

Zoom Out-and-In Network with Map Attention Decision for Region Proposal and Object Detection

Hongyang Li · Yu Liu · Wanli Ouyang · Xiaogang Wang

Received: date / Accepted: date

Abstract In this paper, we propose a zoom-out-and-in network for generating object proposals. A key observation is that it is difficult to classify anchors of different sizes with the same set of features. Anchors of different sizes should be placed accordingly based on different depth within a network: smaller boxes on high-resolution layers with a smaller stride while larger boxes on low-resolution counterparts with a larger stride. Inspired by the conv/deconv structure, we fully leverage the low-level local details and high-level regional semantics from two feature map streams, which are complimentary to each other, to identify the objectness in an image. A map attention decision (MAD) unit is further proposed to aggressively search for neuron activations among two streams and attend the most contributive ones on the feature learning of the final loss. The unit serves as a decision-maker to adaptively activate maps along certain channels with the solely purpose of optimizing the overall training loss. One advantage of MAD is that the learned weights enforced on each feature channel is predicted on-the-fly based on the input context, which is more suitable than the fixed enforcement of a convolutional kernel. Experimental results on three datasets, including PASCAL VOC 2007, ImageNet DET, MS COCO, demonstrate the effectiveness of our proposed algorithm over other state-of-the-arts, in terms of average recall (AR) for region proposal and average precision (AP) for object detection.

Keywords Object Detection · Region Proposals · Zoom Network · Map Attention Decision

H. Li · Y. Liu · X. Wang
Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong, China
E-mail: {yangli, yuliu, xgwang}@ee.cuhk.edu.hk

W. Ouyang
University of Sydney, Sydney, Australia
E-mail: wanli.ouyang@sydney.edu.au

1 Introduction

Object proposal is the task of proposing a set of candidate regions or bounding boxes in an image that may potentially contain an object. In recent years, the emergence of object proposal algorithms (Uijlings et al., 2013; Manén et al., 2013; Arbeláez et al., 2014; Hayder et al., 2016; Kong et al., 2016; Ghodrati et al., 2016; Chavali et al., 2016; Sun et al., 2016; Li et al., 2017a) have significantly boosted the development of many vision tasks, (Liu et al., 2017a,b; Li et al., 2016; Chi et al., 2016; Li et al., 2017b), especially for object detection (Girshick et al., 2014; Dai et al., 2016; Girshick, 2015; Bell et al., 2016; Liu et al., 2016). It is verified by Hosang *et al.* (Hosang et al., 2015) that region proposals with high average recall correlates well with good performance of a detector. Thus generating object proposals has quickly become the de-facto pre-processing step.

Currently, CNN models are known to be effective in generating candidate boxes (Ren et al., 2015; Kuo et al., 2015; Kong et al., 2016). Existing works use deep CNN features at the last layer for classifying whether a candidate box should be an object proposal. The candidate box can come from random seed (Gidaris and Komodakis, 2016), external boxes (selective search (Uijlings et al., 2013), edge box (Zitnick and Dollar, 2014), *etc.*), or sliding windows (Sermanet et al., 2014). Deep CNN-based proposal methods employ a zoom-out network, where down-sampling is used for reducing the resolution of features. This zoom-out design is good for image classification since down-sampling is effective for achieving translation invariance, increasing the receptive field of features, and saving computation.

However, we argue that the zoom out structure faces great challenge by using the same set of features and the same classifier to handle object proposals in different sizes. The learned features have to sacrifice on large objects in order to compromise small ones. It also has two problems for

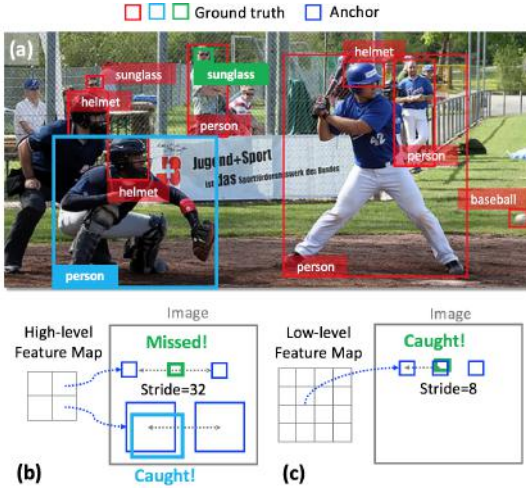


Fig. 1: (a) Input image where blue and green objects are being examined. (b) A larger stride in higher layers miss the green box between two adjacent anchors. (c) This motivates us to place small-scale anchors in the preceding layers with a smaller stride.

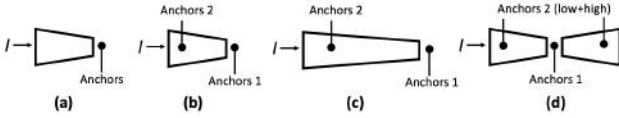


Fig. 2: Different network design for region proposals. (a) RPN (Ren et al., 2015); (b) RPN with split anchors; (c) Deeper RPN with split anchors; (d) Zoom out-and-in network adopted in this paper.

detecting small objects. First, candidate anchors are placed at the final feature map in existing works (Ren et al., 2015). As shown in Figure 1(b), when the down-sampling rate (or total stride) of the feature map is 32, moving the anchor by one step on the feature map corresponds to stepping by 32 pixels in the image. Anchors might skip small objects due to a larger stride. Second, for the case of small objects, the down-sampling operation makes the network difficult to determine whether neurons or features will be activated or not for the subsequent layer. The lack of resolution in feature map is a factor that influences the ability of object proposal methods in finding small objects. As depicted in Figure 1(c), if the resolution of feature map is sufficient, the anchor box with smaller stride can locate small objects. To fit for the zoom-out network design, one could leverage features at shallow layers which has higher resolution. Features from shallow layers are yet weak in extracting high-level information that is essential for object proposal. It would be desirable if high-level semantics can pass information to guide feature learning in the lower counterparts for small objects.

Inspired by the conv/deconv structure (Newell et al., 2016; Long et al., 2015), we propose a Zoom-out-and-In network

for object *Proposals* (ZIP) to both utilize feature maps from high and low streams. Figure 2 shows the network structure of different designs at a glance. Under such a scheme not only the detection of small objects enhances, but also the detection of larger anchors increases. By splitting anchors of different size, we make each level have its own classifier and own set of features to handle a specific range of scales.

In this paper, we devise a *Map Attention Decision* unit (MAD), to actively search for neuron activations among feature maps from low-level and high-level streams. Figure 3 illustrates the pipeline of our algorithm. MAD unit is a side branch taking the feature maps in the highest level as input and generates a column vector, whose length corresponds to two times as the number of channels in maps that MAD enforces on. The weight vector evaluates the contribution of each map on the feature learning in the final RPN loss layer. Such an intuition is derived from the gradient of loss with respect to MAD unit (see Section 3.2). The proposed unit involves parameter update during training and yet is relatively isolated from the weights of the network. In this sense, it resembles an external memory in neural networks (Graves et al., 2016; Wen et al., 2016). One advantage of MAD in selecting neuron activations among feature maps is that it considers the context of input - the weight vector is computed on the fly for each training or test sample, which shares a big distinction in previous work where weights in the learned combination kernel are fixed (see Section 2).

Our algorithm is implemented in Caffe (Jia et al., 2014) with Matlab wrapper and trained on multiple GPUs. The codebase and object proposal results of our method are available online¹. To sum up, our contributions in this work are as follows:

1. A map attention decision unit to actively search for neuron activations among feature maps and attend the most contributive ones on the feature learning of the loss. It adapts the weights among feature maps on the fly based on the context of input.
2. A zoom-out-and-in network that utilizes feature maps at different depth of a network to leverage both low-level details and high-level semantics. Anchors are placed separately based on different resolutions of maps to detect objects of various sizes.
3. Several practical techniques for generating effective proposals and enhancing the object detection performance, including recursive test and training, boosting the classifiers, etc., are investigated. The proposed ZIP algorithm achieves average recall (AR) to 68.8% and 61.2% at top 500 proposals on ILSVRC DET and MS COCO, respectively. Furthermore, the proposed boxes will improve average precision (AP) by around 2% for object detection compared to previous state-of-the-art.

