs.

3. Methodology

In this section, we show the technical details of our proposed Calibrated-Guidance Network (CG-Net) for object detection in aerial images. Specifically, we first revisit the channel attention mechanism on images in Section 3.1. Then, our proposed Calibrated-Guidance (CG) module which can enhance channel communications is described in Section 3.2. After that, we introduce how to implement CG on the base CNNs' feature maps (*i.e.*, Base CG) and on an intra-network feature pyramid (*i.e.*, Rearranged Pyramid CG) for object detection in Section 3.3. Finally, we show the details of the network architecture in Section 3.4.

3.1. Channel attention revisited

A Channel Attention (CA) module utilizes the interdependencies between the channels to emphasize the important ones by weighting the similarity matrix. To be specific, CA operates on queries ($\mathbf{Q}$), keys ($\mathbf{K}$) and values ($\mathbf{V}$) among a set of single-scale feature maps $\mathbf{X}$, and the improved version $\mathbf{X}'$ has the same scale as the original $\mathbf{X}$. For a given set of feature maps $\mathbf{X} \in \mathbb{R}^{W\times H\times C}$, where W, H and C are width, height and channel dimension, respectively, CA implementation can be formulated as:

$$\begin{aligned} \textbf{Input} &: \mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j \\ \textbf{Similarity} &: \mathbf{s}_{i,j} = F_{\text{sim}}(\mathbf{q}_i, \mathbf{k}_j) \\ \textbf{Weight} &: \mathbf{w}_{i,j} = F_{\text{nom}}(\mathbf{s}_{i,j}) \\ \textbf{Output} &: \mathbf{X}'_i = F_{\text{mul}}(\mathbf{w}_{i,j}, \mathbf{v}_j), \end{aligned} \tag{1}$$

where $\mathbf{q}_i = f_q(\mathbf{X}_i) \in \mathbf{Q}$ is the i^{th} query; $\mathbf{k}_j = f_k(\mathbf{X}_j) \in \mathbf{K}$ and $\mathbf{v}_j = f_v(\mathbf{X}_j) \in \mathbf{V}$ are the j^{th} key/value pair; $f_q(\cdot)$, $f_k(\cdot)$ and $f_v(\cdot)$ denote the query/key/value channel transformer functions [37, 3]; $\mathbf{X}_i$ and $\mathbf{X}_j$ denote the i^{th} and j^{th}

channel feature in $\mathbf{X}$; F_{sim} is the dot product similarity function; F_{nom} is the softmax normalization function; F_{mul} denotes matrix dot multiplication; $\mathbf{X}'_i$ is the i^{th} channel feature in the transformed feature map $\mathbf{X}'$. Although CA can enable different channels to obtain different weights, the coarse operation based on the entire channel feature maps (*i.e.*, without the grouped feature representations [56, 37, 33, 51]) cannot enable all the channels to have sufficient communications. As a result, the ability to feature representation is limited.

3.2. Calibrated-Guidance (CG)

We propose CG to enhance channel communications in a feature transformer fashion, which can adaptively determine the calibration weights for the channels based on the global feature affinity-pairs. Its detailed structure is illustrated in Figure 2. CG is inspired by the transformer mechanism and the difference is that we combine the multi-head representations, and concatenate the original and the calibrated features, then use convolution to produce the enhanced feature maps as output.

Multi-head CG. We deploy the multi-head architecture to focus on richer channel feature representations. First, we divide query and key into N parts in the channel dimension. Then, we feed the divided feature with shape $(B, C/N, H, W)$ into each head, where each structure is a CG module (B is batch size). For n^{th} head module, the shape of similarity matrix $\mathbf{s}^n$ is $(B, C/N, C/N)$, which can be expressed as:

$$\mathbf{s}^n = \begin{bmatrix} w^{1,1} & \dots & w^{C/N,1} \\ \vdots & \ddots & \vdots \\ w^{1,C/N} & \dots & w^{C/N,C/N} \end{bmatrix}, \tag{2}$$

where each w denotes the learnable similarity scalar. After that, the outputs of these head modules (*i.e.*, the partial result) are concatenated together to produce the holistic output feature maps, which have the same shape as the original feature maps. The above process can be formulated as:

$$\begin{aligned} \textbf{Weight} &: \mathbf{w}^n_{i,j} = F_{\text{nom}}(\mathbf{s}^n_{i,j}) \\ \textbf{Partial Result} &: \mathbf{X}^n_i = F_{\text{mul}}(\mathbf{w}^n_{i,j}, \mathbf{v}_{j,n}) \\ \textbf{Holistic Output} &: \mathbf{X}' = F_{\text{con}}(\mathbf{X}^n_i), \end{aligned} \tag{3}$$

where $\mathbf{s}^n_{i,j}$ and $\mathbf{w}^n_{i,j}$ denote the n^{th} partial similarity weight and the normalized one. $\mathbf{v}_{j,n}$ denotes the j^{th} value of the n^{th} head. F_{con} is used for feature concatenation in the channel dimension. Compared to the previous transformer-based approaches, the multi-head CG has lower computational complexity, $O(NC^2)$ both in time and space, while the previous ones have the computational complexity of $O(NH^2W^2)$.

