IPG-Net: Image Pyramid Guidance Network for Object Detection

Ziming Liu[1], Guangyu Gao[1], Lin Sun[3]
[1]Beijing Institute of Technology
5 Zhongguancun South Street
Beijing
liuziming.email@gmail.com, guangyugao@bit.edu.cn
[3]Samsung Strategy and Innovation Center
3655 North First Street
San Jose, CA
lin1.sun@samsung.com

Abstract

For Convolutional Neural Network based object detection, there is a typical dilemma: the spatial information is well kept in the shallow layers which unfortunately do not have enough semantic information, while the deep layers have high semantic concept but lost a lot of spatial information, resulting in serious information imbalance. To acquire enough semantic information for shallow layers, Feature Pyramid Networks (FPN) is used to build a top-down propagated path. In this paper, except for top-down combining of information for shallow layers, we propose a novel network called *Image Pyramid Guidance Network(IPG-Net)* to make sure both the spatial information and semantic information are abundant for each layer. Our IPG-Net has three main parts: *the image pyramid guidance sub-network*, *the ResNet based backbone network* and *the fusing module*. To the best of our knowledge, we are the first to introduce an image pyramid guidance sub-network, which supplies spatial information to each scale's feature to solve the information imbalance problem. This sub-network promise even in the deepest stage of the ResNet, there is enough spatial information for bounding box regression and classification. Furthermore, we designed an effective *fusing module* to fuse the features from the image pyramid and features from the feature pyramid. We have tried to apply this novel network to both one stage and two stage models, state of the art results are obtained on the most popular benchmark data sets, i.e. MS COCO and Pascal VOC.

Introduction

Recently, with the development of deep convolution neural network, there have been abundant CNN based methods focusing on object detection task since the emergence of typical network of Faster-RCNN (Ren et al. 2017), YOLO (Redmon and Farhadi 2018), SSD (Liu et al. 2016), RetinaNet (Lin et al. 2017b) etc. However, object detection still suffer several problems, such as the key problem of information imbalance of different feature scales. Because the convolution neural network is designed to output a single output for classification, not for multi-scale task.

Some works have tried to fix this imbalance, such as the most popular Feature Pyramid Network (FPN), which mainly fixed the problem of lacking high semantic information in shallow layers. Although feature pyramid network can supply the semantic information for shallow features, there are still feature misalignment and information lost in deeper features. Feature misalignment refers to that there are some offsets between anchors and convolution features.

In this paper, we argue that good feature extractor for detection should have two common features: i) enough shallow image information for bounding box regression, because object detection is a typical regression task. ii) enough semantic information for classification, which means the output features come from deep layers. To satisfy these features above, we introduce a novel network specific for object detection, namely, the Image Pyramid Guidance Network (IPG-Net). The IPG-Net includes two main parts: the *image pyramid guidance sub-network and the feature pyramid of ResNet*. As shown in Fig. 1, it shows the comparison of a standard ResNet and our IPG-Net. The IPG-Net is designed for extracting better features by fixing the information imbalance problem better.

The deep convolution network will cause the loss of the location or spatial information as the layer becomes deeper. This property maybe not a problem for classification task, while bounding box regression is important for detection task. But, the loss of such spatial information results in the features misalignment in object detection. Here, feature alignment means there are some offsets between anchors and convolution features. Besides the lost of spatial information, small objects will easy to be lost in the deeper convolution layers. We argue that all these problems for object detect are due to the limit of the existed convolution network structure and can't be fixed by just simply modifying typical networks.

Here, we introduce the image pyramid to supply more spatial information into each stage of the feature pyramid of the backbone network. Then the above mentioned problems can be reduced in this way. For each stage of the backbone network, we compute the image pyramid feature of the corresponding level in image pyramid. The image pyramid feature is obtained from a shallow sub-network, e.g. image pyramid guidance sub-network, which has more abundant spatial information especially for small objects. Then we de-