¹ <https://github.com/hli2020/zoom.network>

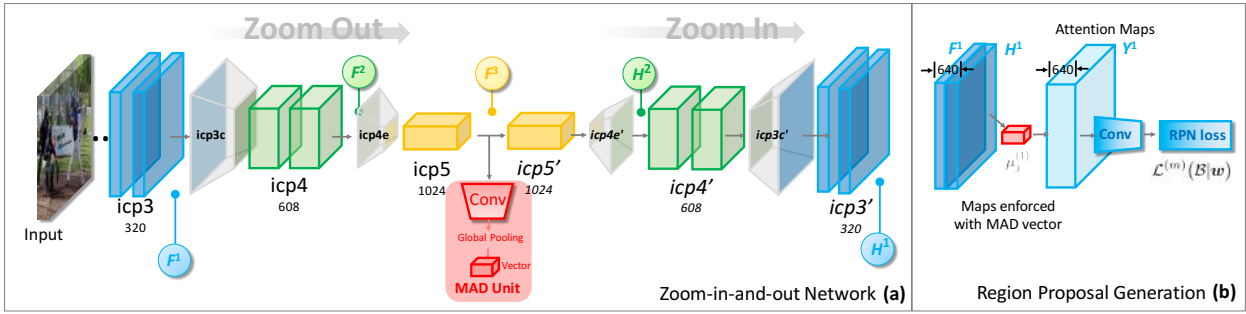


Fig. 3: (a) Zoom network with map attention decision (MAD) unit. Prior to down-sampling in each block, we utilize the feature maps at that depth to place different size of anchors and identify the objectness in the image. Feature maps are actively determined by the MAD unit, using a selective vector μ to decide the attention of neurons from both low-level and high-level layers. (b) The region proposal generation pipeline after the zoom network. It only shows the case on level $m = 1$, where maps F^1, H^1 from two streams, are leveraged by a MAD vector to generate the attention maps Y^1 . Level $m = 2$ shares similar workflow as $m = 1$ whilst level $m = 3$ neither has merging from different sources nor applies MAD. The complete network diagram for region proposal and object detection is depicted in Figure 6.

The rest of the paper is organized as follows: Section 2 states the related work from four aspects: the spirit of utilizing various information from different levels (depth or resolution) in the network, conv/deconv structure walkthrough, region proposal and object detection. Section 3 and 4 depict the detailed description on each component of the proposed framework for region proposal and object detection, respectively. Section 5 verifies the effectiveness of ZIP algorithm by experiments. Section 6 concludes the paper.

2 Related Work

Using features from multi-depth in the network. The idea of utilizing different feature maps at different locations in the network has been investigated and proved to be effective (Liu et al., 2016; Pinheiro et al., 2016; Lin et al., 2017). Pinheiro et al. (Pinheiro et al., 2016) proposed a network to fully leverage all feature maps. Instead of simply combining all feature maps, they devised a upsampling path to gradually refined maps that contained high-level semantics. SSD (Liu et al., 2016) considers the idea of putting different scales of anchors on various depth within the network. However, SSD inherently does not consider the zoom-in (deconvolution) structure to help guide the learning in lower layers. Lin et al. (Lin et al., 2017) introduced a feature pyramid network which up-samples high-level features and generates multiple predictions. However, the top-down path is achieved by convolution kernels: once the training is finished, weights are fixed for test; in ours, the decision of neuron activations are determined by the input’s context, i.e., based on the predictions in the MAD vector.

Conv/deconv structure. The spirit of upsampling feature maps through learnable parameters is known as deconvolution, which is widely applied in other vision domains

(Long et al., 2015; Noh et al., 2015; Hariharan et al., 2014; Newell et al., 2016). Shelhamer et al. (Long et al., 2015) first devise a novel structure to do pixel-wise semantic segmentation via learnable deconvolution. In (Ronneberger et al., 2015), a U-shaped network for segmentation is designed with a contracting path to capture context and a symmetric expanding path to localize objects. Newell et al. (Newell et al., 2016) proposed an hourglass structure where feature maps are passed with skip connections through stacks for pose estimation. DSSD (Fu et al., 2017) embedded the original SSD framework into a similar U-net structure with more discriminative features due to the up-sampling subnetwork; and more sophisticated prediction modules are designed for each detection branch. Our work is inspired by these works and yet have clear distinctions besides different application domain. Existing approaches concatenate all features (Hariharan et al., 2014) or use the final feature map for prediction (Newell et al., 2016), while we utilize specific features at different locations of a network to detect objects of various size. With such a philosophy in mind, we have each object equipped with suitable features at a proper resolution.

Region proposals. In early stages people resort to finding the objectness via multiple cues and larger pixels (called superpixel) in a semantic manner (Uijlings et al., 2013; Alexe et al., 2012; Krahenbuhl and Koltun, 2014). In (Krahenbuhl and Koltun, 2015), a learning method is proposed by training an ensemble of figure-ground segmentation models jointly, where individual models can specialize and complement each other. In recent years, CNN-based approaches (Hayder et al., 2016; Ghodrati et al., 2016; Pont-Tuset and Gool, 2015; He and Lau, 2015) are more popular with a non-trivial margin of performance boost. Jie et al. (Jie et al., 2016) proposed a scale-aware pixel-wise proposal framework where two separate networks are learned to handle

large and small objects, respectively. Pinheiro *et al.* (Pinheiro et al., 2015) devised a discriminative CNN model to generate a mask and predict its likelihood of containing an object. In (Gidaris and Komodakis, 2016), a refine-and-repeat model is formulated to recursively refine the box locations based the revised score from a trained model.

Object detection. Our main pipeline follows the popular object detection prototype (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; Liu et al., 2016; Redmon et al., 2016; Redmon and Farhadi, 2016) of using region proposals as initial point to localize objects and classify them. It is found out that the best practice for detection is to have a set of object boxes passing a RoI-pooling layer, where afterwards we feed the fix-sized feature maps into several additional convolutional layers (He et al., 2016; Kaiming et al., 2014). The final loss is similar to that in region proposal: a classification score and a box location adjustment via regression. Recently, the community has witnessed more advanced versions of the Faster-RCNN detector and its variants. (Dai et al., 2016) took a further step to move the functionality of RoI-pooling to the very last feature layer before the classification: the scheme of position-sensitive score maps addresses the dilemma between invariance in classification and variance in detection under a region-based, fully convolutional network. Dai *et al.* (Dai et al., 2017) motivated that the receptive field of convolutional kernels should also be conditioned on the input RoI boxes and thus formulated the offset reception in an end-to-end learning manner; the deformable unit could be plugged into most popular network structures. Wang *et al.* (Wang et al., 2017) proposed to learn an adversarial network that generates examples with occlusions and deformations, which are hard for the detector to classify. These examples are rare in long-tail distribution and the adversarial part could model such an occurrence.

3 ZIP with Map Attention for Region Proposal

3.1 Backbone Architecture

Figure 3 describes the overview of the proposed zoom out-and-in network with MAD. Most existing network structures (Krizhevsky et al., 2012; He et al., 2016; Simonyan and Zisserman, 2015; Ren et al., 2015; Li et al., 2016) can be viewed and used as a zoom-out network. We adopt the Inception-BN model (Ioffe and Szegedy, 2015) and use the inception module as basic block throughout the paper. Specifically, an image is first fed into three convolutional layers, after which the feature maps are downsampled by a total stride of 8. There are nine inception modules afterwards, denoted as *icp3a-3c*, *icp4a-4e* and *icp5a-5b*. Max-pooling is placed after *icp3c* and *icp4e*. Therefore, we divide the network into three parts based on the stride at 8, 16 and

32. Denote $F^{(m)}$ at depth (or level) m as the output feature maps, where $m = 1, 2, 3$ is the depth index. Note that the spatial size in $F^{(m)}$ is two times the size of that in its subsequent level, which is determined by our network design. The number of feature channels in each block (from *icp3* to *icp5*) are 320, 608 and 1024, respectively.

Inspired by the conv-deconv network in other vision domains (Newell et al., 2016; Noh et al., 2015), we adopt a zoom-in architecture to better leverage the summarized high-level feature maps for refining its low-level counterparts. Such a zoom-in architecture is exactly the mirrored version of the zoom-out part with max-pooling being replaced by deconvolution. We denote the mirrored inception block as *icpX'*. It is found in preliminary experiments that a bilinear upsampling plus a convolution operation achieves better performance than that of an existing deconvolution layer in practice. Denote the output feature maps after the upsampling inception layer *icp4e'* as $H^{(2)}$ and the output maps at layer *icp3'* as $H^{(1)}$. The neuron activations in these feature maps contain the high-level regional semantics, summarized via the zoom-out network and propagated back to guide the feature learning in low-level layer via deconvolution layer by layer. Note that we do not deliberately formulate a $H^{(3)}$ since $F^{(3)}$ alone are suitable enough for identifying large objects.

