Small Object Detection Leveraging on Simultaneous Super-resolution

Hong Ji[1], Zhi Gao[1], Xiaodong Liu[2], Yongjun Zhang[1], Tiancan Mei[3]

1 School of Remote Sensing and Information Engineering, Wuhan University, 430079 Wuhan, China - (2013301220036, zhangyj)@whu.edu.cn, gaozhinus@gmail.com
2 Department of Electrical and Computer Engineering, National University of Singapore, 117583 , Singapore - xiaodongliu@u.nus.edu
3 School of Electronic Information, Wuhan University, 430079 Wuhan, China - mtcwlb@aliyun.com

Abstract

Despite the impressive advancement achieved in object detection, the detection performance of small object is still far from satisfactory due to the lack of sufficient detailed appearance to distinguish it from similar objects. Inspired by the positive effects of super-resolution for object detection, we propose a framework that can be incorporated with detector networks to improve the performance of small object detection, in which the low-resolution image is super-resolved via generative adversarial network (GAN) in an unsupervised manner. In our method, the super-resolution network and the detection network are trained jointly. In particular, the detection loss is back-propagated into the SR network during training to facilitate detection. From Figure 1. we conclude that super-resolution is crucial for detection.

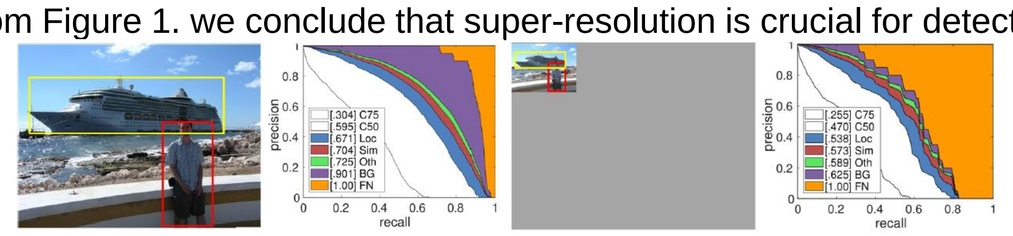

Figure 1. Example of LR image (bottom-left) and its HR counterpart (up-left)and the overall error analysis of the Faster R-CNN++ detector trained on LR images (bottom-right) and HR images (up-right). The comparison demonstrates the large gap between the detection performance of HR images and LR images.

Method & Framework

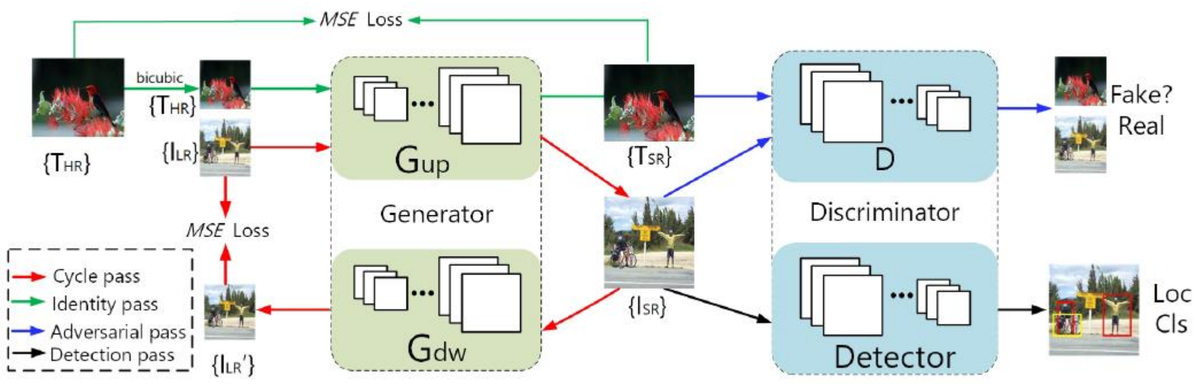

Figure 2. Flowchart of the proposed coarse-to-fine registration framework for multi/hyperspectral images.