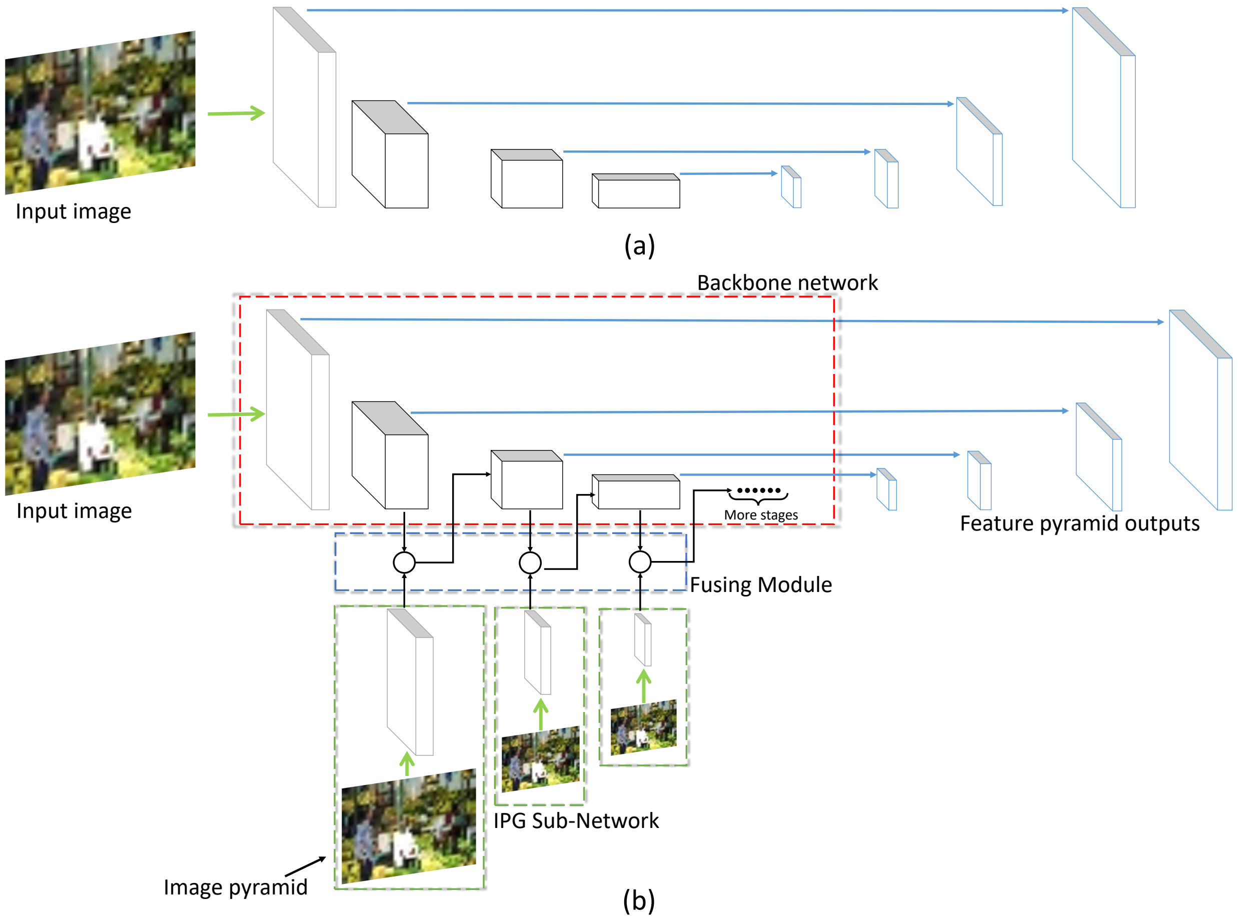

Figure 1: The backbone structure of our model. The first row (a) is a standard feature pyramid network(FPN), the second row (b) is our IPG-Net proposed in this paper, including IPG sub-network, backbone network and the fusing module. The green bounding box denotes the image pyramid guidance sub-network, which responsible to guide the backbone network(within red bounding box). The fusing module is within the blue bounding box. The blue arrow denotes the lateral connection in the FPN and the blue features are outputs of the FPN.

sign a fusion module to fuse the new image pyramid feature into the backbone network.

The fusing module performs two steps to fuse the two kinds of features. Firstly, we transform the original features to align the data size and project them into a hidden space. Secondly, We use common mathematics operation to combine the two features. Sum, product and concatenation are all used in our experiments and improvements of different degree are obtained.

Before going deeper of our proposed methods, we summarize our contributions as below:

- We firstly introduce the concept of image pyramid guidance (IPG) to fix the spatial information and small objects' features lost problem in deep layers.
- We design a new shallow image pyramid guidance subnetwork to extract image pyramid features, which is flexible and light-weighted.
- We also design a flexible fusing module, which is simple but effective.

Related Work

Object detection is a basic task for deeper visual reasoning or visual understanding. The state-of-the-art works based on deep learning for object detection can be classified into one stage model and two stage model(Faster RCNN(Ren et al. 2017), Cascade RCNN(Cai and Vasconcelos 2018), SNIP(Singh and Davis 2018),SNIPER(Singh, Najibi, and Davis 2018) etc.), and one stage model can be further be classified into anchor based methods(Retina net(Lin et al. 2017b), Yolo-v3(Redmon and Farhadi 2018) etc.) and anchor free methods(Center net(Duan et al. 2019), FSAF(Zhu, He, and Savvides 2019) etc.). All of SOTA models are based on the 3 branches, two stage methods are easier to achieve slightly better results while one stage methods have faster speed in practice. There are also some works about design backbone network specific for object detection as what we do here, Detnet is some of them(Li et al. 2018).

Two stage detector Two stage algorithms keep the state of the art results in most popular data sets, such as MS

COCO(Lin et al. 2014), Pascal VOC(Everingham et al. 2010). However, they also suffer from the speed limit and the huge complex of the model building. The information imbalance is also a tough problem for two stage algorithms, although there are some works reduce the imbalance impact in some degree, such as feature pyramid netowrk(Lin et al. 2017a), this is still an unsolved problem.

