Perceptual Generative Adversarial Networks for Small Object Detection

Jianan Li[1] Xiaodan Liang[2] Yunchao Wei[3] Tingfa Xu[1]* Jiashi Feng[3] Shuicheng Yan[3,4]
[1] Beijing Institute of Technology [2] CMU [3] National University of Singapore [4] 360 AI Institute
{20090964, ciom_xtf1}@bit.edu.cn xiaodan1@cs.cmu.edu {eleweiyv, elefjia}@nus.edu.sg yanshuicheng@360.cn

Abstract

Detecting small objects is notoriously challenging due to their low resolution and noisy representation. Existing object detection pipelines usually detect small objects through learning representations of all the objects at multiple scales. However, the performance gain of such ad hoc architectures is usually limited to pay off the computational cost. In this work, we address the small object detection problem by developing a single architecture that internally lifts representations of small objects to "super-resolved" ones, achieving similar characteristics as large objects and thus more discriminative for detection. For this purpose, we propose a new Perceptual Generative Adversarial Network (Perceptual GAN) model that improves small object detection through narrowing representation difference of small objects from the large ones. Specifically, its generator learns to transfer perceived poor representations of the small objects to super-resolved ones that are similar enough to real large objects to fool a competing discriminator. Meanwhile its discriminator competes with the generator to identify the generated representation and imposes an additional perceptual requirement – generated representations of small objects must be beneficial for detection purpose – on the generator. Extensive evaluations on the challenging Tsinghua-Tencent 100K [45] and the Caltech [9] benchmark well demonstrate the superiority of Perceptual GAN in detecting small objects, including traffic signs and pedestrians, over well-established state-of-the-arts.

1. Introduction

Recent great progress on object detection is stimulated by the deep learning pipelines that learn deep representations from the region of interest (RoI) and perform classification based on the learned representations, such as Fast R-CNN [11] and Faster R-CNN [32]. Those pipelines indeed work well on large objects with high resolution, clear appearance and structure from which the discriminative features can be learned. But they usually fail to detect *very small* objects, as rich representations are difficult to learn from their poor-quality appearance and structure, as shown in Figure 1. However, small objects are very common in many real world applications such as traffic sign detection, pedestrian detection for advanced autonomous driving. Small object detection is much more challenging than normal object detection and good solutions are still rare so far.

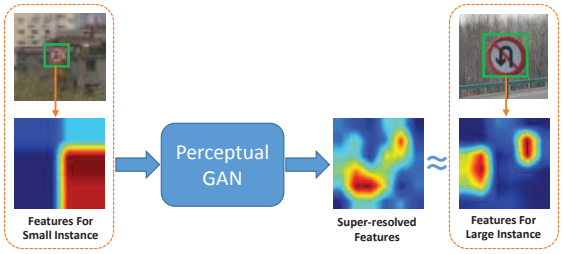

Figure 1. Large and small objects exhibit different representations from high-level convolutional layers of a CNN detector. The representations of large objects are discriminative while those of small objects are of low resolution, which hurts the detection accuracy. In this work, we introduce the Perceptual GAN model to enhance the representations for small objects to be similar to real large objects, thus improve detection performance on the small objects.

Some efforts [4, 25, 18, 39, 23, 1] have been devoted to addressing small object detection problems. One common practice [4, 25] is to increase the scale of input images to enhance the resolution of small objects and produce high-resolution feature maps. Some others [39, 23, 1] focus on developing network variants to generate multi-scale representation which enhances high-level small-scale features with multiple lower-level features layers. However, all of those approaches try to enhance the performance of small object detection by data augmentation or naively increasing the feature dimension. Simply increasing the scale of input images often results in heavy time consumption for training and testing. Besides, the multi-scale representation constructed by the low-level features just works like a black-box and cannot guarantee the constructed features are interpretable and discriminative enough for object detection. In this work, we argue that a preferable way to effectively represent the small objects is to discover the intrinsic structural correlations between small-scale and large-scale objects for each category and then use the transformed representations

*Corresponding author.

to improve the network capability in a more intelligent way.