Compared to CA, our proposed CG has the following three advantages: i) CG is designed for the enhancement of channel communications, while most of the previous ones are used to capture the long-range dependencies in space. ii) CG is based on the multi-head structure, which has its unique tendency of feature representation in different feature spaces [37, 2]. Hence CG can provide an enhanced feature representation. iii) CG is designed for object detection in aerial images. Experimental results (in Section 4.3) show that CG can improve the state-of-the-art performance swimmingly on both oriented and horizontal tasks.

3.3. Pipeline

Base CG. Given an arbitrary aerial image, we can extract a set of feature maps by a fully convolution network. For these feature maps, CG can directly achieve calibrated-guidance practice to enhance channel communications and adaptively determine the calibration weight for each channel. Its detailed architecture in a level of the feature pyramid (*i.e.*, feature maps with the same scale) is illustrated in Figure 2 (b). Since this CG implementation is performed on the basic feature maps, we call it Base CG. Base CG is a general unit, which works on the backbone network. Compared to other existing head-network-based task-specific methods [50, 22], it is more universal and can facilitate a wide range of downstream recognition tasks. Our Base CG improves feature extraction, and the results can be seen from the ablation experiments shown in Section 4.2.

Rearranged Pyramid CG. Feature pyramid has shown its effectiveness in a wide range of computer vision tasks [20, 21, 53]. In this section, we show how to implement our Calibrated-Guidance on a feature pyramid (*i.e.*, the proposed Rearranged Pyramid CG (RP-CG)). Compared to the existing feature calibration methods on the in-network feature pyramid [30, 10, 32], our RP-CG has lower computational complexity and fewer model parameters (details are shown in Section 4.1). The RP-CG module works on an extracted feature pyramid from the feature pyramid network [20], whose architecture is illustrated in Figure 2 (c).

From the perspective of levels inside the feature pyramid, each level can been seen as local features, *i.e.*, only part of the features of the input image are captured. In order to emphasize the most suitable feature in the channel dimension of the feature pyramid, combining global and local information is crucial in feature extraction. In our work, RP-CG focuses on weighting different features among pyramid levels $\mathbf{X}_{P2-P6}$. As illustrated in Figure 2 (c), we apply CG between the levels of the feature pyramid to communicate levels' information. In our implementation, firstly, we reduce the channel dimension and launch interpolation on pyramid features $\mathbf{X}_{P2-P6}$ to generate the same scale features (same scale as the largest one: $P2$) and then concatenate them as

$\overline{\mathbf{X}}_{P2-P6}$, which is expressed as:

$$\overline{\mathbf{X}}_{P2-P6} = F_{\text{intp}}(\mathbf{X}_{P2-P6}), \tag{4}$$

where F_{intp} is a channel dimension reduction and scale interpolation function. The shape of output feature $\overline{\mathbf{X}}_{P2-P6}$ is $(B, 5, H_{p2}, W_{p2})$. Then, same as Base CG, RP-CG produces the output $\overline{\mathbf{X}}_i^{'}$ from input $\mathbf{q}_i, \mathbf{k}_j$ and $\mathbf{v}_j$ by learning the weight between the query and the key. The interaction is formulated as:

$$\begin{aligned}
\textbf{Input} &: \mathbf{X}_{P2-P6} \\
\textbf{Interpolation} &: \overline{\mathbf{X}}_{P2-P6} \\
\textbf{Extraction} &: \mathbf{q}_i, \mathbf{k}_j, \mathbf{v}_j \\
\textbf{Similarity} &: \mathbf{s}_{i,j} = F_{\text{sim}}(\mathbf{q}_i, \mathbf{k}_j) \\
\textbf{Weight} &: \mathbf{w}_{i,j} = F_{\text{nom}}(\mathbf{s}_{i,j}) \\
\textbf{Output} &: \overline{\mathbf{X}}_i^{'} = F_{\text{mul}}(\mathbf{w}_{i,j}, \mathbf{v}_j) \\
\textbf{Holistic Output} &: \overline{\mathbf{X}}_{P2-P6}^{rpcg} = F_{\text{con}}(\overline{\mathbf{X}}_i^{'}),
\end{aligned} \tag{5}$$

where $\overline{\mathbf{X}}_i^{'}$ is the i^{th} level feature in transformed feature map $\overline{\mathbf{X}}_{P2-P6}^{rpcg}$ with shape $(B, 5, H_{p2}, W_{p2})$. $\overline{\mathbf{X}}_{P2-P6}^{rpcg}$ realizes global channel communication in pyramid features, but we need to find the right way to feed back to pyramid features.

In addition, there have been multitudes of methods to verify the effectiveness of the combination of global and local information in visual recognition, and our method is global in essence. To this end, combining our RP-CG with the existing local channel attention method is a natural choice. In this work, the classical channel attention [16] is chosen. Based on this, the overall structure of our proposed Rearrange Pyramid Calibrated-Guidance module can be expressed as:

$$\begin{aligned}
\textbf{Weight} &: \overline{\mathbf{X}}_{P2-P6}^{\text{mean}(rpcg)} = F_{mean}(\overline{\mathbf{X}}_{P2-P6}^{rpcg}) \\
\textbf{Scale} &: \overline{\mathbf{X}}_{P2-P6}^{'} = \overline{\mathbf{X}}_{P2-P6}^{mean(rpcg)} \otimes \mathbf{X}_{P2-P6} \\
\textbf{Output} &: \overline{\mathbf{X}}_{P2-P6}^{final} = F_{conv}(\overline{\mathbf{X}}_{P2-P6}^{'} \oplus \mathbf{X}_{P2-P6}).
\end{aligned} \tag{6}$$