3.2 Zoom Network Training with MAD Unit

At the core of the zoom out-and-in network is a feature map attention unit to deterministically choose neuron activations among channels from both high and low level streams, and to help supervise the feature learning in the final RPN layer to identify objectness in the image. To this end, we propose a feature map selection unit, called MAD, to actively help find useful neuron activations among feature channels (see red area in Figure 3).

The unit $\text{MAD}(\cdot)$ is fed with the feature maps F^3 as input, goes through three convolutions with kernel size of 3, stride of 1 and the number of channels being exactly two times as that on level m at each layer, where $m = 1, 2$. Since the zoom network is a fully convolution design and the size of input image varies, the spatial size of output after the unit also changes. Therefore, we take a global max-pooling operation to dynamically alters the spatial size to 1×1 . We denote the output of the unit on level m as MAD vector:

$$\mu^{(m)} = \text{MAD}(F^3, m) \in \mathbb{R}^{1 \times 1 \times 2C^{(m)}}, \quad (1)$$

where $C^{(m)}$ is the channel number of the input maps.

Given the inputs from two streams, we first concatenate and denote them as $X^{(m)} = [F^{(m)}, H^{(m)}]$. Then we have

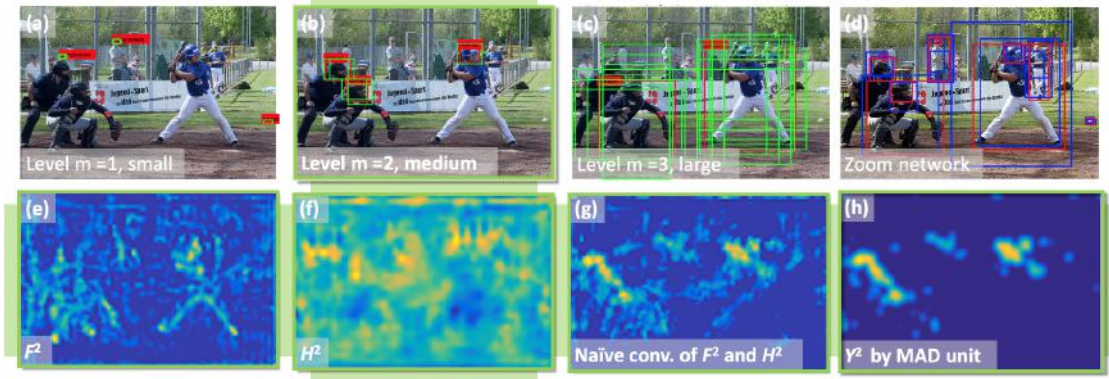


Fig. 4: Effect of MAD unit in learning crucial attention information of features. (a)-(c): Anchor templates (green boxes) with different scales and aspect ratios are placed at different levels in the network, where ground truth annotations are marked in red. (d): Our proposal results (in blue). (e)-(h): Lower neuron activation in F^2 ; Higher neuron activation in H^2 ; naive convolution output from F^2 and H^2 ; attention map using MAD unit.

the attention maps $\mathbf{Y}^{(m)}$, which is an outcome of the concatenated maps multiplied by the MAD vector:

$$\mathbf{y}_j^{(m)} = \mu_j^{(m)} \mathbf{x}_j^{(m)}, \quad (2)$$

where $\mathbf{x}_j, \mathbf{y}_j$ are the vector representation of $\mathbf{X}^{(m)}, \mathbf{Y}^{(m)}$ in channel j , respectively. Note that $\mu_j^{(m)}$ is a learning variable and also involves the parameter update rule. There are many possible alternative designs of MAD unit besides the one stated above and we have included an ablative investigation in the experiments (see Section 5.2).

Now the key question is how to derive the gradient of MAD vector and how to interpret its functionality. For simplicity, we drop notation m and use the vector form of \mathbf{X} and \mathbf{Y} in the following discussion. Suppose we have the upper gradient $\nabla_{\mathbf{y}_j} \mathcal{L}$ flowing back from the loss layer, the gradients with respect to the MAD vector and input are achieved by chain rule²:

$$\frac{\partial \mathcal{L}}{\partial \mu_j} = \nabla_{\mathbf{y}_j} \mathcal{L} \cdot \frac{\partial \mathbf{y}_j}{\partial \mu_j} = \nabla_{\mathbf{y}_j} \mathcal{L} \cdot \mathbf{x}_j, \quad (3)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}_j} = \nabla_{\mathbf{y}_j} \mathcal{L} \cdot \frac{\partial \mathbf{y}_j}{\partial \mathbf{x}_j} = \nabla_{\mathbf{y}_j} \mathcal{L} \cdot \mu_j. \quad (4)$$

The update rule of μ_j and \mathbf{x}_j are defined as follows:

$$\mu_j = \mu_j - \alpha \frac{\partial \mathcal{L}}{\partial \mu_j}, \quad \mathbf{x}_j = \mathbf{x}_j - \alpha \frac{\partial \mathcal{L}}{\partial \mathbf{x}_j}, \quad (5)$$

where $\alpha > 0$ is the learning rate in the unit.

Analysis. The gradient $\nabla_{\mathbf{y}_j} \mathcal{L}$ is the propagated error from the final loss. If the data and parameters of the network follow the update rule, the loss decreases and we say the features are well learned. Keep in mind that weights or

data in the network can increase or decrease as long as such an alternation optimizes the loss, or equivalently, benefits the feature learning. If the direction of map \mathbf{x}_j matches the upper gradient³, $\nabla_{\mu_j} \mathcal{L}$ is positive with μ_j decreased, and based on the update rule of \mathbf{x}_j , it remains a relatively high value, meaning the features does not alter much during update; similarly if the direction of \mathbf{x}_j departs from the upper gradient, $\nabla_{\mu_j} \mathcal{L}$ is negative with μ_j increased, and \mathbf{x}_j changes to a relatively low value since $\nabla_{\mathbf{x}_j} \mathcal{L}$ is large. Therefore, MAD vector μ_j serves as an adapter to alter the magnitude within the network: feature maps whose direction are in accordance with the gradient (thus the optimization goal) should remain unchanged; feature maps whose direction are different from the gradient should alter quickly and remain at a low state (inactive neuron). One advantage of the MAD unit is that it merges feature maps from two streams based on the context of input image, *i.e.*, the vector μ_j are on-the-fly predictions of the network, whereas in previous work, *e.g.*, (Lin et al., 2017), they opt to fixed kernel parameters in the convolutional layer to combine features from different channels during test (see Section 2).

The squeeze-excitation network (SENet) (Hu et al., 2017) demises a block module that first squeezes the spatial size of feature maps along the channel dimension, then extracts the summarized information (adaptive recalibration) and at last imposes the output vector of excitation on the feature maps. The recalibration output resembles the MAD unit. However, the biggest difference lies in the source where MAD, called “excitation recalibration vector” in (Hu et al., 2017), comes from. In their scheme, the map attention decision vector is generated from the very feature map itself; whereas in our design, the MAD decision comes from the higher feature

² The first row is the inner product of two vectors, resulting in a scalar gradient; while the second is the common vector multiplication by a scalar, resulting in a vector also.

³ Direction ‘matches’ means the included angle between two vectors in multi-dimensional space is within 90 degrees; and ‘departs’ means the angle falls at [90, 180].

maps (level $m = 3$) to guide or supervise the feature importance learning in the lower layers (on level $m = 1$ or 2). We deem SENet and our proposed one as concurrent works.