Therefore, we propose a novel Perceptual Generative Adversarial Network (Perceptual GAN) to generate super-resolved representations for small objects for better detection. The Perceptual GAN aims to enhance the representations of small objects to be similar to those of large object, through fully exploiting the structural correlations between objects of different scales during the network learning. It consists of two subnetworks, *i.e.*, a generator network and a perceptual discriminator network. Specifically, the generator is a deep residual based feature generative model which transforms the original poor features of small objects to highly discriminative ones by introducing fine-grained details from lower-level layers, achieving "super-resolution" on the intermediate representations. The discriminator network serves as a supervisor and provides guidance on the quality and advantages of the generated fine-grained details. Different from the vanilla GAN, where the discriminator is only trained to differentiate fake and real representations, our proposed Perceptual GAN includes a new perceptual loss tailored for the detection purpose. Namely, the discriminator network is trained not only to differentiate between the generated super-resolved representations for small objects and the original ones from real large objects with an adversarial loss, but also to justify the detection accuracy benefiting from the generated super-resolved features with a perceptual loss.

We optimize the parameters of the generator and the discriminator network in an alternative manner to solve the min-max problem. In particular, the generator network is trained with the goal of fooling the discriminator by generating the most large-object like representations from small objects as well as benefiting the detection accuracy. On the other hand, the discriminator is trained to improve its discriminative capability to correctly distinguish the generated super-resolved representations from those from real large objects, and also provides feedback about the localization precision to the generator. Through competition between these two networks, generator is effectively trained to enhance the representations for small objects to super-resolved ones capable of providing high detection accuracy.

We evaluate our Perceptual GAN method on the challenging Tsinghua-Tencent 100K [45] and the Caltech benchmark [9] for traffic sign and pedestrian detection respectively. Small instances are common on these two datasets, thus they provide suitable testbed for evaluating methods on detecting small objects. Our proposed method shows large improvement over state-of-the-art methods and demonstrates its superiority on detecting small objects.

To sum up, this work makes the following contributions. (1) We are the first to successfully apply GAN-alike models to solve the challenging small-scale object detection problems. (2) We introduce a new conditional generator model that learns the additive residual representation between large and small objects, instead of generating the complete representations as before. (3) We introduce a new perceptual discriminator that provides more comprehensive supervision beneficial for detections, instead of barely differentiating fake and real. (4) Successful applications on traffic sign detection and pedestrian detection have been achieved with the state-of-the-art performance.

2. Related Work

2.1. Small Object Detection

Traffic Sign Detection Traffic sign detection and recognition has been a popular problem in intelligent vehicles, and various methods [20, 15, 34, 19, 38, 45] have been proposed to address this challenging task. Traditional methods for this task includes [20] [15]. Recently, CNN-based approaches have been widely adopted in traffic sign detection and classification due to their high accuracy. In particular, Sermanet *et al.* [34] proposed to feed multi-stage features to the classifier using connections that skip layers to boost traffic sign recognition. Jin *et al.* [19] proposed to train the CNN with hingle loss, which provides better test accuracy and faster stable convergence. Wu *et al.* [38] used a CNN combined with fixed and learnable filters to detect traffic signs. Zhu *et al.* [45] trained two CNNs for simultaneously localizing and classifying traffic signs.

Pedestrian Detection The hand-crafted features achieve great success in pedestrian detection. For example, Dollár *et al.* proposed Integral Channel Features (ICF) [8] and Aggregated Channel Features (ACF) [7], which are among the most popular hand-crafted features for constructing pedestrian detectors. Recently, deep learning methods have greatly boosted the performance of pedestrian detection [29, 33, 28, 36, 41]. Ouyang *et al.* [29] proposed a deformation hidden layer for CNN to model mixture poses information, which can further benefit the pedestrian detection task. Tian *et al.* [36] jointly optimized the pedestrian detection with semantic tasks. Sermanet *et al.* [33] utilized multi-stage features to integrate global shape information with local distinctive information to learn the detectors.