The output from $\overline{\mathbf{X}}_{P2-P6}$ are divided into 5 parts ($P2 - P6$). $\overline{\mathbf{X}}_{P2-P6}^{rpcg}$ is the overall feature after we have weighted $\overline{\mathbf{X}}_{P2-P6}$. We use F_{mean} to derive the weighting parameter to distinguish different scales' features, and it includes the operation of using the mean value as the weighting parameter for each pyramid's levels, which is then resized to the same scale of the original level feature. $\otimes$ is matrix cross multiplication, and $\oplus$ is channel concatenation. $\overline{\mathbf{X}}_{P2-P6}^{'}$ is the calibrated feature with the same size as the original feature pyramid. We get final output $\overline{\mathbf{X}}_{P2-P6}^{final}$ from convolution F_{conv}, which is to reduce the channel to the original size.

3.4. Network architecture

We build a Calibrated-Guidance network (CG-Net) for both oriented and horizontal object detection tasks of aerial images. The overall architecture is illustrated in Figure 2. CG-Net is based on our proposed **Base CG** (in Figure 2 (b)) and **RP-CG** (in Figure 2 (c)) for transforming pyramid features. Specifically, we deploy ResNet [14] as backbone following [8], which has been pre-trained on the ImageNet [7]. Then, we produce a feature pyramid from the feature pyramid network [20]. For this feature pyramid, we firstly apply **Base CG** in the feature maps from each level of the pyramid. After that, we deploy the **RP-CG** to produce a new feature pyramid that realizes global and local communication in the feature pyramid. Then, we concatenate the original feature maps with the calibrated ones together in the channel dimension and reduce the dimensionality of the concatenated feature maps into 256 channels by a 3×3 convolution. Finally, we use the head network from the RoI transformer [8] for oriented object detection and a standard Faster R-CNN [35] for horizontal object detection.

4. Experiments

4.1. Settings

Datasets. ***DOTA*** [43] is one of the largest datasets for object detection in aerial images with both oriented and horizontal bounding box annotations. It contains $2,806$ aerial images with $188,282$ annotated instances from different sensors and platforms. The image size ranges from around 800×800 to $4,000 \times 4,000$ pixels and contains objects exhibiting in a wide variety of scales, orientations, and shapes. DOTA contains 15 common object categories, including Plane (PL), Baseball diamond (BD), Bridge (BR), Ground track field (GTF), Small vehicle (SV), Large vehicle (LV), Ship (SH), Tennis court (TC), Basketball court (BC), Storage tank (ST), Soccer-ball field (SBF), Roundabout (RA), Harbor (HA), Swimming pool (SP), and Helicopter (HC). In our experiments, following [43, 50], $3/6$ of the original images are randomly selected as the training set, $1/6$ as the validation set, and $2/6$ as the testing set. ***HRSC2016*** [25] is a challenging dataset for ship detection in aerial images with large aspect ratios and arbitrary orientations, which contains 1061 images and more than 20 categories of ships in various appearances. These images are collected from Google Earth. The image size ranges from 300×300 to 1500×900. In our work, following [25], the training, validation, and test sets include 436 images, 181 images, and 444 images, respectively. For HRSC2016, only oriented object detection can be carried out.

Image size. For DOTA and HRSC2016, we generate a series of $1,024 \times 1,024$ patches from the original images with a stride of 824 for training, validation, and test sets.

Backbone	+Ours	GFLOPs / FPS	#Params (M)	mAP (%)
ResNet-50		211.30 / 25.0	41.20	73.26
	✓	366.32 / 14.8	42.99	74.21$_{+0.95}$
ResNet-101		289.26 / 20.9	60.19	73.06
	✓	444.21 / 13.4	61.98	**74.30**$_{+1.24}$
ResNet-152		367.23 / 17.0	75.83	72.78
	✓	522.19 / 12.1	77.63	73.53$_{+0.75}$

Table 1. The effectiveness of our proposed methods with different backbone networks on the test set of DOTA [43] for oriented object detection. "+ Ours" indicates the implementation of our proposed Base CG and RP-CG units on the backbone networks.

Baseline setup. We use the standard two-stage detector Faster R-CNN [35] as the baseline. It utilizes ResNet-101 as the backbone. FPN [20] is adopted to construct a feature pyramid. Predefined horizontal anchors are set on each feature level, *i.e.*, P2 - P6. Here, we do not use any rotation anchor. For oriented object detection, we add the rotated head developed in RoI-Transformer [8] which transforms the horizontal proposals to the rotated ones. For a fair comparison, all the experimental data and parameter settings are strictly consistent as those reported in [8, 43, 25].

Hyper-parameters. For hyper-parameters, following [8, 27], in DOTA and HRSC2016, only three horizontal anchors are set with aspect ratios of $\{1/2, 1, 2\}$, the base anchor scale is set as $\{8^2\}$, and the anchor strides of each level of the feature pyramid are set as $\{4, 8, 16, 32, 64\}$. In order to verify the effectiveness of our method, we performed ablation studies on DOTA dataset, and we avoid utilizing any bells-and-whistles training strategy and data augmentation in the ablation study. But for the peer comparison on DOTA and HRSC2016, like [8, 50, 27], we only conduct a rotation augmentation randomly from 4 angles $(0, 90, 180, 270)$.