One stage detector To achieve faster inference speed, a lot of one stage algorithms were proposed and achieved as good performance as two stage models. The initial SOTA one stage models are based on anchor mechanism, but more efficient algorithms of anchor free are proposed recently. The most typical works including center net which motivated by key point detection(Duan et al. 2019), WSMA-Seg which is motivated by segmentation(Cheng et al. 2019), FSAF(Zhu, He, and Savvides 2019). Unfortunately, the information imbalance and the feature misalignment also impact the one stage methods' performance, especially the anchor based detectors.

Information imbalance and Feature alignment There are also some works to solve the imbalance problem in feature level. PANet(Liu et al. 2018) add a bottom-up path on previous FPN to shorten the information propagate path between lower feature and the topmost feature. Pang etc. propose Libra R-CNN which contains balanced feature pyramid to reduce the imbalance in feature level, e.g. the outputs of the feature pyramid network(FPN)(Pang et al. 2019). All of the works above are focusing on fixing the imbalance and misalignment problem, but there is still no one that can solve the problem completely in object detection. Here we propose a novel network, IPG-Net, which is based on image pyramid, the introducing of image pyramid to solve the information imbalance problem is a new path.

Image Pyramid Guidance Network(IPG-Net)

Challenges to be Solved

As mentioned in the last subsection, FPN reduce information imbalance of different scales' features in some degree, but we think there still some challenges eager to be solved. we summary these challenges in this part.

Deep CNNs blur the feature. Deeper convolution network enable better semantic features are extracted in classification task, which don't need to localize the object. However, deep convolution is adverse for object detection, because the location of objects in deep features is not align with the location in the original image. But anchor based detection algorithms rely heavily on the assumption that the location of object is aligned with original images for a any feature. So there is serious misalignment between the anchor and the feature. The phenomenon becomes more serious with depth increase.

FPN suffers the misalignment. Feature pyramid network fuse the deep features and the shallow features, resulting better detection performance. However, because of the blur of deep features, there must be misalignment between the deep features and the shallow features. For example, the spatial position (i, j) corresponds to the object k in the shallow layer, but the spatial position (i, j) corresponds to the object w in the deep layer, k is not equal to w.

Deep CNNs lose small objects. Deep CNNs achieve high performance in classification due to the large stride of 32 respect to initial image size. However, large stride also leads to the miss of the detail information of the input image, e.g. the small object information. Small objects in detection task depend on the detail information of input images, so keeping the detail of small objects is essential for the backbone network. We usually detect small objects in shallow features which lack the high semantic information. Feature pyramid network is often used to build a top to down path to supply semantic information for shallow layers' features. Although FPN introduces the semantic information, the information or features of small objects has been lost in deeper layers, so FPN can't fix the missing problem of small objects.

Overall Structure

The overall structure of our network is shown in Fig. 1. We use the ResNet(He et al. 2016) as the baseline to build our new backbone network, *image pyramid guidance net*, including *image pyramid guidance sub-network, backbone network and fusing module*, which provide a fair comparison with the existing methods.

The image pyramid guidance sub-network accepts a set of images from image pyramid and extracts the image pyramid features for fusing. The function of the sub-network is to extract shallow features to supply the spatial information and the detail information. The image pyramid features are used to guide the backbone network to keep the spatial information and small objects' features. We use a fusing module to perform the guidance. The fusing module's function is to fuse the deep features in backbone network and the shallow features in image pyramid guidance sub-network, the formulation and variants will be talk about in the next subsection. The idea of fusing module is to transform the two types of features and then combine them together to achieve the augment effect for the object detection, especially small objects detection.

Image Pyramid Guidance Sub-Network

Traditionally, we introduce the image pyramid to obtain more scales to reduce the impact of image scale, because convolution network don't have the scale-invariant ability. The performance can be significantly improved in this way, but the computation is also too large to afford in training stage with deep neural network. Different from the traditional purpose, here we use image pyramid to guide the backbone network to learn better features used for the detection. Better features mean that all of the features of different scales have abundant spatial information and enough semantic information, e.g. there are no feature misalignment and information imbalance.

The input of the image pyramid guidance sub-network is a simple image pyramid, which can be formulated as:

$$I = \{I_{H\times W}, I_{\frac{H}{2^1}\times\frac{W}{2^1}}, I_{\frac{H}{2^2}\times\frac{W}{2^2}}, I_{\frac{H}{2^i}\times\frac{W}{2^i}}, ...\}_n \qquad (1)$$

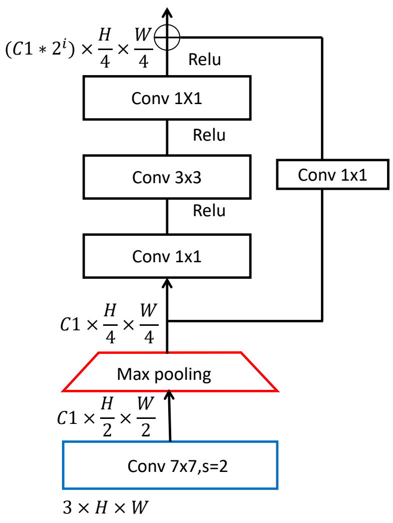
$(C1 * 2^i) \times \frac{H}{4} \times \frac{W}{4}$
Relu
Conv 1X1
Relu
Conv 3x3
Relu
Conv 1x1
Conv 1x1
$C1 \times \frac{H}{4} \times \frac{W}{4}$
Max pooling
$C1 \times \frac{H}{2} \times \frac{W}{2}$
Conv 7x7,s=2
$3 \times H \times W$