2.2. Generative Adversarial Networks

The Generative Adversarial Networks (GANs) [14] is a framework for learning generative models. Mathieu *et al.* [26] and Denton*et al.* [6] adopted GANs for the application of image generation. In [22] and [40], GANs were employed to learn a mapping from one manifold to another for style transfer and inpainting, respectively. The idea of using GANs for unsupervised representation learning was described in [31]. GANs were also applied to image super-resolution in [21]. To the best of our knowledge, this work makes the first attempt to accommodate GANs on the object detection task to address the small-scale problem by generating super-resolved representations for small objects.

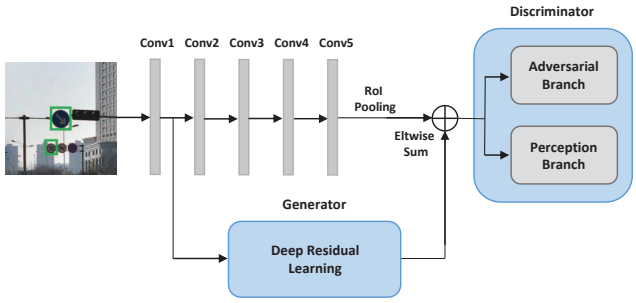

Figure 2. Training procedure of object detection network based on the Perceptual GAN. The perception branch of the discriminator network along with the bottom convolutional layers is first trained using the training images that contain only large objects. Then with the training images that contain only small objects, the generator network is trained to generate super-resolved large-object like representations for small objects. The adversarial branch of the discriminator network is trained to differentiate between the generated super-resolved representations for small objects and the original ones for real large objects. By iteratively boosting the abilities of the generator network and the discriminator network through alternative training, the detection accuracy especially for small objects can be improved.

3. Perceptual GANs

We propose a new Perceptual GAN network to address the challenging small object detection problems. We introduce new designs on the generator model that is able to generate super-resolved representations for small objects, and also a new discriminator considering adversarial loss and perceptual loss to "supervise" the generative process. In this section, we first present the alternative optimization for perceptual GAN from a global view. Then, the details of the generator for super-resolved feature generation and the discriminator for adversarial learning are given.

3.1. Overview

The learning objective for vanilla GAN models [14] corresponds to a minimax two-player game, which is formulated as

$$\min_G \max_D L(D,G) \triangleq \mathbb{E}_{x\sim p_{\text{data}(x)}} \log D(x) + \mathbb{E}_{z\sim p_z(z)} \left[\log(1-D(G(z)))\right],$$

where G represents a generator that learns to map data z from the noise distribution $p_z(z)$ to the distribution $p_{\text{data}(x)}$ over data x, and D represents a discriminator that estimates the probability of a sample coming from the data distribution $p_{\text{data}(x)}$ rather than G. The training procedure for G is to maximize the probability of D making a mistake.

In our case, x and z are the representations for large objects and small objects, *i.e.*, F_l and F_s respectively. We aim to learn a generator function G that transforms the representations of a small object F_s to a super-resolved one $G(F_s)$ that is similar to the original one of the large object F_l. Learning the representation $G(F_s)$ for small objects matching the distribution of large object feature F_l may be difficult due to the limited information contained in F_s. We thus introduce a new conditional generator model which is conditioned on the extra auxiliary information, *i.e.*, the low-level features of the small object f from which the generator learns to generate the residual representation between the representations of large and small objects through residual learning instead.

$$\min_G \max_D L(D,G) \triangleq \mathbb{E}_{F_l\sim p_{\text{data}(F_l)}} \log D(F_l) + \mathbb{E}_{F_s\sim p_{F_s}(F_s|f)}[\log(1-D(\underbrace{F_s+G(F_s|f)}_{\text{residual learning}}))].$$

In this case, the generator training can be substantially simplified over directly learning the super-resolved representations for small objects. For example, if the input representation is from a large object, the generator only needs to learn a zero-mapping. Besides, we introduce a perceptual loss on the discriminator to benefit the detection task as detailed below.