Implementation details. In our work, SGD is used as the optimizer in the training stage. The initial learning rate is set to 0.005 and is divided by 10 at each decay step. Weight decay and momentum are set to 0.0001 and 0.9, respectively. Following [43, 25], the total iterations of DOTA and HRSC2016 are 80k and 20k, respectively. We train the models on RTX 2080Ti with a batch size of 1.

Evaluation and metrics. Following [43], the standard mean Average Precision (mAP) is used as the primary evaluation metric in accuracy. Moreover, to verify the model efficiency, the model Parameters (#Params), and GFLOPs / FPS are also taken into consideration. The results of DOTA reported in our work are obtained by submitting our predictions to the official DOTA evaluation server[2].

4.2. Ablation study

Our ablation study is carried out on DOTA [43] for oriented object detection with ResNet-101 [14], which aims to: 1) verify the effectiveness of our method on different backbone networks; 2) verify the effectiveness of the two proposed units on base CNN feature maps (*i.e.*, Base CG) and a feature pyramid (*i.e.*, RP-CG); 3) reveal mismatching error rates on different scales; and 4) show some visual comparisons.

[2]https://captain-whu.github.io/DOTA/

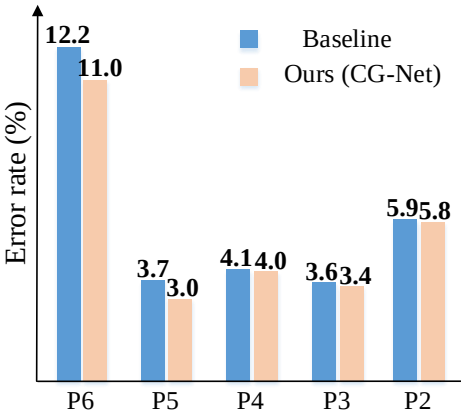

Figure 3. Mismatching error rate comparison between the baseline and the proposed method on DOTA [43] with ResNet-101 [14] for oriented object detection. The lower the better.

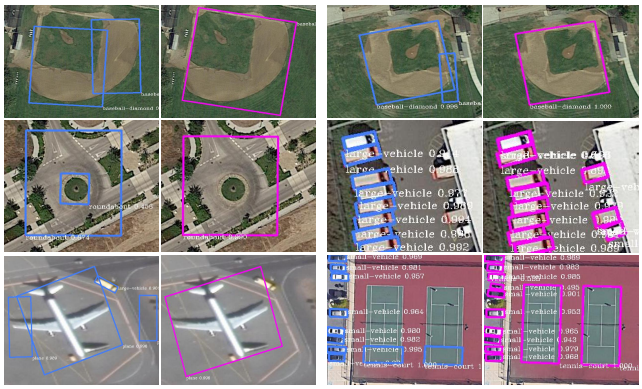

Figure 4. Comparison to the baseline on DOTA [43] for oriented object detection with ResNet-101 [14]. The figures with blue boxes are the results of the baseline and pink boxes are the results of our proposed CG-Net.

Effectiveness on different backbones. In Table 1, we show the experimental results of different backbone networks with our proposed units on the test set of DOTA. We contrast mAP and its improvement from the combination of our proposed module with ResNet-50, ResNet-101, and ResNet-152, respectively. We can observe that adding our proposed module to the backbone increases mAP by 0.95%, 1.24%, and 0.75%. Besides, we also report the model #Params and GFLOPs / FPS for comparisons of model efficiency. Using Base CG and RP-CG increases computational costs; for example, average increases on these three backbones are around 1.80 M model #Params with around 155 GFLOPs, and with around 5-10 FPS reduction. Consider-

Baseline	Base CG	RP-CG	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	GFLOPs / FPS	#Params (M)	mAP (%)
✓			88.53	77.70	51.59	68.80	74.02	76.85	86.98	90.24	84.89	77.68	53.91	63.56	**75.88**	69.48	55.50	289.26 / 20.9	60.19	73.06
✓	✓		88.38	81.09	53.43	68.35	74.23	76.95	86.55	90.67	84.79	**78.17**	54.20	65.23	75.84	69.01	57.64	340.79 / 19.0	60.78	73.64$_{+0.58}$
✓		✓	88.49	77.18	**54.02**	68.85	74.49	76.67	**87.09**	**90.79**	86.17	77.58	54.54	65.03	75.87	69.11	56.94	341.15 / 17.2	60.80	73.52$_{+0.46}$
✓	✓	✓	**88.65**	**82.75**	53.02	**69.65**	**74.77**	**77.48**	86.99	90.32	**86.38**	77.23	**57.05**	**66.33**	75.46	**70.22**	**58.23**	444.21 / 13.4	61.98	**74.30**$_{+1.24}$

Table 2. Ablation study on the test set of DOTA [43] for oriented object detection. ResNet-101 [14] is the backbone.