The pipeline of our proposed method is shown in Figure 2. The processing starts by forwarding the original LR image. I_{LR} is the input LR image, I_{SR} is the super-resolved HR image from I_{LR}, I'_{LR} is of LR generated from I_{SR}. T_{HR} is the HR image provided as reference from other high-quality dataset. T_{LR} is down-sampled version of T_{HR}. T_{SR} is the super-resolved HR image from T_{LR}.

CycleGAN-based SR network

CycleGAN strategy is shown as up figure of Figure. 3. we use CycleGAN-like architecture for image super-resolution, which is shown as bottom of Figure. 3. Training loss is defined as:

$$\mathcal{L}_{cycGAN} = \mathcal{L}_{GAN} + \lambda_1 \mathcal{L}_{cyc} + \lambda_2 \mathcal{L}_{Idt}$$

where:

$$\mathcal{L}_{GAN} = E_{I_{HR}\sim P_{data}(I_{HR})}[\log(D(I_{HR}))] + E_{I_{LR}\sim P_{data}(I_{LR})}[\log(1 - DG_{UP}(I_{LR}))]$$

$$\mathcal{L}_{cyc} = E_{I_{LR}\sim P_{data}(I_{LR})}[\|G_{dw}(G_{UP}(I_{LR})) - I_{LR}\|_2]$$

$$\mathcal{L}_{Idt} = E_{T_{LR}\sim P_{data}(T_{LR})}[\|G_{UP}(T_{LR}) - T_{HR}\|_2]$$

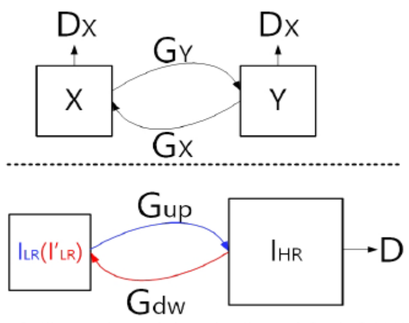

Figure 3. Illustration of the pipeline of CycleGAN(up), CycleGAN-like SR network(bottom).

Discriminator networks

Firstly, architecture of discriminator network D is shown in Table 1, which if for distinguishing the real HR images from the generated super-resolved images.

Secondly, we employ detector as another discriminator for object localization and classification. We study the naive Faster R-CNN using VGG16 as backbone and predicts objects in the last convolutional layer. The training loss is defined as:

$$\mathcal{L}_{Det} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg}$$

where:

$$\mathcal{L}_{cls} = E_{I_{LR}\sim P_{data}(I_{LR})}[-\log(Det_{cls}(G_{UP}(I_{LR})))]$$

$$\mathcal{L}_{reg} = E_{I_{LR}\sim P_{data}(I_{LR})}[smooth_{L_1}(Det_{loc}(G_{UP}(I_{LR})), \boldsymbol{t}_*)]$$

layer	conv	conv	BN	conv	BN	conv	BN	conv
Kernel size	4	4	-	4	-	4	-	4
Kernel num	64	128	-	256	-	512	-	1
stride	2	2	-	2	-	1	-	1

Table 1. Architecture of discriminator D.

Finally, we train the generators and discriminators alternatively

Experiment & Result

Image Data

We first perform experiments on PASCAL VOC that has 20 object categories. We train all the models on VOC 2012 trainval and VOC 2007 trainval respectively, and perform inference on their corresponding test datasets, VOC 2012 test (11k) and VOC 2007 test (5k) respectively. To demonstrate the effect of our work, we down-sample the PASCAL VOC datasets using bicubic kernel to generate LR images. We focus on the resulting detection accuracy in terms of mAP, and consider the efficiency issue in future. Therefore, we apply the Faster R-CNN networks, both its basic version and its improved version which are termed as Naive Faster R-CNN and Faster R-CNN++ respectively, as our detectors.