As shown in Figure 2, the generator network aims to generate super-resolved representation for the small object. The discriminator includes two branches, *i.e.* the adversarial branch for differentiating between the generated super-resolved representation and the original one for the large object and the perception branch for justifying the detection accuracy benefiting from the generated representation. We optimize the parameters embedded in the generator and the discriminator network in an alternative manner to solve the adversarial min-max problem.

Denote G_{Θ_g} as the generator network with parameters Θ_g. We obtain Θ_g by optimizing the loss function L_{dis}

$$\Theta_g = \arg\min_{\Theta_g} L_{dis}(G_{\Theta_g}(F_s)), \quad (1)$$

where L_{dis} is the weighted combination of the adversarial loss L_{dis_a} and the perceptual loss L_{dis_p} produced by the discriminator network, which is detailed in Section 3.3. We train the adversarial branch of the discriminator network to maximize the probability by assigning the correct label to both the generated super-resolved feature for the small object $G_{\Theta_g}(F_s)$ and the feature for the large object F_l.

Suppose D_{Θ_a} is the adversarial branch of the discriminator network parameterized by Θ_a. We obtain Θ_a by optimizing a specific loss function L_a:

$$\Theta_a = \arg\min_{\Theta_a} L_a(G_{\Theta_g}(F_s), F_l), \quad (2)$$

where the loss L_a is defined as

$$L_a = -\log D_{\Theta_a}(F_l) - \log(1-D_{\Theta_a}(G_{\Theta_g}(F_s))). \quad (3)$$

Eventually, L_a encourages the discriminator network to distinguish the difference between the currently generated super-resolved representation for the small object and the original one from the real large object.

To justify the detection accuracy benefiting from the generated super-resolved representation, the perception branch should be first well trained based on the features of large

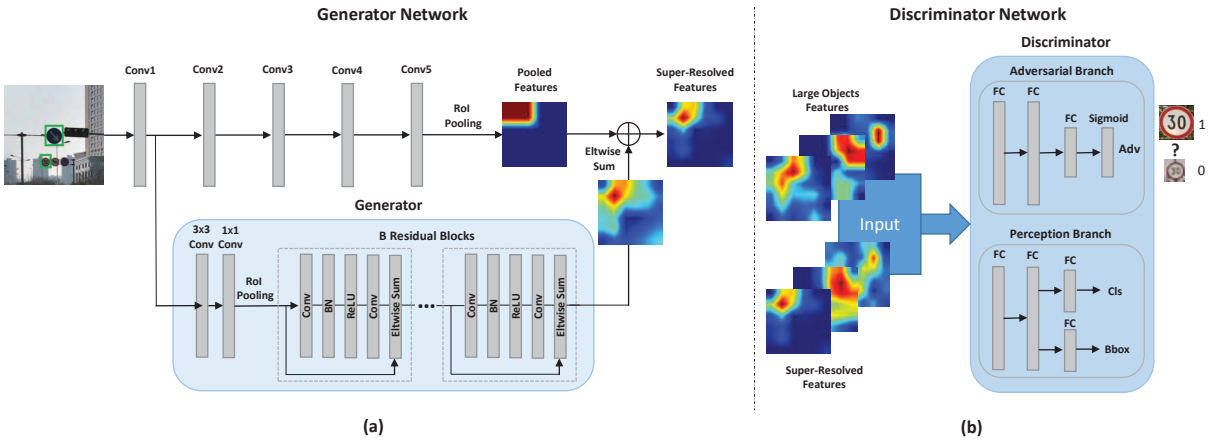

Figure 3. Details of the proposed Perceptual Generative Adversarial network. (a) The generator is a deep residual network which takes the features with fine-grained details from lower-level layer as input and passes them to 3×3 convolutional filters followed by 1×1 convolutional filters to increase the feature dimension to be aligned with that of "Conv5". Then B residual blocks each of which consists of convolutional layers followed by batch normalization and ReLU activation are employed to learn the residual representation, which is used to enhance the pooled features from "Conv5" for small objects to super-resolved representation through element-wise sum operation. (b) The discriminator takes the features of large object and the super-resolved representation of small object as inputs and splits into two branches. The adversarial branch consists of three fully connected layers followed by sigmoid activation, which is used to estimate the probability that the current input representation belongs to that of real large object. The perception branch consists of two fully connected layers followed by two output sibling layers, which are used for classification and bounding box regression respectively to justify the detection accuracy benefiting from the generated super-resolved representation.