Figure 2: The structure of the image pyramid guidance sub-network in level i, the sub-network has different parameters in different levels. The level i determines the output feature's channel dimension is $(C1 * 2^i)$, i start from 0. The channel dimension is consistent with the backbone feature.

where H and W is the image size which is same as the common input image in object detection, n is the number of levels in the image pyramid. We set $n = 4$ in our experiments to be consistent with the depth of the standard ResNet.

Next, we will describe what's the image pyramid guidance sub-network look like, the image pyramid guidance sub-network is shown in Fig. 2. The structure of image pyramid guidance sub-network is component with two parts, one is a 7×7 convolution followed with a 2×2 max pooling, another is a residual block, which is kept same with the design in (He et al. 2016). The residual block accept features with same dimensions and output features with different dimensions that are same as dimensions of features in backbone network. There are two reasons of why we use a shallow network to extract image pyramid feature. On the one hand, the function of IPG is to obtain spatial or detail information, deep convolution will lost these information. On the other hand, the computation complex will not increase too much with the light-weighted design.

The outputs of the image pyramid guidance sub-networks with image pyramid can be formulated as:

$$F = \{F_{H\times W}, F_{\frac{H}{2^1}\times\frac{W}{2^1}}, F_{\frac{H}{2^2}\times\frac{W}{2^2}}, F_{\frac{H}{2^i}\times\frac{W}{2^i}}, ...\}_n \tag{2}$$

$$F_{\frac{H}{2^i}\times\frac{W}{2^i}} = f(I_{\frac{H}{2^i}\times\frac{W}{2^i}}) \tag{3}$$

where the $f(\cdot)$ denotes the image pyramid guidance sub-network, as shown in Fig. 2, $F_{\frac{H}{2^i}\times\frac{W}{2^i}}$ denotes the image pyramid feature of the level i. All of the features from different level of image pyramid form image pyramid features F.

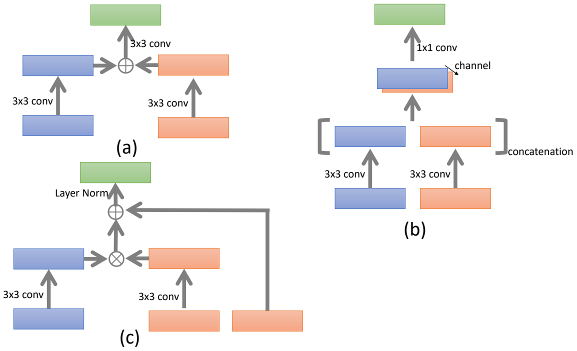

Figure 3: Three instances of the fusing module, (a) is the sum strategy, (b) is the concatenation strategy, (c) is the residual product strategy. All of them are the variants of the Eq. 4.

Backbone Network

The backbone network is modified from the standard ResNet which contains *res 1-5*. We add new stages at the end of the ResNet, each new stage contains two Bottleneck modules, same as ResNet. Our ablation studies suggest adding one new stage can perform better than the other conditions. Too deep backbone network also is harmful for the detection, We argue that the backbone which is too deep has difficulty in training.

The reason why we design deeper convolution network than the standard ResNet is the image pyramid guidance sub-network supply enough spatial information or detail information into backbone network, which reduce the impact of feature misalignment or detail lost. The advantage of the deep backbone network is that the backbone network can generate better semantic information which is good for the classification. Another advantage is the network can cover larger range of the scales of the object.

Fusing Module

Formulation. The fusing module in this paper is a flexible enough module, we first formulate it as following. The $f(\cdot)$ and $g(\cdot)$ correspond to the network of IPG sub-network and backbone network separately. The function of β can be flexible with different versions.

$$O_i = \beta(f_i(I_i), g_i(I_0)), i \in [0, n-1] \tag{4}$$

where O_i is the output feature the of fusing module in level i. I_0 and I_i are images in the image pyramid in level 0 and level i separately. The $\beta(\cdot)$ denotes the fusing function of the fusing module. The $f_i(\cdot)$ denotes the output of the image pyramid guidance sub-network in level i and the $g_i(\cdot)$ denotes the output of the backbone network in level i. If there are n images in image pyramid, the number of levels is n.

The fusing module is shown in the blue box in the Fig. 1. In this case, there are two inputs, image pyramid features from image pyramid guidance sub-network and the feature from the backbone network. We propose several different variants to demonstrate the effectiveness of image pyramid guidance. Sum, Product and Concatenation are the three types of fusing modules we use in our experiments. We believe that other similar design of the fusing module will also works well in our IPG-Net, especially some attention design, but we will not focus on that in this paper, we will follow the direction in future works. Next, we will describe the details of three type of variants.