Training loss. After obtaining the attention maps from MAD, we further feed them into a subsequent convolution layer and conduct the anchor detection in the final RPN layer. Let $\mathcal{L}^{(m)}(\mathcal{B}\{\mathbf{y}_j(i)|\mathbf{w}\})$ be the RPN loss on depth m , where \mathcal{B} is the mini-batch; \mathbf{w} denotes the weights in the final RPN layer. $\mathbf{y}_j(i) = \{y_q(i) \in \mathbb{R}^{jd}\}$ denotes the i -th sample from preceding layer. We extend the vector form of \mathbf{y}_j to the element-wise notation y_q for gradient derivation later, i.e., $q = jd$, where d indexes the spatial location in \mathbf{y}_j ; q is the index considering the width, height and channel in the map. The zoom network on level m is trained using the cross-entropy loss plus a regression loss:

$$\mathcal{L}^{(m)}(\mathcal{B}|\mathbf{w}) = \sum_i -\log p_{ik^*} + \delta(l_{ik^*}) \|\mathbf{t}_i - \mathbf{t}_i^*\|^2, \quad (6)$$

where $\delta(\cdot)$ is the indicator function. $\mathbf{p}_i = \{p_{ik}\} \in \mathbb{R}^K$ is the estimated probability with K being the total number of classes. In our task, $k = 0, 1$. $\mathbf{l}_i = \{l_{ik}\}$ denotes the label vector with its correct label index k^* being 1 and others being 0. $\mathbf{t}_i, \mathbf{t}_i^* \in \mathbb{R}^4$ indicates the estimated and ground truth regression offset (Girshick, 2015), respectively. The total loss of the zoom network is optimized as the summed loss across all levels: $\mathcal{L}_{\text{zoom-net}} = \sum_m \mathcal{L}^{(m)}(\mathcal{B}\{\mathbf{y}_j(i)|\mathbf{w}\})$. Note that the number of samples \mathcal{B} in a mini-batch varies in each level, since there are way more anchors in lower depth than those in higher one.

The upper gradient assumed to be existent in Eqn.(3) and (4) is therefore derived as:

$$\begin{aligned} \nabla_{\mathbf{y}_j} \mathcal{L} &= \frac{\partial \mathcal{L}}{\partial y_{jd}} \triangleq \frac{\partial \mathcal{L}}{\partial y_q} = \\ &= \sum_k \frac{\partial \mathcal{L}}{\partial p_k} \frac{\partial p_k}{\partial y_q} + \delta(l_{k^*}) \sum_r \frac{\partial \mathcal{L}}{\partial t_r} \frac{\partial t_r}{\partial y_q}, \\ &= \sum_k (p_k - l_k) w_{qk} + \delta(l_{k^*}) \sum_r (t_r - t_r^*) w_{qr}, \end{aligned} \quad (7)$$

where we drop notations of level m and sample i for brevity. w_{qk} and w_{qr} are the ‘abstract’ weights from input y_q to the loss layer \mathcal{L} . Note that there could be several different layers between the attention maps and the final RPN loss.

Figure 4 visualizes the effect of using MAD unit on acquiring more accurate neuron activations among feature maps. Take the feature maps at level $m = 2$ (for medium objects) as an illustration, we can see the low-level features contain many local details and might be noisy for detection; while raw high-level features in H^2 highlight some region around the helmets, there are still many disturbing areas. A naive combination of high-level and low-level maps by convolution can effectively remove unnecessary background details and magnify objects on medium size. However, we can observe from (h) that by enforcing a MAD unit to actively look

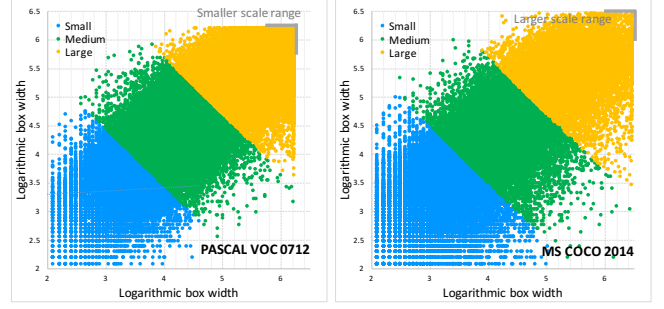


Fig. 5: Anchor distribution in logarithmic width and height. Left: PASCAL VOC 0712 trainval data; right: MS COCO 2014 test data. Region proposals share similar properties across datasets; on COCO the anchors vary slightly more in terms of size and aspect ratio.

for neuron attention among feature channels, the outcome could have clear and strong signal as to where to find potential objects, which could benefit the classification and regression in the RPN layer in a great deal.

3.3 Training Strategies and Anchor Design

There are several remarks regarding the training of the zoom network. First, *adjust input scale dynamically*, known as `dyTrainScale` in Table 2. Each sample is resized to the extent where at least one of the ground truth boxes is covered by the medium anchors. This ensures there are always positive samples in each batch and the model can relatively see a fixed range of anchor sizes - achieving multi-scale training at low computational cost. Second, *control the number of negative samples*, denoted as `equilibrium`. In preliminary experiments, we find that the training loss converges slowly if we fill in the rest of a batch with negative samples. This will cause the unbalance of training data when the number of positive samples is small. Thus, we strict the number of negative samples to be at most twice the number of the positive. Third, *add an additional gray class*, indicated as `grayCls`. Adding an additional gray label into the training stage better separates the positive from the negative. For the positive class, IoU threshold is above 0.6; for the gray class, IoU ranges from 0.35 to 0.55; and for the negative, the criterion is below 0.25. The number of gray samples is set to be half of the total number of positive and negative ones. Note that these strategies are not mentioned in the original region proposal networks (Ren et al., 2015).

As stated in Section 1, we place anchors of various sizes at different depth in the network. Figure 5 depicts the anchor distribution in PASCAL VOC and COCO. The training and test sets share similar data properties *within* and *across* dataset. There are 30 anchors in total and each depth holds

10 box templates. Since the scale problem is handled by the various resolutions in the feature map, we set the number of different aspect ratios (which is 5) more than the number of scales (which is 2 in our setting) in each level. We manually group the anchors into three clusters based on the scale distribution across datasets. Specifically, the scale of anchors in each level are $\{16, 32\}$, $\{64, 128\}$, $\{256, 512\}$, respectively; and the aspect ratio set is 0.25, 0.5, 0.75, 1, 2, 4. Note that these anchor templates are unanimously set across all datasets to be evaluated.

In SSD (Liu et al., 2016), the design of anchors is heuristic and exhaustive: it places too many templates in (relative) size from 0.2 to 0.9 of the fixed input image on six locations. In our work, we bear in mind the computational cost and design the anchor size carefully: we divide the anchor space into three domains and vary the input image size such that there will always be objects of size around 128 (the medium size); the anchors are placed only at three locations (depth) in the network, compared to the six prediction modules in SSD. The size of feature maps in each branch of our method is exactly halved to the previous branch, thus covering a vast majority of anchor space at a lower computation budget.

3.4 Zoom Network Prediction

During inference, we merge proposal results from all the three levels. Prior to merging, an initial NMS (Alexe et al., 2012) process (with top boxes 2000 and IoU threshold 0.5) is launched to first filter out low confidence and redundant boxes. Since the raw scores at each level is not comparable, we tune a bias term for each level to add on the raw scores for better combination among results. Note that these thresholds are cross validated on a smaller set different from the one used for evaluation. The second NMS (with top boxes 300) is conducted in a multi-threshold manner as does in (Gidaris and Komodakis, 2016) to automatically set the optimal IoU threshold under different number of proposals to be evaluated. Since the scale varies dynamically during training, we also forward the network in several image scales, ranging from 1000 to 250 with interval 250.

4 Zoom Network for Object Detection

We now verify our proposed method in the context of object detection and embed the MAD unit, which is for selecting feature maps from various branches in the network, with region proposal generation in an end-to-end learning system. Figure 6 depicts the detection follow-up pipeline. Since the channel number of attended maps on different level varies (640, 1216 and 1024 from level 1 to 3, respectively), we have an additional convolution layer for each level, which transfers the RoI-pooled feature maps to have the same number

of channels after the RoI-pooling (1024 in our setting, not shown in the figure). The network architecture for the detection follow-up layers consists of 9 blocks, corresponding to the higher modules in ResNet-101 model (He et al., 2016), *i.e.*, res4b15 to res4b20, res5a to res5c. The detection loss, denoted as $\mathcal{D}(\mathcal{B}|w)$, resembles Eqn. (6), where the summation traverses over the number of RoIs and the class label k ranges from 0 (background) to the number of total object classes in one dataset. Note that we have the parameter sharing strategy among RoI-pooled features from various levels, which allows us to have more follow-up layers afterwards (crucial to the classification confidence score) in the detection pipeline.