objects to achieve high detection accuracy. Denote D_{Θ_p} as the perception branch of the discriminator network parameterized by Θ_p. We obtain Θ_p by optimizing a specific loss function L_{dis_p} with the representation for the large object:

$$\Theta_p = \arg\min_{\Theta_p} L_{dis\_p}(F_l), \tag{4}$$

where L_{dis_p} is the multi-task loss for classification and bounding-box regression, which is detailed in Section 3.3.

With the average size of all instances, we obtain two subsets containing small objects and large objects, respectively. For overall training, we first learn the parameters of bottom convolutional layers and the perception branch of the discriminator network based on the subset containing large objects. Guided by the learned perceptual branch, we further train the generator network based on the subset containing small objects and the adversarial branch of the discriminator network using both subsets. We alternatively perform the training procedures of the generator and the adversarial branch of the discriminator network until a balance point is finally achieved, *i.e.* large-object like super-resolved features can be generated for the small objects with high detection accuracy.

3.2. Conditional Generator Network Architecture

The generator network aims to generate super-resolved representations for small objects to improve detection accuracy. To achieve this purpose, we design the generator as a deep residual learning network that augments the representations of small objects to super-resolved ones by introducing more fine-grained details absent from the small objects through residual learning.

As shown in Figure 3, the generator takes the feature from the bottom convolutional layer as the input that preserves many low-level details and is informative for feature super-resolution. The resulting feature is first passed into the 3×3 convolution filters followed by the 1×1 convolution filters to increase the feature dimension to be the same as that of "Conv5". Then, B residual blocks with the identical layout consisting of two 3×3 convolutional filters followed by batch-normalization layer and ReLU activation layer are introduced to learn the residual representation between the large and the small objects, as a generative model. The learned residual representation is then used to enhance the feature pooled from "Conv5" for the small object proposal through RoI pooling [11] by element-wise sum operation, producing super-resolved representation.

3.3. Discriminator Network Architecture

As shown in Figure 3, the discriminator network is trained to not only differentiate between the generated super-resolved feature for the small object and the original one from the real large object, but also justify the detection accuracy benefiting from the generated super-resolved feature. Taking the generated super-resolved representation as input, the discriminator passes it into two branches, *i.e.*, the adversarial branch and the perception branch. The adversarial branch consists of two fully-connected layers followed by a sibling output layer with the sigmoid activation, which produces an adversarial loss. The perception branch consists of two fully-connected layers followed by two sibling output layers, which produces a perceptual loss to justify the detection performance contributing to the super-

resolved representation. The output units number of the first two fully-connected layers for both branches are 4096 and 1024 respectively.

Given the adversarial loss L_{dis_a} and the perceptual loss L_{dis_p}, a final loss function L_{dis} can be produced as weighted sum of both individual loss components. Given weighting parameters w_1 and w_2, we define $L_{dis} = w_1 \times L_{dis\_a} + w_2 \times L_{dis_p}$ to encourage the generator network to generate super-resolved representation with high detection accuracy. Here we set both w_1 and w_2 to be one.