Methods	Backbone	PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC	mAP (%)
Oriented object detection																	
FR-O [43] (CVPR 2018)	R-101	79.09	69.12	17.17	63.49	34.20	37.16	36.20	89.19	69.60	58.96	49.40	52.52	46.69	44.80	46.30	52.93
R-DFPN [48] (ISO4 2018)	R-101	80.92	65.82	33.77	58.94	55.77	50.94	54.78	90.33	66.34	68.66	48.73	51.76	55.10	51.32	35.88	57.94
R^{2}CNN [17] (preprint 2017)	R-101	80.94	65.67	35.34	67.44	59.92	50.91	55.81	90.67	66.92	72.39	55.06	52.23	55.14	53.35	48.22	60.67
RRPN [26] (TMM 2018)	R-101	88.52	71.20	31.66	59.30	51.85	56.19	57.25	90.81	72.84	67.38	56.69	52.84	53.08	51.94	53.58	61.01
ICN [1] (ACCV 2018)	R-101	81.36	74.30	47.70	70.32	64.89	67.82	69.98	90.76	79.06	78.02	53.64	62.90	67.02	64.17	50.23	68.16
RoI Trans [8] (CVPR 2019)	R-101	88.64	78.52	43.44	**75.92**	68.81	73.68	83.59	90.74	77.27	81.46	58.39	53.54	62.83	58.93	47.67	69.56
CAD-Net [54] (TGRS 2019)	R-101	87.80	82.40	49.40	73.50	71.10	63.50	76.70	**90.90**	79.20	73.30	48.40	60.90	62.00	67.00	62.20	69.90
DRN [29] (CVPR 2020)	H-104	88.91	80.22	43.52	63.35	73.48	70.69	84.94	90.14	83.85	84.11	50.12	58.41	67.62	68.60	52.50	70.70
O^2-DNet [41] (ISPRS 2020)	H-104	89.31	82.14	47.33	61.21	71.32	74.03	78.62	90.76	82.23	81.36	60.93	60.17	58.21	66.98	61.03	71.04
SCRDet [50] (ICCV 2019)	R-101	89.98	80.65	52.09	68.36	68.36	60.32	72.41	90.85	**87.94**	86.86	65.02	66.68	66.25	68.24	65.21	72.61
R^{3}Det [47] (preprint 2019)	R-152	89.49	81.17	50.53	66.10	70.92	78.66	78.21	90.81	85.26	84.23	61.81	63.77	68.16	69.83	67.17	73.74
CSL [49] (ECCV 2020)	R-152	**90.25**	**85.53**	54.64	75.31	70.44	73.51	77.62	90.84	86.15	86.69	**69.60**	**68.04**	73.83	71.10	68.93	76.17
DAL [27] (AAAI 2021)	R-50	89.69	83.11	55.03	71.00	78.30	81.90	**88.46**	90.89	84.97	**87.46**	64.41	65.65	76.86	72.09	64.35	76.95
R^{3}Det-DCL [46] (CVPR 2021)	R-152	89.26	83.60	53.54	72.76	**79.04**	82.56	87.31	90.67	86.59	86.98	67.49	66.88	73.29	70.56	69.99	77.37
Ours (CG-Net)	R-101	88.75	85.18	**57.41**	71.88	73.23	**82.68**	88.14	**90.90**	86.00	85.37	62.99	66.74	**77.98**	**79.90**	**71.27**	**77.89**$_{+0.52}$
Horizontal object detection																	
SSD [24] (ECCV 2016)	R-101	44.74	11.21	6.22	6.91	2.00	10.24	11.34	15.59	12.56	17.94	14.73	4.55	4.55	0.53	1.01	10.94
YOLOv2 [34] (CVPR 2016)	R-101	76.90	33.87	22.73	34.88	38.73	32.02	52.37	61.65	48.54	33.91	29.27	36.83	36.44	38.26	11.61	39.20
R-FCN [6] (NIPS 2016)	R-101	79.33	44.26	36.58	53.53	39.38	34.15	47.29	45.66	47.74	65.84	37.92	44.23	47.23	50.64	34.90	47.24
FR-H [43] (CVPR 2018)	R-101	80.32	77.55	32.86	68.13	53.66	52.49	50.04	90.41	75.05	59.59	57.00	49.81	61.69	56.46	41.85	60.46
FPN [20] (CVPR 2017)	R-101	88.70	75.10	52.60	59.20	69.40	78.80	84.50	90.60	81.30	82.60	52.50	62.10	76.60	66.30	60.10	72.00
ICN [1] (ACCV 2018)	R-101	90.00	77.70	53.40	**73.30**	73.50	65.00	78.20	90.80	79.10	84.80	57.20	62.10	73.50	70.20	58.10	72.50
SCRDet [50] (ICCV 2019)	R-101	**90.18**	81.88	55.30	73.29	72.09	77.65	78.06	**90.91**	82.44	**86.39**	**64.53**	63.45	75.77	78.21	60.11	75.35
Ours (CG-Net)	R-101	88.76	**85.36**	**60.81**	71.88	**74.04**	**83.43**	**88.29**	90.89	**86.00**	85.55	63.51	**67.02**	**78.78**	**80.86**	**68.70**	**78.26**$_{+2.91}$

Table 3. Result comparisons with state-of-the-art methods on the test set of DOTA [43] for both oriented and horizontal object detection in aerial images. By "Ours" we mean that implementing Base CG and RP-CG on the baseline model at the same time. "R-" in the Backbone column denotes the ResNet [14], and "H-" denotes the Hourglass network [28].

ing the model performance and the amount of calculation, in the following experiments, we select ResNet-101 as our backbone network.