MAD extension and feature boosting. To further generalize the MAD unit and boost the detection performance, we have the following two extensions. In the original formulation (1), the length of the MAD unit is exactly twice as the number of the incoming feature maps, since we merge the two sources from high-level and low-level, respectively. The number of incoming feature maps can be various. Therefore, we extend the formulation of MAD to:

$$\mu^{(m)} = \text{MAD}(\mathbf{F}^3, m) \in \mathbb{R}^{1 \times 1 \times \lambda C^{(m)}}, \quad (8)$$

where λ could be 2, 4, \dots . The combination of map sources could be various and manually designed. Second, we modify the detection confidence score from the neural network output p to the classifier decision of the boosting algorithm. Specifically, for each RoI, we extract features from the follow-up detection blocks (marked in red in Figure 6), concatenate them and feed into the Adaboost classifier, in favor of leveraging different aspects of features in the network and forging into a stronger classifier for detection. Details for the configuration of these two extensions are provided in Section 5.

5 Experiments

We evaluate the effectiveness of our algorithm on generating region proposals and performing generic object detection. To this end, we conduct ablative evaluation of individual components in our approach on PASCAL VOC (Everingham et al., 2015) and compare results with state-of-the-arts on two challenging datasets, ILSVRC DET 2014 (Deng et al., 2009) and Microsoft COCO (Lin et al., 2014).

5.1 Datasets and Setup

The PASCAL VOC dataset (Everingham et al., 2015) is designed for object detection and contains 20 object classes. To fast verify our algorithm’s design, we conduct experiments on this dataset, where we use the training and validation set in 2007 and 2012 (16551 images) during training and evaluate on the 2007 test set (4952 images). The ILSVRC DET

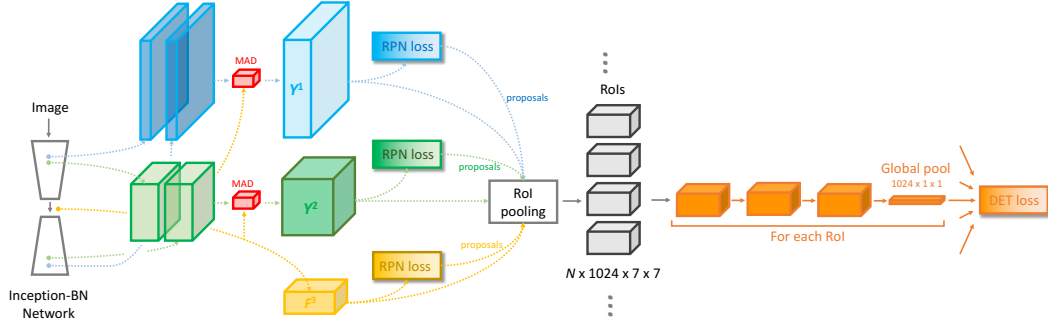


Fig. 6: The complete pipeline of ZIP and MAD algorithm for region proposal and object detection. The system can be trained end-to-end. Based on the proposals generated from the zoom network, we conduct a RoI-pooling on each level after the MAD unit. N is the number of RoIs in one mini-batch. Note that the detection follow-up layers (marked in orange after RoI-pooling) share the *same* parameter across different levels, which is different from Fig. 3 (b).

2014 dataset (Deng et al., 2009) is a subset of the whole ImageNet database and consists of more than 170,000 training and 20,000 validation images. Since some training images has only one object with simple background, which has a distribution discretion with the validation set, we follow the practice of (Girshick, 2015) and split the validation set into two parts. The training set is the `train_14` with 44878 images and `val1` with 9205 images. We use `val2` as the validation set for evaluation. The Microsoft COCO 2014 dataset (Lin et al., 2014) contains 82,783 training and 40,504 validation images, where most images have various shapes surrounded by complex scenes. We use all the training images, without any data augmentation, to learn our model, and follow (Pinheiro et al., 2015) to evaluate on the first 5000 validation images (denoted as `val_5k`).

Implementation details. We reimplement an Inception BN (Ioffe and Szegedy, 2015) model on the ImageNet classification dataset, which could achieve around 94% top-5 classification accuracy. The zoom network is finetuned using the pretrained model and the zoom-in part is also initialized by coping the weights from its mirror layers. The convolution layers in MAD unit is trained from scratch. The base learning rate is set to 0.0001 with a 50% drop every 7,000 iterations. The momentum and weight decay is set to be 0.9 and 0.0005, respectively. The maximum training iteration are roughly 8 epochs for each dataset. We utilize stochastic gradient descent for optimization. The mini-batch size B during each iteration is set to be 300 with each class having at most 100 samples. Note that during training, there are always some cases where the actual batch size is lower than the pre-setting due to the sample equilibrium scheme stated in Section 3.3. The network architecture in the following experiments are derived (zoom-out part) and mirrored (zoom-in part) from the Inception-BN model, which is explicitly stated in Section 3.1.

Evaluation metric. We use recall (correctly retrieved ground truth boxes over all annotations) under different IoU

Table 1: Ablation on the network structure. AR on PASCAL VOC is reported. We use 30 anchors and treat training as a two-class problem. ‘sp’ means splitting the anchors. Four out of five cases below (a, b, d, e) are illustrated in Figure 2.

Structure	AR@100	AR@S	AR@M	AR@L
(a) Zoom-out, F-RCNN	62.31	40.77	57.97	71.05
(b) Zoom-out_sp	65.98	43.81	60.35	73.87
(c) Zoom-out_sp, all	66.02	44.52	61.77	74.13
(d) ZIP	68.51	49.07	62.93	75.64
(e) Deeper Zoom-out_sp	67.21	47.92	61.53	74.02

thresholds and number of proposals as metric to evaluate region proposals; and precision (correctly identified boxes over all predictions) to evaluate object detection. The mean value of recall and precision from IoU 0.5 to 0.95 is known as average recall (AR) and precision (AP). AR summarizes the general proposal performance and is shown to correlate with the average precision (AP) performance of a detector better than other metrics (Hosang et al., 2015). Moreover, we compute AR and AP of different sizes of objects to further investigate on a specific scale of targets. Following COCO-style evaluation, we denote three types of instance size, @small ($\alpha < 32^2$), @medium ($32^2 \leq \alpha < 96^2$) and @large ($96^2 \leq \alpha$), where α is the area of an object.

5.2 Ablative Evaluation for Region Proposals

All the experiments in this subsection are conducted on the PASCAL VOC dataset for the region proposal task.

Network structure design. Table 1 reports AR for different network design. All images are fed into the network with a fixed size on shorter dimension of 600. The anchor design is mentioned in Section 3.3. Other settings are by default as that in (Ren et al., 2015). There are some observations: (a) if we employ a zoom out structure and place

Table 2: Ablation study on MAD unit and training strategies. AR on PASCAL VOC is reported. We use the zoom-out-and-in network structure. Baseline model corresponds to the (d) setting in Table 1. `all` means adopting all three training strategies. ER vector denotes the Excitation Recalibration idea implemented in SENet (Hu et al., 2017). See context in the paper for setting details.

Strategy	AR@100	AR@S	AR@M	AR@L
Baseline	68.51	49.07	62.93	75.64
dyTrainScale	72.58	53.41	66.34	78.83
equilibrium	69.85	50.49	63.55	76.88
grayCls	69.77	50.21	63.19	77.04
ZIP + all	74.22	54.39	68.47	81.53
ZIP + all + MAD	76.51	57.28	70.05	84.21
spatial MAD	76.50	57.11	70.13	84.09
ER vector	71.87	51.84	65.49	76.21

all anchors at the last feature map, the recall is 62.31%, which is adopted by most popular detectors, *e.g.*, FasterRCNN (Ren et al., 2015); (b) then we split the anchors into three groups and insert them at layer `icp_3b`, `icp_4d` and `icp_5b`. Such a modification will increase recall to 65.98%. It indicates that extracting object proposals at different depth helps; (c) if we consider another comparison where all anchors are placed at each desired location in the network, *i.e.*, `icp_3b`, `icp_4d` and `icp_5b`, the average recall is 66.12. The result is better than (b) because it places all anchors, no matter big or small, densely into the network - boosting the search space of object boxes. However, the computational cost of (c) is around 3.5 times than (b) due to the heavy burden on dense anchors in lower layers; (d) the zoom-out-and-in design (ZIP) further increases performance of recall to 68.51%. ZIP increases the network depth and has about 40 layers; (e) by simply stacking layers of the network to 40 layers via a zoom-out design, we do not witness an obvious increase compared with ZIP, which verifies that the gain in our method does not come from depth or parameter increase of the model. Compared with the baseline zoom-out structure, ZIP achieves a larger improvement of AR on small objects (8.3%) than medium-sized objects (4.96%) and large objects (4.59%). Such a zoom-out-and-in structure could enhance the quality of proposals in all ranges of scales.