Adversarial Loss Denote D_{Θ_a} as the adversarial branch of the discriminator network with parameters Θ_a. Taking the generated representation $G_{\Theta_g}(F_s)$ for each object proposal as input, this branch outputs the estimated probability of the input representation belonging to a real large object, denoted as $D_{\Theta_a}(G_{\Theta_g}(F_s))$. By trying to fool the discriminator network with the generated representation, an adversarial loss is introduced to encourage the generator network to produce the super-resolved representation for the small object similar as that of the large object. The adversarial loss L_{dis_a} is defined as

$$L_{dis\_a} = -\log D_{\Theta_a}(G_{\Theta_g}(F_s)). \quad (5)$$

Perceptual Loss Taking the super-resolved representation for each proposal as input, the perception branch outputs the category-level confidences $p = (p_0, ..., p_k)$ for $K+1$ categories and the bounding-box regression offsets, $r_k = (r_x^k, r_y^k, r_w^k, r_h^k)$ for each of the K object classes, indexed by k. Following the parameterization scheme in [12], r_k specifies a scale-invariant translation and log-space height/width shift relative to an object proposal. Each training proposal is labeled with a ground-truth class g and a ground-truth bounding-box regression target r^*. The following multi-task loss L_{dis_p} is computed to justify the detection accuracy benefiting from the generated super-resolved features for each object proposal:

$$L_{dis\_p} = L_{cls}(p, g) + \mathbf{1}[g \geq 1] L_{loc}(r_g, r^*), \quad (6)$$

where L_{cls} and L_{loc} are the losses for the classification and the bounding-box regression, respectively. In particular, $L_{cls}(p, g) = -\log p_g$ is log loss for the ground truth class g and L_{loc} is a smooth L_1 loss proposed in [11]. For background proposals (*i.e.* $g = 0$), the L_{loc} is ignored.

4. Experiments

4.1. Datasets and Evaluation Metrics

4.1.1 Traffic-sign Detection Datasets

The Tsinghua-Tencent 100K [45] is a large traffic-sign benchmark, which contains 30,000 traffic-sign instances. The images are of resolution 2,048×2,048. Following [45], we ignore the classes whose instances are less than 100 and have 45 classes left. The performance is evaluated using the same detection metrics as for the Microsoft COCO benchmark. We report the detection performance on difference sizes of objects, including small objects (area $< 32 \times 32$ pixels), medium objects ($32 \times 32 <$ area $< 96 \times 96$) and large objects (area $> 96 \times 96$). The numbers of instances corresponding to the three kinds of division are 3270, 3829 and 599, respectively. This evaluation scheme helps us understand the ability of a detector on objects of different sizes.

4.1.2 Pedestrian Detection Datasets

The Caltech benchmark [9] is the most popular pedestrian detection dataset. About 250,000 frames with a total of 350,000 bounding boxes and 2,300 unique pedestrians are annotated. We use dense sampling of the training data (every 4th frame) as adopted in [44, 27]. Following the conventional evaluation setting [9], the performance is evaluated on pedestrians over 50 pixels tall with no or partial occlusion, which are often of very small sizes. The evaluation metric is log-average Miss Rate on False Positive Per Image (FPPI) in $[10^{-2}, 10^0]$ (denoted as MR following [42]).

4.2. Implementation Details

For traffic sign detection, we use the pretrained VGG-CNN-M-1024 model [3] as adopted in [24] to initialize our network. For pedestrian detection, we use the pretrained VGG-16 model [35] as adopted in [41]. For the generator and the discriminator network, the parameters of newly added convolutional layers and fully connected layers are initialized with "Xavier" [13]. We resize the image to 1600 pixels and 960 pixels on the shortest side as input for traffic sign detection and pedestrian detection respectively. Following [16], we perform down-sampling directly by convolutional layers with a stride of 2. The implementation is based on the publicly available Fast R-CNN framework [11] built on the Caffe platform [17].

The whole network is trained with Stochastic Gradient Descent (SGD) with momentum of 0.9, and weight decay of 0.0005 on a single NVIDIA GeForce GTX TITAN X GPU with 12GB memory. For training the generator network, each SGD mini-batch contains 128 selected object proposals from each training image. Following [11], in each mini-batch, 25% of object proposals are foreground that overlap with a ground truth bounding box with at least 0.5 IoU, and the rest are background. For training the discriminator network, each SGD mini-batch contains 32 selected foreground object proposals from four training images. The number of residual blocks in the generator network B is set as 6. For the Tsinghua-Tencent 100K [45] benchmark, we train a Region Proposal Network (RPN) as proposed in [32] to generate object proposals on the training and testing images. For the Caltech benchmark [9], we utilize the ACF pedestrian detector [7] trained on the Caltech training set for object proposals generation. For testing, on average, the Perceptual GAN processes one image within 0.6 second (excluding object proposal time).