Effectiveness of the proposed units. In Table 2, we show the units and their combined performance on ResNet-101. We can observe that Base CG and RP-CG respectively bring 0.58% and 0.46% improvements for the bounding box mAP. Combining Base CG and RP-CG together (*i.e.*, our proposed CG-Net), the model can increase mAP by at most 1.24%, in which some categories have large improvements, such as BD (Baseball diamond) 5.05%, SBF (Soccer-ball field) 3.14%, and RA (Roundabout) 2.77%. These results indicate that the feature presentation capabilities have been further improved by Base CG and RP-CG. As for the model efficiency, we can observe that Base CG and RP-CG respectively brings 0.59 and 0.61 M model #Params with 51.53 and 51.89 GFLOPs. When these two models are deployed together, there are 1.79 M model #Params and 154.95 GFLOPs increment.

Mismatching error rates on different scales. To reveal the effect of the proposed method on each level of the feature map, we define mismatching error rates on different scales in the feature pyramid, *i.e.*, the selected level of each object is not consistent with the ground-truth level. It can be seen from Figure 3 that the mismatching error rate of each layer in the feature pyramid has been reduced after deploying our proposed method (*i.e.*, the joint implementation of Base CG and RP-CG). Compared with the low-level feature in the feature pyramid that is more suitable for small objects, the reduction of error rates in high-level is obvious. For example, there are 0.1%, 0.2%, 0.1%, 0.7%, and 1.2% error rate reduction from level $P2$ to $P6$. Therefore, the effectiveness of our method can be further confirmed.

Visualizations. We show some of the visual comparisons between the baseline and the proposed method in Figure 4. From the qualitative experimental results, we can observe that the proposed method achieves notably better precision for relative ambiguous object detection, such as baseball diamond, crossroad, airplane, and tennis court.

4.3. Comparisons with state-of-the-arts

Results on DOTA. The experimental result on the test set of DOTA is shown in Table 3, and the proposed CG-Net ranks first among all the methods. CG-Net achieves the best score

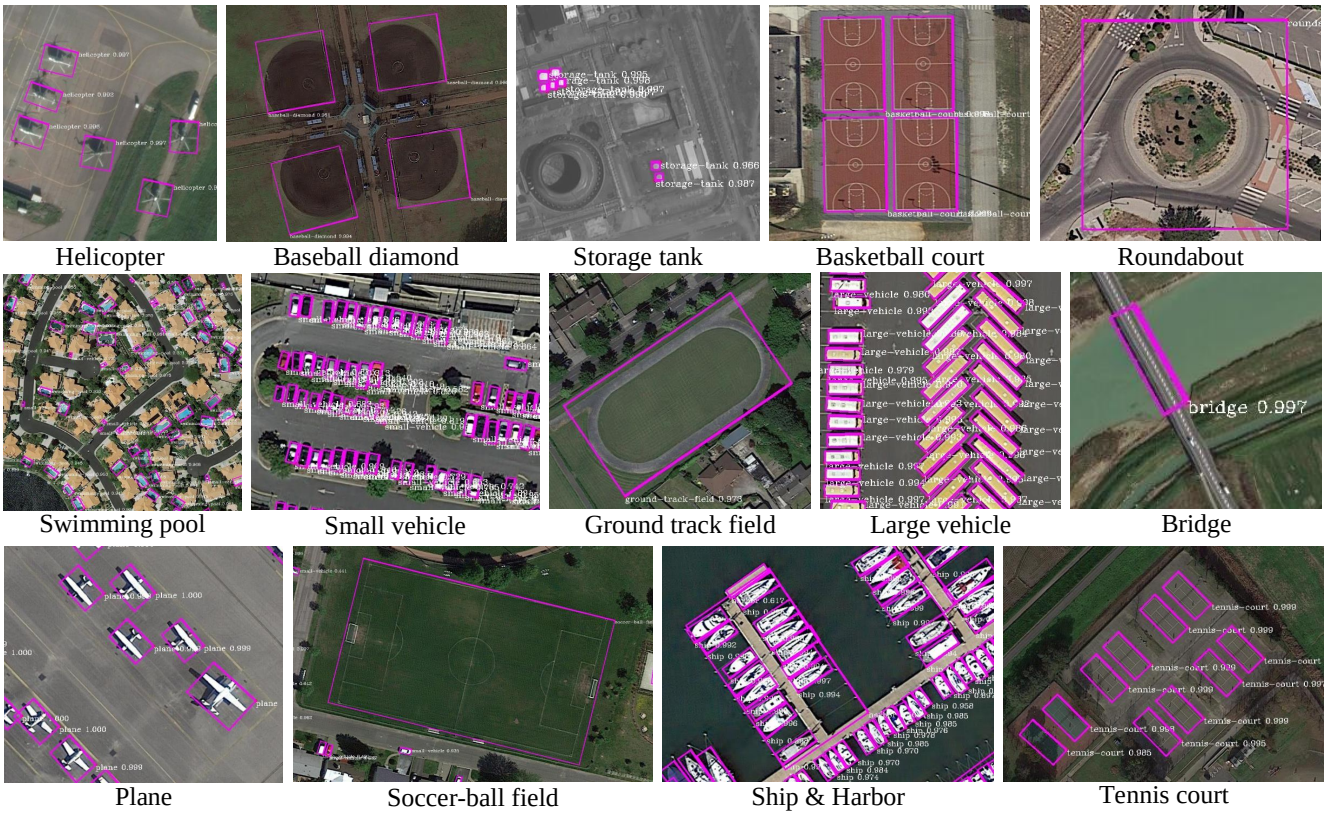

Figure 5. Visualization results for oriented object detection on the test set of DOTA [43]

by 77.89% mAP for oriented object detection and 78.26% mAP for horizontal object detection. Considering the 15 categories, CG-Net ranks at the top for 6 for oriented object detection and 10 for horizontal object detection. Moreover, CG-Net surpasses the state-of-the-art by 0.52% mAP for oriented object detection with a weaker backbone network (ResNet-101 *vs* ResNet-152) and 2.91% mAP for horizontal object detection with the same backbone. Visualization results on the test set of DOTA are shown in Figure 5.