Training strategies. Table 2 shows the effect of MAD unit and different training strategies stated in Section 3. Starting from the baseline, we add each training component individually and investigate their contribution to AR. It is observed that the dynamic alternation of input scale weighs more among these strategies. The overall AR improvement of merging all strategies could obtain a recall of 74.22%. This result is the outcome of imposing a simple convolution kernel to combine features from two streams (qualitatively

visualized in Figure 4(g)). Furthermore, the MAD unit can actively attend important neuron activations and we verify its effectiveness via the last experiment: enhancing AR to 76.51%.

Other alternatives to MAD. A natural thought could be to skip the global pooling after the convolution operation at the output of `icp5`. The MAD vector $\mu^{(m)}$ could be spatially sized and proportional to each input image and level depth. In help of deconvolution, we can align spatially the MAD vector with feature maps to be weighted. In this manner, each pixel, instead of all pixels on a channel, could be evaluated in terms of importance towards the final loss. Such a design is denoted as ‘spatial MAD’ in Table 2. We can see the performance is similar compared to the global pooling version. Since it could increase the computational cost by adding more parameters in the model, we do not resort to this alternative. Such a result indicates that for weighing the importance of feature maps, the learned information along the channel dimension is more important, or at least enough, than the information learned along both the channel and spatial dimensions. We also try the excitation recalibration idea to replace MAD. Specifically, $Y^{(m)}$ is generated from the squeeze-and-excitation unit alone. For $m = 1, 2$, the input is the concatenation of $F^{(m)}$ and $H^{(m)}$; for level $m = 3$, the input is $F^{(3)}$. As is shown in the last case of Table 2, the replacement is inferior to ours (71.87 vs 76.51). This is probably because that the excitation vector derives from information directly in the preceding layer - an inherent design in (Hu et al., 2017); whilst our MAD comes from higher-level summation, leveraging the guidance of high-level, region-based semantics to merge local details in lower layers.

Runtime analysis. During training the image is resized to a much smaller size based on the medium anchor size (64 – 128) and thus the total training time is one third less than that without such a constraint. For inference, the runtime per image is 1.13s (Titan X, code partially optimized), compared with AttractionNet (Gidaris and Komodakis, 2016) 1.63s and DeepMask (Pinheiro et al., 2015) 1.59s.

In the following experiments, we denote our method as ‘ZIP’ and by default it includes MAD if not specified.

5.3 Average Recall for Region Proposals

Figure 7 illustrates recall under different IoU thresholds and number of proposals. Our algorithm is superior than or on par with previous state-of-the-arts, including: BING (Cheng et al., 2014), EdgeBox (Zitnick and Dollar, 2014), GOP (Krahenbuhl and Koltun, 2014), Selective Search (Uijlings et al., 2013), MCG (Arbeláez et al., 2014), Endres (Endres and Hoiem, 2014), Prims (Manén et al., 2013), Rigor (Humayun et al., 2014), Faster RCNN (Ren et al., 2015), AttractionNet (Gidaris and Komodakis, 2016), DeepBox (Kuo et al., 2015), CoGen (Hayder et al., 2016), DeepMask (Pinheiro et al.,

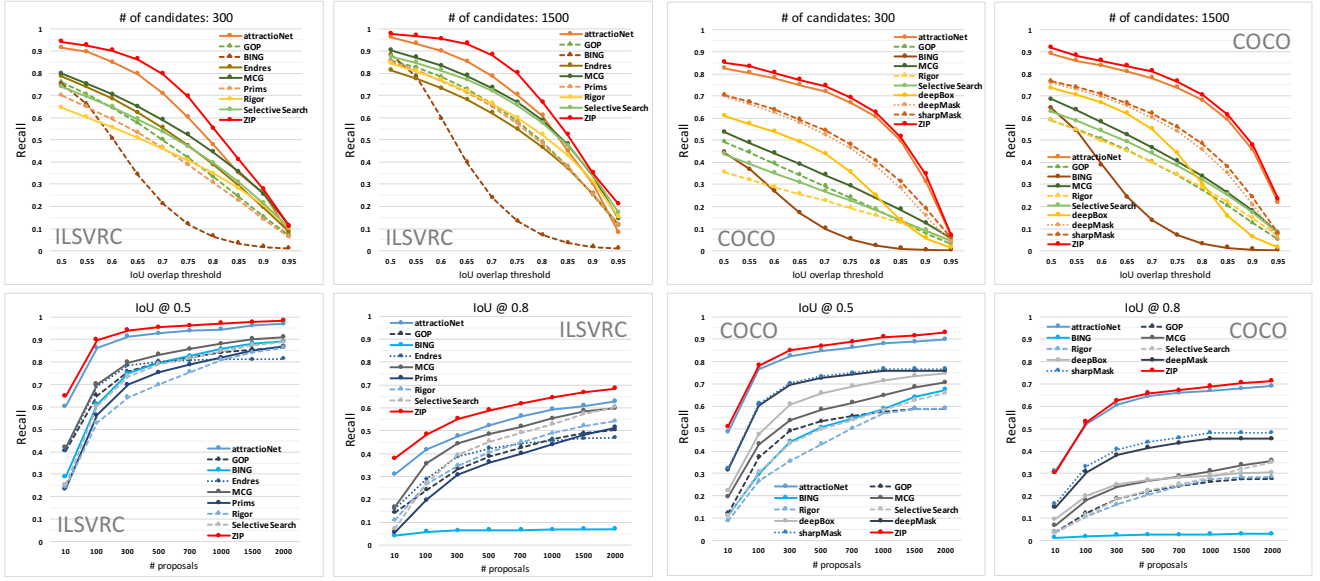


Fig. 7: Recall at different IoU thresholds (0.5, 0.8) and number of proposals (300, 1500). Left: ILSVRC DET 2014 val2. Right: MS COCO 2014 val_5k. For method abbreviations, please refer to the context.

2015), SharpMask (Pinheiro et al., 2016), and FPN (Lin et al., 2017). Table 3 reports the average recall vs. the number of proposals (from 10 to 1000) and the size of objects on ILSVRC. Our method performs better on identifying different size of objects. This proves that the zoom-out-and-in structure with a MAD unit is effective. For MS COCO, similar phenomenon of AR is also observed via Table 4. We include results from the latest work (Lin et al., 2017) which shares similar spirit of a conv-deconv structure and combining feature maps from different sources. The main distinction from ours to (Lin et al., 2017) is the introduction of MAD unit to dynamically decide the attention weights of feature channels based on the context of the input image.

Compared with Faster RCNN (Ren et al., 2015) (which also employs CNN, has a zoom-out structure with all anchors placed at the final layer and thus serves as our baseline), ZIP has achieved a very large improvement over the baseline on small objects (21.8%) and medium objects (22.4%). The performance improvement for large objects is also satisfying (8.4%). AttractionNet (Gidaris and Komodakis, 2016) also has large improvements on small and medium object proposals, since it employs the recursive regression and active box generation. These techniques are also beneficial for detecting small objects.

Generalization for unseen categories. In order to evaluate the generalization capability, we show the average recall @10 boxes in Table 5, where our ZIP model is trained on COCO and tested on ILSVRC val2 set. Although COCO has fewer classes, which is 80, the scales and aspect ratios of objects varies in great extent and thus the trained ZIP model

can handle most categories on ILSVRC DET dataset, which has 200 classes.

5.4 Ablative Evaluation for Object Detection

All the experiments in this subsection are conducted on the PASCAL VOC dataset for object detection. The backbone structure (inception-BN) for detection is the same as that in region proposal. We have conducted experiments based on the VGG-16 model and the performance is slightly inferior (76.4%, *c.f.* 76.8% denoted as ‘ZIP MAD’ in Table 6). Hence we employ the inception model thereafter.