Table 1. Comparisons of detection performance for different sizes of traffic signs on Tsinghua-Tencent 100K. (R): Recall, (A): Accuracy. (In %)

Object size	Small	Medium	Large
Fast R-CNN [11] (R)	46	71	77
Fast R-CNN [11] (A)	74	82	80
Faster R-CNN [32] (R)	50	84	91
Faster R-CNN [32] (A)	24	66	81
Zhu *et al.* [45] (R)	87	94	88
Zhu *et al.* [45] (A)	82	91	91
Ours (R)	**89**	**96**	**89**
Ours (A)	**84**	**91**	**91**

4.3. Performance Comparison

4.3.1 Traffic-sign Detection

Table 1 provides the comparison of our approach with other state-of-the-arts in terms of average recall and accuracy on traffic-sign detection. It can be observed that the proposed Perceptual GAN outperforms the previous state-of-the-art method of Zhu *et al.* [45] in terms of average recall and accuracy: 89% and 84% vs 87% and 82%, 96% and 91% vs 94% and 91%, 89% and 91% vs 88% and 91% on three subsets of different object sizes. Specifically, our approach makes a large improvement, *i.e.*, 2% and 2% in average recall and accuracy on the small-size subset, demonstrating its superiority in accurately detecting small objects. Table 2 shows the comparisons of recall and accuracy for each category. Our approach achieves the best performance in most categories such as "p3" and "pm55" in which small instances are most common. More comparisons of accuracy-recall curves in terms of different object sizes are provided in Figure 5, which can further demonstrate the effectiveness of the proposed generative adversarial learning strategy.

Several examples of the detection results for small objects are visualized in Figure 7. We compare our visual results with those from Zhu *et al.* [45]. Note that Zhu *et al.* [45] take the original image of resolution $2,048 \times 2,048$ as input, which may cause heavy time consumption for training and testing. In contrast, the Perceptual GAN only takes image of resolution 1600×1600 as input. In addition, no data augmentation as adopted by Zhu *et al.* [45] has been applied. As shown in Figure 7, generally, our method can accurately classify and localize most objects in small scales, while Zhu *et al.* [45] fails to localize some instances due to serious small-scale problem.

4.3.2 Pedestrian Detection

Since the pedestrian instances on the Caltech benchmark [9] are often of small scales, the overall performance on it can be used to evaluate the capability of a method in detecting small objects. We compare the result of Perceptual GAN with all the existing methods that achieved best performance on the Caltech testing set, including VJ [37], HOG [5], LDCF [27], Katamari [2], SpatialPooling+ [30], TA-CNN [36], Checkerboards [43], CompACT-Deep [44] and RPN+BF [41]. As shown in Figure 4, the proposed method outperforms all the previous methods and achieves the lowest log-average miss rate of 9.48%, validating its superiority in detecting small objects.

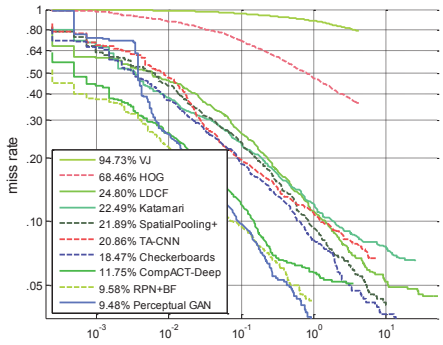

Figure 4. Comparisons of detection performance with the state-of-the-arts on the Caltech benchmark.

4.4. Ablation Studies

We investigate the effectiveness of different components of Perceptual GAN. All experiments are performed on the Tsinghua-Tencent 100K [45] dataset. The performance achieved by different variants of Perceptual GAN and parameter settings on small objects and all the objects of different sizes are reported in the following.