Results on HRSC2016. Result comparisons with the state-of-the-art methods on the test set of HRSC2016 [25] are shown in Table 4. We can observe that our method achieves the state-of-the-art performance in mAP by 90.58%, which surpasses the previous best model by 1.12%. Particularly, in our experiments, our CG-Net uses only 3 horizontal anchors with aspect ratios of $\{1/2, 1, 2\}$, but outperforms the frameworks with a large number of anchors. This shows that it is critical to effectively utilize the predefined anchors to generate high-quality feature, and it may be no need to preset a large number of rotated anchors.

Methods	Backbone	mAP (%)
R^2CNN [17] (preprint 2017)	R-101	73.07
RCI&RC2 [25] (ICPRAM 2017)	V-16	75.70
RRPN [26] (TMM 2018)	R-101	79.08
R^2PN [57] (GRSL 2018)	V-16	79.60
RRD [19] (CVPR 2018)	V-16	84.30
RoI Trans [8] (CVPR 2019)	R-101	86.20
Gliding Vertex [44] (TPAMI 2020)	R-101	88.20
R-RetinaNet [21] (ICCV 2017)	R-101	89.18
R^3Det [47] (preprint 2019)	R-101	89.26
RetinaNet-DAL [27] (AAAI 2021)	R-101	89.77
R^3Det-DCL [46] (CVPR 2021)	R-101	89.46
Ours (CG-Net)	R-101	**90.58**$_{+1.12}$

Table 4. Result comparisons with state-of-the-art methods on the test set of HRSC2016 [25] for oriented object detection in aerial images. "R-" in the Backbone column denotes the ResNet [14], and "V-" denotes the VGG network [36]. mAP is obtained on the VOC 2007 evaluation metric.

5. Conclusion

In this work, a simple yet effective CG operation was proposed to enhance channel communications in a feature transformer fashion, which can adaptively determine the calibration weights for each channel. To demonstrate its effectiveness, we implemented CG on the standard object detection backbone network with a feature pyramid network (*i.e.*, CG-Net) and conducted extensive experiments

on both oriented and horizontal object detection of aerial images. Experimental results on the challenging DOTA and HRSC2016 indicated that the proposed CG-Net could achieve state-of-the-art performance in accuracy with a fair computational overhead. In the future, we will explore applying CG-Net to a broader range of natural scenes. Besides, exploring how to use CG-Net in other visual tasks (*e.g.*, semantic segmentation, person re-identification, and image generation) is also an important direction.

References

[1] Seyed Majid Azimi, Eleonora Vig, Reza Bahmanyar, Marco Körner, and Peter Reinartz. Towards multi-class object detection in unconstrained remote sensing imagery. In *ACCV*, 2018. 7

[2] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, 2019. 4

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2, 3

[4] Jie Chen, Li Wan, Jingru Zhu, Gang Xu, and Min Deng. Multi-scale spatial and channel-wise attention for improving object detection in remote sensing imagery. In *IEEE Geosci Remote Sens Lett*, 2019. 1, 2, 3

[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017. 2, 3

[6] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *NIPS*, 2016. 3, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[8] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, 2019. 1, 3, 5, 6, 7, 8

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2, 3

[10] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *CVPR*, 2019. 4

[11] Ross Girshick. Fast r-cnn. In *ICCV*, 2015. 1

[12] Juan Mario Haut, Ruben Fernandez-Beltran, Mercedes E Paoletti, Javier Plaza, and Antonio Plaza. Remote sensing image superresolution using deep residual channel attention. In *IEEE Trans Geosci Remote Sens*, 2019. 1, 2, 3

[13] Chaoyang He, Haishan Ye, Li Shen, and Tong Zhang. Milenas: Efficient neural architecture search via mixed-level reformulation. In *CVPR*, 2020. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 5, 6, 7, 8

[15] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *ICCV*, 2019. 2

[16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2, 3, 5

[17] Yingying Jiang, Xiangyu Zhu, Xiaobing Wang, Shuli Yang, Wei Li, Hua Wang, Pei Fu, and Zhenbo Luo. R2cnn: rotational region cnn for orientation robust scene text detection. In *arXiv*, 2017. 7, 8

[18] Mirosław Kamiński, Piotr Zientara, and Mirosław Krawczyk. Electrical resistivity tomography and digital aerial photogrammetry in the research of the "bachledzki hill" active landslide–in podhale (poland). In *Eng Geol*, 2021. 1

[19] Minghui Liao, Zhen Zhu, Baoguang Shi, Gui-song Xia, and Xiang Bai. Rotation-sensitive regression for oriented scene text detection. In *CVPR*, 2018. 8

[20] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 2, 4, 5, 6, 7

[21] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1, 3, 4, 8

[22] Youtian Lin, Pengming Feng, and Jian Guan. Ienet: Interacting embranchment one stage anchor free detector for orientation aerial object detection. In *arXiv*, 2019. 4