We designed several variants of the fusing module to prove the robust of our image pyramid guidance mechanism. The details of them are shown in the following sections.

Element-wise Sum. In this version, we regard the image pyramid information as an additional information, so the aim is to sum the image pyramid features $f_i(I_i)$ and main features $g_i(I_0)$. Due to the parameters of image pyramid network are all shared, so we need to align the channel dimension of the two types of features. Here, we use channel-dimension linear interpolate operation to perform the $CT(channeltransform)$.

$$O_i = W \cdot [W_s \cdot CT(f_i(I_i)) + W_m \cdot g_i(I_0)] \quad (5)$$

Where the W, W_s, W_m denotes the linear transforms, f_i denotes the image pyramid guidance sub-network and g_i denotes the backbone network from $stage0 - i$.

Residual Product. Here we use the product $W_s \cdot CT(f_i(I_i)) * W_m \cdot g_i(I_0)$ to represent the lost information in main-features $g_i(I_0)$. After adding the missing information into main-features, we use a ”layer norm” operation to normalize the processed main-feature O_i.

$$O_i = LN\{[W_s \cdot CT(f_i(I_i)) * W_m \cdot g_i(I_0)] + g_i(I_0)\} \quad (6)$$

Where the LN denotes the Layer Norm operation.

Concatenation. We also try to use concatenation operation to realize the fusing of the image pyramid feature and the main-feature, which is similar to the fusing operation in U-net(Ronneberger, Fischer, and Brox 2015). The formulation is shown as following.

$$O_i = W \cdot Cat[W_s \cdot CT(f_i(I_i)), W_m \cdot g_i(I_0)] \quad (7)$$

Where the Cat denotes the concatenation operation.

Experiments

Experiment Details

Datasets We conduct ablation experiments on two data sets, MSCOCO(Lin et al. 2014) and Pascal VOC(Everingham et al. 2010). MSCOCO is the most common benchmark for object detection, the COCO data set is divided into train, validation, including more than 200,000 images and 80 object categories. Following common practice, we train on the COCO *train2017*(i.e. *trainval 35k* in 2014) and test on the COCO *val 2017* data set(i.e. *minival* in 2014) to conduct ablation studies. Finally, we also report our state of the art results in MS COCO *test-dev*, the test is finished in CodaLab[1] platform. We also apply our algorithm on another popular data set, Pascal VOC. Pascal VOC 2007 has 20 classes and 9,963 images containing 24,640 annotated objects and Pacal VOC 2012 also has 20 classes and 11,530 images containing 27,450 annotated objects and 6,929 segmentation. We train our model with Pascal VOC 2007 *trainval* set and Pascal VOC 2012 *trainval* set and test the model with Pascal VOC2007 *test*.

Training We follow the common training strategies for object detection, 12 epoch with 4 mini-batch in each GPU. All of the experiments are conducted in 8 NVIDIA P100 GPUs, optimized by SGD(stochastic gradient descent) and default parameters of SGD in pytorch framework are adopted. The learning rate is set as 0.01 at the beginning and decrease by a factor of 0.1 in epoch 7 and epoch 11. The linear warm-up strategy is also used, the number of warm-up iterations is 500 and the warm-up ratio is 1.0/3. All of the input images are resized into 1333×800 in COCO and 1000×800 in Pascal VOC, which is consist with the common practice. The image pyramid is obtained by down-sampling(linear interpolate) the input image into four levels with a factor of 2.

Inference The image size of image pyramid keep same with the training stage. The IOU threshold of NMS is 0.5, and the score threshold of predicted bounding box is 0.05. The max number of the bounding box of each image is set as 100.

MS COCO

Which fusing strategy is better. We propose three different strategies to fusing the features from image pyramid and the features of the backbone network in this paper. To compare the effectiveness and the difference of them, we perform different strategies in a same baseline and report the AP of small, middle and large objects separately. The results in Table. 3 shows that all of three versions have similar results for small objects$(20.8vs18.9vs19)$, but the results for middle objects and large objects have large margin$(2\% - 4\%)$ between them. Table. 3 shows that the sum operation achieve much better performance in all metrics. We argue that the sum operation is easy to optimize, while product and concatenation are those operations with more tricks, e.g. hard to optimize. Here, we perform the rest experiments with sum fusing module.

How deep is the IPG-Net. The Table. 4 shows that the mAP is not always increase with the depth increase, and we also notice that the improvement comes from the large objects, while the small objects slightly decrease, $0.3\%(21.2vs21.1vs20.8)$. We also study the effect of keeping spatial size of the last 3 stages, as the (Li et al. 2018) proposed. The results shows that there is slightly improvement for small objects $(20.8vs21)$ and middle objects $(39.3vs39.6)$, but the performance improve in mAP is not significant. Considering the computation complex and the model performance, the depth of 5 stages is the best choice for the IPG RCNN. Here, we construct the IPG RCNN with a 4 levels image pyramid guidance sub-network and a Faster RCNN head, the backbone network is a ResNet50 which ranges from stage 1 to stage 4.