Table 6 reports the detailed ablation investigation on the object detection task. For reference, we also list the results from popular methods, RFCN (Dai et al., 2016), SSD (Liu et al., 2016), Faster-RCNN (Ren et al., 2015), in the first three rows. We can see the zoom-in architecture alone could enhance the performance in a great deal (*c.f.*, zoom-out and ZIP baseline, around 2%). The MAD unit could further improve the result by leveraging feature weights among different channels in the network. Invoking the boosting method alone into the pipeline (denoted as ‘ZIP MAD B’, using all features in the nine blocks) could enhance the detection performance to some extent. Empirically we find the value of λ in Eqn. 8 to be 4 best fits the task, which is noted as ‘ZIP MAD E’ in the table. At last, after incorporating both the boosting and extension schemes, we have the final result on PASCAL to be 79.8% mAP; such an improvement bears from the zoom-in structure, the auto-selected MAD unit and the stronger classifier from boosting.

Table 3: Average recall (AR) analysis on ILSVRC val12. The AR for small, medium and large objects are computed for 100 proposals. The top two results in each metric are in **bold** and *italic*, respectively.

ILSVRC DET 2014	AR@10	AR@100	AR@500	AR@1000	AR@Small	AR@Medium	AR@Large
BING	0.114	0.226	0.287	0.307	0.000	0.064	0.340
EdgeBox	0.188	0.387	0.512	0.555	0.021	0.156	0.559
GOP	0.208	0.349	0.486	0.545	0.022	0.185	0.482
Selective Search	0.118	0.350	0.522	0.588	0.006	0.103	0.526
MCG	0.229	0.435	0.553	0.609	0.050	0.215	0.604
Endres	0.221	0.393	0.508	0.531	0.029	0.209	0.543
Prims	0.101	0.296	0.456	0.523	0.006	0.077	0.449
Rigor	0.139	0.325	0.463	0.551	0.027	0.092	0.485
Faster RCNN	0.356	0.475	0.532	0.560	0.217	0.407	0.571
AttractionNet	<i>0.412</i>	<i>0.618</i>	<i>0.672</i>	<i>0.748</i>	<i>0.428</i>	<i>0.615</i>	<i>0.623</i>
Zoom Network (ZIP)	0.420	0.635	0.688	0.761	0.435	0.631	0.655

Table 4: Average recall (AR) analysis on COCO val5k. The AR for small, medium and large objects are computed for 100 proposals. The top two results in each metric are in **bold** and *italic*, respectively.

MS COCO 2014	AR@10	AR@100	AR@500	AR@1000	AR@Small	AR@Medium	AR@Large
BING	0.042	0.100	0.164	0.189	0.000	0.026	0.269
EdgeBox	0.074	0.178	0.285	0.338	0.009	0.086	0.423
GOP	0.058	0.187	0.297	0.339	0.007	0.141	0.401
Selective Search	0.052	0.163	0.287	0.351	0.003	0.063	0.407
MCG	0.098	0.240	0.342	0.387	0.036	0.173	0.497
Endres	0.097	0.219	0.336	0.365	0.013	0.164	0.466
DeepBox	0.127	0.270	0.376	0.410	0.043	0.239	0.511
CoGen	0.189	0.366	-	0.492	0.107	0.449	0.686
DeepMask	0.183	0.367	0.470	0.504	0.065	0.454	0.555
SharpMask	0.196	0.385	0.489	0.524	0.068	0.472	0.587
FPN	-	0.440	-	0.563	-	-	-
AttractionNet	<i>0.328</i>	<i>0.533</i>	<i>0.601</i>	<i>0.662</i>	<i>0.315</i>	<i>0.622</i>	<i>0.777</i>
Zoom Network (ZIP)	0.335	0.539	0.612	0.670	0.319	0.630	0.785

Table 5: Generalization ability of ZIP when the training and test set are from different datasets. The test set is fixed as ILSVRC. When the training set is changed from ILSVRC to COCO, AR@10 only drops slightly. Definition of seen and unseen categories are the same with AttractionNet (Gidaris and Komodakis, 2016).

Train Data	All	Seen	Unseen
ZIP on ImageNet	42.0	48.5	33.3
ZIP on COCO	41.9	48.23	31.7
AttractionNet on COCO	41.2	47.41	29.9
MCG on COCO	21.9	22.8	20.5

300 proposals - only the region proposals are different. It can be seen that our object proposal provides a better mAP (AP@0.5) on the ILSVRC and are suitable for detecting objects of different sizes. We reimplement the results of RFCN in the table using RPN proposals, which is to ensure that the AP performance difference descends from region proposals only. It is worth noticing that although there a big gain in region proposals from ours compared with previous methods (ours vs. SS, around 30% recall boost), the gain in terms of object detection is not that obvious (around 2%). We believe the discrepancy descends from the classification inaccuracy.

5.5 Average Precision for Object Detection

Comparison using the R-FCN detector. We follow the detection pipeline as that in the R-FCN detector (Dai et al., 2016), use the ResNet-50 (He et al., 2016) model and run the public code. The batch size is 4 and the total iteration is 220,000. Table 7 reports the detection performance of different proposal methods using the *same* detector at top

5.6 One Stage vs Two Stage

We also embed the spirit of deconvolution and MAD unit in a single stage manner, where the SSD detector is invoked as the prototype. Compared with the 76.8% mAP (SSD in Table 6) baseline on PASCAL, the single stage trained detector with our design achieves better performance at 78.1%, we believe the enhancement derive from better feature rep-

Table 6: Ablation study for object detection on the PASCAL VOC 2007 test set. The training data comes from both 2007 and 2012 training and validation set. ‘Zoom-out’ indicates that the network structure is the standard convolution type without up-sampling. E means MAD extension; B means boosting and EB denotes the combination.

method	data	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster R-CNN	07+12	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	56.7	87.8	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72.0
SSD512	07+12	76.8	82.4	84.7	78.4	73.8	53.2	86.2	87.5	86.0	57.8	83.1	70.2	84.9	85.2	83.9	79.7	50.3	77.9	73.9	82.5	75.3
R-FCN	07+12	79.5	82.5	83.7	80.3	69.0	69.2	87.5	88.4	88.4	65.4	87.3	72.1	87.9	88.3	81.3	79.8	54.1	79.6	78.8	87.1	79.5
Zoom-out	07+12	71.3	72.1	77.3	73.4	62.8	50.1	79.6	81.2	82.7	49.6	82.6	65.4	83.7	81.4	73.5	75.4	38.3	74.1	73.6	81.1	68.5
ZIP baseline	07+12	73.1	73.8	79.2	75.8	64.2	53.7	81.3	83.0	85.2	51.4	85.7	66.2	84.4	83.7	75.1	76.8	40.5	76.2	75.4	83.1	67.9
ZIP MAD	07+12	76.8	74.1	80.1	76.0	65.7	55.4	83.5	85.3	87.5	54.2	87.3	68.6	87.4	85.2	77.2	77.7	40.9	77.8	77.5	84.6	70.3
ZIP MAD _B	07+12	77.1	75.6	88.2	76.1	66.7	57.9	86.2	86.7	90.5	56.8	88.3	69.3	89.9	90.3	79.4	78.5	45.6	78.0	78.2	85.1	75.6
ZIP MAD _E	07+12	77.4	75.8	87.5	76.4	67.0	58.1	86.6	87.2	90.1	57.2	89.1	72.5	87.3	91.2	80.1	78.9	46.0	78.5	78.2	85.4	75.8
ZIP MAD _{EB}	07+12	79.8	79.8	85.5	78.5	72.1	57.7	88.8	88.1	93.1	58.9	93.7	73.3	91.8	92.4	84.9	81.2	47.5	81.3	81.5	88.3	75.5

Table 7: Average precision (AP) for object detection on the ILSVRC val2 set. To evaluate the performance of different proposals, we use the *same* R-FCN detector (Dai et al., 2016) across approaches. Note that the first two methods are listed for reference only since they have their own network structures and detection pipelines.

Method	AP@0.50	AP@0.75	AP@0.5:0.95
SSD	43.4	-	-
ResNet	60.5	-	-
EdgeBox	49.94	35.24	31.47
Selective Search	51.98	35.13	34.22
AttractionNet	52.07	35.93	35.47
RFCN	53.11	36.41	35.63
ZIP	54.08	36.72	35.66
Method	AP@small	AP@medium	AP@large
EdgeBox	3.98	20.61	60.98
Selective Search	4.86	23.24	61.78
AttractionNet	4.90	26.37	63.11
RFCN	4.93	26.81	64.77
ZIP	4.97	27.04	65.36

resentation by deconvolution and the automatic feature selection via MAD unit. However, compared to our final two-stage detector, which has a 79.8% mAP, the single stage counterpart is inferior. In the two-stage scheme, region proposals are uniquely generated and separated from the second stage, which provides a more accurate set of candidate boxes (hypothesis) for the detection process in the second stage. Such an observation is in accordance with findings in (Huang et al., 2017).