[23] Lei Liu, Zongxu Pan, and Bin Lei. Learning a rotation invariant detector with rotatable bounding box. In *arXiv*, 2017. 3

[24] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 7

[25] Zikun Liu, Liu Yuan, Lubin Weng, and Yiping Yang. A high resolution optical satellite image dataset for ship recognition and some new baselines. In *ICPRAM*, 2017. 3, 5, 6, 8

[26] Jianqi Ma, Weiyuan Shao, Hao Ye, Li Wang, Hong Wang, Yingbin Zheng, and Xiangyang Xue. Arbitrary-oriented scene text detection via rotation proposals. In *TMM*, 2018. 7, 8

[27] Qi Ming, Zhiqiang Zhou, Lingjuan Miao, Hongwei Zhang, and Linhao Li. Dynamic anchor learning for arbitrary-oriented object detection. In *AAAI*, 2021. 1, 6, 7, 8

[28] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 7

[29] Xingjia Pan, Yuqiang Ren, Kekai Sheng, Weiming Dong, Haolei Yuan, Xiaowei Guo, Chongyang Ma, and Changsheng Xu. Dynamic refinement network for oriented and densely packed object detection. In *CVPR*, 2020. 7

[30] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. In *CVPR*, 2019. 4

[31] Wen Qian, Xue Yang, Silong Peng, Yue Guo, and Chijun Yan. Learning modulated loss for rotated object detection. In *AAAI*, 2021. 3

[32] Zheng Qin, Zeming Li, Zhaoning Zhang, Yiping Bao, Gang Yu, Yuxing Peng, and Jian Sun. Thundernet: Towards real-time generic object detection on mobile devices. In *CVPR*, 2019. 4

[33] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *CVPR*, 2020. 2, 4

[34] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1, 3, 7

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2016. 1, 3, 5, 6

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 8

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3, 4

[38] Chen Wang, Xiao Bai, Shuai Wang, Jun Zhou, and Peng Ren. Multiscale visual attention networks for object detection in vhr remote sensing images. In *IEEE Geosci Remote Sens Lett*, 2018. 1, 2, 3

[39] Qi Wang, Shaoteng Liu, Jocelyn Chanussot, and Xuelong Li. Scene classification with recurrent attention of vhr remote sensing images. In *IEEE Trans Geosci Remote Sens*, 2018. 1, 2, 3

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3

[41] Haoran Wei, Yue Zhang, Zhonghan Chang, Hao Li, Hongqi Wang, and Xian Sun. Oriented objects as pairs of middle lines. In *ISPRS*, 2020. 7

[42] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 2

[43] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *CVPR*, 2018. 1, 3, 5, 6, 7, 8

[44] Yongchao Xu, Mingtao Fu, Qimeng Wang, Yukang Wang, Kai Chen, Gui-Song Xia, and Xiang Bai. Gliding vertex on the horizontal bounding box for multi-oriented object detection. In *TPAMI*, 2020. 3, 8

[45] Xue Yang, Kun Fu, Hao Sun, Jirui Yang, Zhi Guo, Menglong Yan, Tengfei Zhan, and Sun Xian. R2cnn++: Multi-dimensional attention based rotation invariant detector with robust anchor strategy. In *arXiv*, 2018. 3

[46] Xue Yang, Liping Hou, Yue Zhou, Wentao Wang, and Junchi Yan. Dense label encoding for boundary discontinuity free rotation detection. In *arXiv*, 2020. 1, 7, 8

[47] Xue Yang, Qingqing Liu, Junchi Yan, Ang Li, Zhiqiang Zhang, and Gang Yu. R3det: Refined single-stage detector with feature refinement for rotating object. In *arXiv*, 2019. 7, 8

[48] Xue Yang, Hao Sun, Kun Fu, Jirui Yang, Xian Sun, Menglong Yan, and Zhi Guo. Automatic ship detection in remote sensing images from google earth of complex scenes based on multiscale rotation dense feature pyramid networks. In *ISO4*, 2018. 7

[49] Xue Yang and Junchi Yan. Arbitrary-oriented object detection with circular smooth label. In *ECCV*, 2020. 1, 7

[50] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. Scrdet: Towards more robust detection for small, cluttered and rotated objects. In *ICCV*, 2019. 1, 2, 4, 5, 6, 7

[51] Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W Cohen. Breaking the softmax bottleneck: A high-rank rnn language model. In *ICLR*, 2018. 2, 4

[52] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. In *arXiv*, 2018. 2

[53] Dong Zhang, Hanwang Zhang, Jinhui Tang, Meng Wang, Xiansheng Hua, and Qianru Sun. Feature pyramid transformer. In *ECCV*, 2020. 4

[54] Gongjie Zhang, Shijian Lu, and Wei Zhang. Cad-net: A context-aware detection network for objects in remote sensing imagery. In *IEEE Trans Geosci Remote Sens*, 2019. 7

[55] Wenxuan Zhang, Dong Zhang, and Xinguang Xiang. Cascaded and dual: Discrimination oriented network for brain tumor classification. In *ACML*, 2019. 3

[56] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *CVPR*, 2018. 2, 4

[57] Zenghui Zhang, Weiwei Guo, Shengnan Zhu, and Wenxian Yu. Toward arbitrary-oriented ship detection with rotated region proposal and discrimination networks. In *IEEE Geosci Remote Sens Lett*, 2018. 8