[1] https://competitions.codalab.org/competitions/20794

model	backbone	AP	$AP50$	$AP75$	AP_S	AP_M	AP_L
Two Stage Det							
R-FCN(Dai et al. 2016)	ResNet-101	29.9	51.9	-	10.8	32.8	45.0
Faster RCNN++(He et al. 2016)	ResNet-101	34.9	55.7	37.4	15.6	38.7	50.9
Faster RCNN w FPN(Lin et al. 2017a)	ResNet-101	36.2	59.1	39.0	18.2	39.0	48.2
DeNet-101(wide)(Ghodrati et al. 2015)	ResNet-101	33.8	53.4	36.1	12.3	36.1	50.8
CoupleNet(Zhu et al. 2017)	ResNet-101	34.4	54.8	37.2	13.4	38.1	50.8
Deformable R-FCN(Dai et al. 2016)	Aligned-Inception-ResNet	37.5	58.0	40.8	19.4	40.1	52.5
Mask-RCNN(He et al. 2017)	ResNeXt-101	39.8	62.3	43.4	22.1	43.2	51.2
Cascade RCNN(Cai and Vasconcelos 2018)	ResNet-101	42.8	62.1	46.3	23.7	45.5	55.2
SNIP++(Singh and Davis 2018)	ResNet-101	43.4	**65.5**	48.4	**27.2**	46.5	54.9
SNIPER(2scale)(Singh, Najibi, and Davis 2018)	ResNet-101	43.3	63.7	48.6	27.1	44.7	56.1
Grid-RCNN(Lu et al. 2018)	ResNeXt-101	43.2	63.0	46.6	25.1	46.5	55.2
Anchor based One Stage Det							
SSD512(Liu et al. 2016)	VGG-16	28.8	48.5	30.3	10.9	31.8	43.5
YOLOv2(Redmon and Farhadi 2017)	DarkNet-19	21.6	44.0	19.2	5.0	22.4	35.5
DSSD513(Fu et al. 2017)	ResNet-101	31.2	50.4	33.3	10.2	34.5	49.8
RetinaNet80++(Lin et al. 2017b)	ResNet-101	39.1	59.1	42.3	21.8	42.7	50.2
RefineDet512(Zhang et al. 2018)	ResNet-101	36.4	57.5	39.5	16.6	39.9	51.4
M2Det800	VGG-16	41.0	59.7	45.0	22.1	46.5	53.8
Anchor Free One Stage Det							
CornetNet511(Law and Deng 2018)	Hourglass-104	40.5	56.5	43.1	19.4	42.7	53.9
FCOS(Tian et al. 2019)	ResNeXt-101	42.1	62.1	45.2	25.6	44.9	52.0
FSAF(Zhu, He, and Savvides 2019)	ResNeXt-101	42.9	63.8	46.3	26.6	46.2	52.7
CenterNet511(Duan et al. 2019)	Hourglass-104	44.9	62.4	48.1	25.6	47.4	57.4
IPG RCNN	IPG-Net101	**45.7**	64.3	**49.9**	26.6	**48.6**	**58.3**

Table 1: The state of the art of the performance on the MS COCO *test-dev*, '++' denotes that the inference is performed with multi-scales etc.

Where to perform fusing. Here we conduct ablation experiments using a IPG-Net or a ResNet with 4 stages. Firstly, we only add one image pyramid feature into backbone network. Secondly, we also increase the level of the image pyramid to find out if more levels are better. The Table. 5 shows that IPG-Net with different configures all achieve slightly improvement compared with baseline ResNet. The best mAP of them is 36.6%, which is only 0.1% improvement from the others. We conclude that the IPG-Net is not sensible enough for the position of image pyramid features and the increase of the image pyramid level also has little effect. All in all, the experiment here indeed improve the performance.

The effect on deep layers. As we claimed in this paper, the function of image pyramid guidance is to supply the spatial information and the image details information of small objects in to deep features. Here, we conduct a simple comparison experiment to prove the effectiveness of IPG in deep layers. The configure of the experiment is simple but persuasive. The depth of the IPG-Net and the ResNet is 7 stages but we only use 4 outputs of the last four stages, which are all deep features without enough detail information. The detector we use here is RetinaNet(Lin et al. 2017b), which relies on each level of the feature pyramid.

The Table. 6 shows that IPG-Net achieve higher performance than ResNet backbone in almost all metrics. The increase of AP reaches $0.6\%(24.8vs24.2, 39.9vs39.3)$. The results of Table. 6 also suggest that the IPG-Net works on RetinaNet(Lin et al. 2017b)(a one stage detector). We also notice that the IPG have more significant effect on RetinaNet(0.6%) than Faster RCNN $(< 0.6\%)$, because the two stage model perform ROI Pooling in shallow layers' features while the one stage model consider features of both shallow features and deep features.