However, we still believe the single-stage spirit of object detection will dominate the field and will be prevalent. The one-stage framework is trained end-to-end without con-

sidering the two tasks separately and seems more neat and unified. We leave such an investigation as future work.

6 Conclusions

In this work, we devise a zoom-out-and-in network that both utilizes low-level details and high-level semantics. The information from top layers is gradually up-sampled by deconvolution to reach suitable resolution for small-sized objects. Such a strategy could alleviate the drawback of identifying small objects on feature maps with a large stride. We further propose a map attention decision (MAD) unit to actively search for neuron activations and attend to specific maps that could weigh more during the feature learning in the final RPN layer. Several training strategies are also proposed and investigated to enhance the quality of region proposals. Experiments for both the region proposal generation and object detection tasks show that our proposed algorithm (ZIP with map attention decision) performs superior against previous state-of-the-arts on popular benchmarks, including PASCAL VOC 2007, ImageNet DET and MS COCO.

Acknowledgment

We would like to thank reviewers for helpful comments, S. Gidaris, X. Tong and K. Kang for fruitful discussions along the way, W. Yang for proofreading the manuscript. H. Li is funded by the Hong Kong Ph.D. Fellowship scheme. We are also grateful for SenseTime Group Ltd. donating the resource of GPUs at time of this project.

References

Alexe, B., Deselaers, T., and Ferrari, V. (2012). Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2189–2202.

- Arbeláez, P., Pont-Tuset, J., Barron, J., Marques, F., and Malik, J. (2014). Multiscale combinatorial grouping. In *CVPR*.
- Bell, S., Zitnick, C. L., Bala, K., and Girshick, R. (2016). Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*.
- Chavali, N., Agrawal, H., Mahendru, A., and Batra, D. (2016). Object-proposal evaluation protocol is 'gameable'. In *CVPR*.
- Cheng, M., Zhang, Z., Lin, W., and Torr, P. H. S. (2014). BING: binarized normed gradients for objectness estimation at 300fps. In *CVPR*.
- Chi, Z., Li, H., Lu, H., and Yang, M.-H. (2016). Dual deep network for visual tracking. *arXiv preprint: 1612.06053*.
- Dai, J., Li, Y., He, K., and Sun, J. (2016). R-FCN: Object Detection via Region-based Fully Convolutional Networks. In *NIPS*.
- Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In *arXiv preprint: 1703.06211*.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*.
- Endres, I. and Hoiem, D. (2014). Category-independent object proposals with diverse ranking. *IEEE Trans. on PAMI*, 36:222–234.
- Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136.
- Fu, C.-Y., Liu, W., Ranga, A., Tyagi, A., and Berg, A. C. (2017). Dssd : Deconvolutional single shot detector. In *arXiv preprint: 1701.06659*.
- Ghodrati, A., Diba, A., Pedersoli, M., Tuytelaars, T., and Gool, L. V. (2016). DeepProposals: Hunting objects and actions by cascading deep convolutional layers. *arXiv preprint: 1606.04702*.
- Gidaris, S. and Komodakis, N. (2016). Attend Refine Repeat: Active box proposal generation via in-out localization. In *BMVC*.
- Girshick, R. (2015). Fast R-CNN. In *ICCV*.
- Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*.
- Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., and Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, advance online publication.
- Hariharan, B., Arbelaz, P., Girshick, R., and Malik, J. (2014). Hypercolumns for object segmentation and fine-grained localization. In *CVPR*.
- Hayder, Z., He, X., and Salzmann, M. (2016). Learning to co-generate object proposals with a deep structured network. In *CVPR*.
- He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
- He, S. and Lau, R. W. (2015). Oriented object proposals. In *ICCV*.
- Hosang, J., Benenson, R., Dollár, P., and Schiele, B. (2015). What makes for effective detection proposals? *IEEE Trans. on PAMI*.
- Hu, J., Shen, L., and Sun, G. (2017). Squeeze-and-excitation networks. In *arXiv preprint: 1709.01507*.
- Huang, J., Rathod, V., Sun, C., Zhu, M., Korattikara, A., Fathi, A., Fischer, I., Wojna, Z., Song, Y., Guadarrama, S., and Murphy, K. (2017). Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*.
- Humayun, A., Li, F., and Rehg, J. M. (2014). Rigor: Reusing inference in graph cuts for generating object regions. In *CVPR*.
- Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*.
- Jie, Z., Liang, X., Feng, J., Lu, W. F., Tay, E. H. F., and Yan, S. (2016). Scale-aware pixelwise object proposal networks. *IEEE Trans. on Image Processing*, 25.
- Kaiming, H., Xiangyu, Z., Shaoqing, R., and Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*.
- Kong, T., Yao, A., Chen, Y., and Sun, F. (2016). Hypernet: Towards accurate region proposal generation and joint object detection. In *CVPR*.
- Krahenbuhl, P. and Koltun, V. (2014). Geodesic object proposals. In *ECCV*.
- Krahenbuhl, P. and Koltun, V. (2015). Learning to propose objects. In *CVPR*.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114.
- Kuo, W., Hariharan, B., and Malik, J. (2015). DeepBox: Learning objectness with convolutional networks. In *ICCV*.
- Li, H., Liu, Y., Ouyang, W., and Wang, X. (2017a). Zoom out-and-in network with recursive training for object proposal. In *arXiv preprint: 1702.05711*. arXiv.
- Li, H., Liu, Y., Zhang, X., An, Z., Wang, J., Chen, Y., and Tong, J. (2017b). Do we really need more training data

- for object localization. In *IEEE International Conference on Image Processing*. IEEE.
- Li, H., Ouyang, W., and Wang, X. (2016). Multi-bias non-linear activation in deep neural networks. In *ICML*.
- Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). Feature pyramid networks for object detection. *CVPR*.
- Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollar, P. (2014). Microsoft COCO: Common Objects in Context. *arXiv preprint:1405.0312*.
- Liu, W., Anguelov, D., Erhan, D., Szegedy, C., and Reed, S. (2016). SSD: Single shot multibox detector. In *ECCV*.
- Liu, Y., Li, H., and Wang, X. (2017a). Learning deep features via congenerous cosine loss for person recognition. In *arXiv preprint: 1702.06890*. arXiv.
- Liu, Y., Li, H., Yan, J., Wei, F., Wang, X., and Tang, X. (2017b). Recurrent scale approximation for object detection in cnn. In *IEEE International Conference on Computer Vision*.
- Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.
- Manén, S., Guillaumin, M., and Van Gool, L. (2013). Prime Object Proposals with Randomized Prim's Algorithm. In *ICCV*.
- Newell, A., Yang, K., and Deng, J. (2016). Stacked hour-glass networks for human pose estimation. In *ECCV*.
- Noh, H., Hong, S., and Han, B. (2015). Learning deconvolution network for semantic segmentation. In *ICCV*.
- Pinheiro, P. O., Collobert, R., and Dollar, P. (2015). Learning to segment object candidates. In *NIPS*.
- Pinheiro, P. O., Lin, T.-Y., Collobert, R., and Dollr, P. (2016). Learning to refine object segments. In *ECCV*.
- Pont-Tuset, J. and Gool, L. V. (2015). Boosting object proposals: From pascal to coco. In *CVPR*.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *CVPR*.
- Redmon, J. and Farhadi, A. (2016). Yolo9000: Better, faster, stronger. In *arXiv preprint: 1612.08242*.
- Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint: 1505.04597*.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., and LeCun, Y. (2014). Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*.
- Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.
- Sun, C., Paluri, M., Collobert, R., Nevatia, R., and Bourdev, L. (2016). ProNet: Learning to propose object-specific boxes for cascaded neural networks. In *CVPR*.
- Uijlings, J., van de Sande, K., Gevers, T., and Smeulders, A. (2013). Selective search for object recognition. *International Journal of Computer Vision*.
- Wang, X., Shrivastava, A., and Gupta, A. (2017). A-fast-rcnn: Hard positive generation via adversary for object detection. *CVPR*.
- Wen, Y., Zhang, K., Li, Z., and Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *ECCV*.
- Zitnick, L. and Dollar, P. (2014). Edge Boxes: Locating object proposals from edges. In *ECCV*.