Comparisons and results

We compare our framework with other settings:

- FASR and FASR++ are with Naïve Faster R-CNN and its higher version
- Original/FASR: results on the original (gt) high-resolution images
- Bicubic/FASR, EDSR/FASR, CysSR/FASR represent results on the SR images obtained using bicubic, EDSR, and CycleGAN respectively

Figure 4. Examples of detection results.

method \ dataset(mAP)	VOC2007	VOC2012
Original/FASR++	0.795	0.738
Bicubic/FASR++	0.595	0.544
EDSR/FASR++	0.646	0.592
Cyc-SR/FASR++	0.665	0.607
Ours/FASR++	**0.669**	**0.637**
Original/FASR	0.759	0.695
Bicubic/FASR	0.535	0.485
EDSR/FASR	0.605	0.550
Cyc-SR/FASR	0.624	0.566
Ours/FASR	**0.658**	**0.594**

Table 2. Detection results of different methods (IoU=0.5).

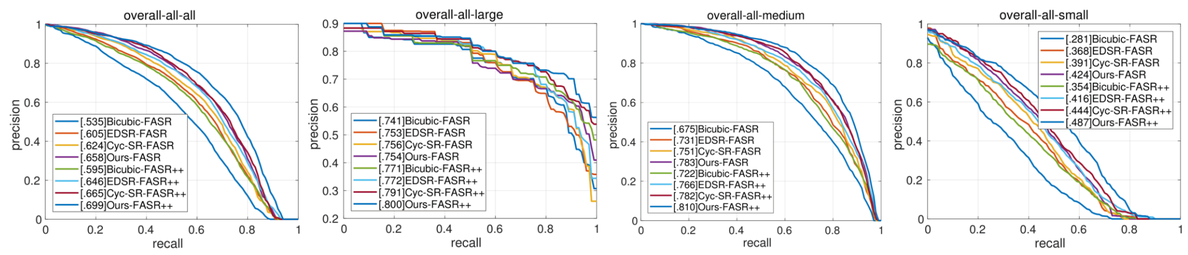

Figure 5. Overall detection performance on all/large/medium/small objects.

Table 2 reports overall results on VOC2007 and VOC2012 datasets. Figure 4. gives some visualizations of detection results. Figure 5. shows the curves of the recall-precision on VOC2007 dataset *w.r.t* all, large, medium, and small size objects, respectively. Our framework exceeds other methods in all scales.

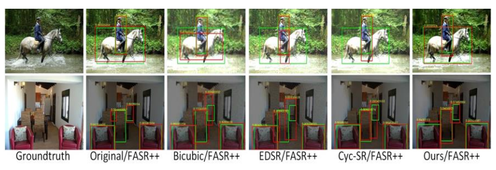

Figure 5. Examples of the detection results on challenging images

We conduct inference on more challenging scenarios shown in Figure 5. All the models never see those images during training, and our method achieves the best results(Table 3).

We report the average PSNR values in our previous experiments(Table 4). Clearly, Cyc-SR outperforms our SR results in terms of PSNR. This result demonstrates that the SR network of our method is detection-driven, which contributes more than other SR components.

Dataset*	Original	Bicubic	EDSR	Cyc-SR	Ours
VOC2007/L	0.755	0.520	0.576	0.544	0.590
VOC2012/L	0.711	0.466	0.523	0.488	0.533
VOC2007/E	0.616	0.359	0.298	0.257	0.615
VOC2012/E	0.569	0.304	0.247	0.200	0.564

Table 3. mAP results on challenging images(IoU=0.5).

Method*	Bicubic	EDSR	Cyc-SR	Ours
PSNR	18.49	18.55	25.38	22.42

Table 4. Averaging PSNR values of different methods.

Discussion & Conclusion

In this work, we propose a framework to facilitate small object detection leveraging on simultaneous super-resolution in an end-to-end manner. Our SR network and detection network are trained jointly. Particularly, the detection loss is back-propagated into the super-resolution network during training to facilitate detection. Compared with the available simultaneous super-resolution and detection methods which heavily rely on low-/high-resolution image pairs, our work breaks through such restriction via applying the CycleGAN strategy, achieving increased generality and applicability. We are going to extend our work to realize instance segmentation of small object, which could provide more valuable information to facilitate precise scene analysis.