Comparison with the state of the art results in MS COCO *test-dev* Finally, we also test our IPG RCNN in MS COCO *test-dev* to make a comparison with the state of the art detectors. We construct a modidified IPG RCNN with a IPG-Net101 and a cascade RCNN head(Cai and Vasconcelos 2018). The image pyramid guidance sub-network choose stage 3 as the level to perform fusing module, because the IPG-Net is not sensible with position of fusing. The depth of the IPG-Net is four stages to make full use of the pre-trained parameters of standard ResNet in ImageNet. The IPG RCNN achieve $45.7mAP$ in MS COCO *test-dev*, which is the state of the art result compared with other detectors in the condition of single scale inference.

Pascal VOC

Comparison with the state of the art results in Pascal VOC. To valid the results more properly, we also test the new *IPG RCNN*(based on Faster RCNN(Ren et al. 2017)) in Pascal VOC data set. The baseline is a faster RCNN with the ResNet-50 as backbone network, the performance of the baseline Faster RCNN is much better than the original paper(Ren et al. 2017), reaching $79.8\%mAP$. Then we add the fusing module into *stage 3* following the ablation studies to construct a IPG RCNN with a IPG-Net50 and a faster RCNN head. The Table. 2 shows that the IPG-Net-50 obtains $80.5\%mAP$, we further apply multi-scale inference strategy$((800, 500), (1000, 600), (1333, 800))$ to test the effort of the IPG-Net-50, resulting in $81.6\%mAP$. Furthermore, to keep consist with the previous works, we also use a 101 layers IPG-Net to get the state of the art result, the IPG-Net-101 is also fine-turned with pre-trained parameters of COCO data set. The results of single scale and multi-scale all tested in Pascal VOC2007 *test*. Table 2 shows that IPG RCNN101 achieves 84.8 in single scale and 85.9 in multi-

model	backbone	input size	mAP
Two Stage Det			
Faster RCNN(He et al. 2016)	ResNet-101	1000x600	**76.4**
R-FCN(Dai et al. 2016)	ResNet-101	1000x600	80.5
OHEM(Shrivastava, Gupta, and Girshick 2016)	VGG-16	1000x600	74.6
HyperNet(Kong et al. 2016)	VGG-16	1000x600	76.3
R-FCN w DCN(Dai et al. 2017)	ResNet-101	1000x600	82.6
CoupleNe(Zhu et al. 2017)t	ResNet-101	1000x600	82.7
DeNet512(wide)(Ghodrati et al. 2015)	ResNet-101	512x512	77.1
FPN-Reconfig(Kong et al. 2018)	ResNet-101	1000x600	82.4
One Stage Det			
SSD512(Liu et al. 2016)	VGG-16	512x512	79.8
YOLOv2(Redmon and Farhadi 2017)	Darknet	544x544	78.6
RefineDet512(Zhang et al. 2018)	VGG-16	512x512	81.8
RFBNet512(Liu, Huang, and Wang 2018)	VGG-16	512x512	82.2
CenterNet(Zhou, Wang, and Krhenbhl 2019)	ResNet-101	512x512	78.7
CenterNet(Zhou, Wang, and Krhenbhl 2019)	DLA(Zhou, Wang, and Krhenbhl 2019)	512x512	80.7
Ours			
Faster RCNN(Ren et al. 2017)	ResNet-50	1000x600	79.8
IPG RCNN	IPGnet-50	1000x600	80.5
IPG RCNN++	IPGnet-50	1000x600	**81.6**
IPG RCNN	IPGnet-101	1000x600	84.8
IPG RCNN++	IPGnet-101	1000x600	**85.9**

Table 2: The state of the art of the performance on the Pascal VOC 2007 *test*, '++' denotes that inference is performed with three scales.

model	fusing strategy	AP_S	AP_M	AP_L
IPG RCNN	sum	**20.8**	**39.6**	**46.2**
	product	**18.9**	36.3	43.4
	concatenation	**19**	35.5	42.6

Table 3: The ablation study of the fusing module on the MS COCO *minival*.

model	N stages	mAP	AP50	AP75	AP_S	AP_M	AP_L
IPG RCNN	4	35.4	57.9	37.8	21.2	39.2	44.9
	5	**35.7**	**58.2**	**38.2**	*21.1*	**39.6**	**45.7**
	6	35.7	58.2	38.3	*20.8*	39.3	45.8
	7(keep)	35.7	58	38.3	*21*	39.6	45.8

Table 4: The ablation study of the depth of backbone on the COCO *minival*, $7(keep)$ denotes the depth of backbone is 7 stages and the spatial size of the last 3 stages keeps constant.

scale. The results on two popular benchmark show that the IPG RCNN is robust enough and effective.

Conclusion

In this paper, the main problem we concentrate on is the information imbalance of the object detection. In the previous backbone of detection, there is serious information imbalance between the shallow layer and the deep layer. In this paper, we propose a novel *image pyramid guidance net(IPG-Net)*, including a new sub-network based on image pyramid, a fusing module and a backbone network based on ResNet. The new sub-network can extract proper features full of the spatial information and small objects' information. The image pyramid feature from sub-network and the feature from backbone network are fused together by a fusing module to reduce the feature misalignment problem and small objects' missing problem in deep layers. We conduct abundant ablation experiment to prove the effectiveness of the new image pyramid guidance net. The work also can be extend to video object detection task further with the natural advantage of the image pyramid guidance.

stage1	stage2	stage3	stage4	mAP	AP50	AP75
-	-	-	-	36.3	58.1	39.0
√	-	-	-	36.5	58.4	39.3
-	√	-	-	36.2	58.1	39.0
-	-	√	-	**36.6**	58.4	**39.4**
-	-	-	√	36.5	58.4	39.2
-	√	√	√	36.5	58.4	**39.4**

Table 5: The ablation study of the position of the fusing module in IPG-Net, we add only one fusing module into one stage and also add multi-modules into multi-stages.

model	mAP	AP50	AP75	AP_S	AP_M	AP_L
IPG-Net	**23.9**	40.2	**34.8**	**4**	**28.7**	**39.9**
ResNet	23.6	40.2	24.2	3.9	28.3	39.3

model	AR	AR50	AR75	AR_S	AR_M	AR_L
IPG-Net	**23.7**	**36.2**	**38.5**	**12.4**	**43.1**	**60.9**
ResNet	23.4	35.7	38	12	42.3	60.7

Table 6: The effect of IPG in deep layers on the COCO val based on RetinaNet-50.

References

Cai, Z., and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Cheng, Z.; Wu, Y.; Xu, Z.; Lukasiewicz, T.; and Wang, W. 2019. Segmentation is all you need.

Dai, J.; Li, Y.; He, K.; and Sun, J. 2016. R-fcn: Object detection via region-based fully convolutional networks.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; and Tian, Q. 2019. Centernet: Keypoint triplets for object detection.

Everingham, M.; Van Gool, L.; Williams, C. K. I.; Winn, J.; and Zisserman, A. 2010. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* 88(2):303–338.

Fu, C.-Y.; Liu, W.; Ranga, A.; Tyagi, A.; and Berg, A. C. 2017. Dssd : Deconvolutional single shot detector.

Ghodrati, A.; Diba, A.; Pedersoli, M.; Tuytelaars, T.; and Gool, L. V. 2015. Deepproposal: Hunting objects by cascading deep convolutional layers. *2015 IEEE International Conference on Computer Vision (ICCV)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Kong, T.; Yao, A.; Chen, Y.; and Sun, F. 2016. Hypernet: Towards accurate region proposal generation and joint object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kong, T.; Sun, F.; Huang, W.; and Liu, H. 2018. Deep feature pyramid reconfiguration for object detection. *Lecture Notes in Computer Science* 172188.

Law, H., and Deng, J. 2018. Cornernet: Detecting objects as paired keypoints. *Lecture Notes in Computer Science* 765781.

Li, Z.; Peng, C.; Yu, G.; Zhang, X.; Deng, Y.; and Sun, J. 2018. Detnet: A backbone network for object detection.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollr, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. *Lecture Notes in Computer Science* 740755.

Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017a. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollar, P. 2017b. Focal loss for dense object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; and Berg, A. C. 2016. Ssd: Single shot multibox detector. *Lecture Notes in Computer Science* 2137.

Liu, S.; Qi, L.; Qin, H.; Shi, J.; and Jia, J. 2018. Path aggregation network for instance segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Liu, S.; Huang, D.; and Wang, Y. 2018. Receptive field block net for accurate and fast object detection. *Lecture Notes in Computer Science* 404419.

Lu, X.; Li, B.; Yue, Y.; Li, Q.; and Yan, J. 2018. Grid R-CNN. *CoRR* abs/1811.12030.

Pang, J.; Chen, K.; Shi, J.; Feng, H.; Ouyang, W.; and Lin, D. 2019. Libra r-cnn: Towards balanced learning for object detection.

Redmon, J., and Farhadi, A. 2017. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Redmon, J., and Farhadi, A. 2018. Yolov3: An incremental improvement.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2017. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):11371149.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention MICCAI 2015* 234241.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Singh, B., and Davis, L. S. 2018. An analysis of scale invariance in object detection - snip. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Singh, B.; Najibi, M.; and Davis, L. S. 2018. Sniper: Efficient multi-scale training.

Tian, Z.; Shen, C.; Chen, H.; and He, T. 2019. Fcos: Fully convolutional one-stage object detection.

Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; and Li, S. Z. 2018. Single-shot refinement neural network for object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhou, X.; Wang, D.; and Krhenbhl, P. 2019. Objects as points.

Zhu, Y.; Zhao, C.; Wang, J.; Zhao, X.; Wu, Y.; and Lu, H. 2017. Couplenet: Coupling global structure with local parts for object detection. *2017 IEEE International Conference on Computer Vision (ICCV)*.

Zhu, C.; He, Y.; and Savvides, M. 2019. Feature selective anchor-free module for single-shot object detection.