4.4.1 The Effectiveness of Super-resolved Features by Generator

To verify the superiority of the generated super-resolved representation in detecting small objects, we compare our method with several other feature enhancement solutions, including combining low-level features, improving the image resolution by simply increasing the input scales, taking images with multi-scales as input. All these methods are implemented based on the base convolutional layers and the perceptual branch with end-to-end training. As shown in Table 3, "**Skip Pooling**" indicates the model trained by combining low-level features through skip pooling as proposed in [1]. Our Perceptual GAN outperforms this approach by 13% and 2% in average recall and accuracy on small-size objects respectively, which validates that our method can effectively incorporate fine-grained details from low-level layers to improve small object detection. "**Large Scale Images**" represents the model trained with images of higher resolution by simply increasing the scale of input images to 2048×2048. "**Multi-scale Input**" indicates the model trained with input images with multi-scale settings ($s \in 1120, 1340, 1600, 1920, 2300$) as adopted in [11]. One can observe that our Perceptual GAN outperforms both approaches in performance on small objects. This shows that our method is more effective in boosting small object detection than simply increasing the input image scale or using multi-scale settings.

We further visualize some of the generated super-

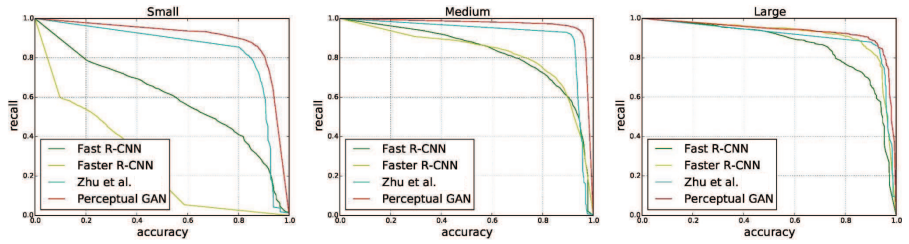

Figure 5. Comparisons of overall detection performance on Tsinghua-Tencent 100K, for small, medium and large traffic signs.

Table 2. Comparisons of detection performance for each class on Tsinghua-Tencent 100K. (R): Recall, (A): Accuracy. (In %)

Class	i2	i4	i5	il100	il60	il80	io	ip	p10	p11	p12	p19	p23	p26	p27
Fast R-CNN [11] (R)	51	74	84	44	61	10	70	73	54	71	21	42	65	63	36
Fast R-CNN [11] (A)	82	86	85	85	70	91	75	80	72	73	47	48	79	74	100
Faster R-CNN [32] (R)	60	76	80	74	89	77	72	64	62	61	53	73	75	78	81
Faster R-CNN [32] (A)	44	46	45	41	57	62	41	39	45	38	60	59	65	50	79
Zhu *et al.* [45] (R)	82	94	95	97	91	94	89	92	95	91	89	94	94	93	96
Zhu *et al.* [45] (A)	72	83	92	100	91	93	76	87	78	89	88	53	87	82	78
Ours (R)	**84**	**95**	**95**	**95**	**92**	**95**	**92**	**91**	**89**	**96**	**97**	**97**	**95**	**94**	**98**
Ours (A)	**85**	**92**	**94**	**97**	**95**	**83**	**79**	**90**	**84**	**85**	**88**	**84**	**92**	**83**	**98**

Class	p3	p5	p6	pg	ph4	ph4.5	ph5	pl100	pl120	pl20	pl30	pl40	pl5	pl50	pl60
Fast R-CNN [11] (R)	50	78	8	88	32	77	18	68	39	14	18	58	69	34	41
Fast R-CNN [11] (A)	85	87	100	86	92	82	88	86	92	89	59	78	88	65	73
Faster R-CNN [32] (R)	55	82	54	84	57	80	46	86	77	46	61	68	69	62	65
Faster R-CNN [32] (A)	48	57	75	80	68	58	51	68	67	51	43	52	53	39	53
Zhu *et al.* [45] (R)	91	95	87	91	82	88	82	98	98	96	94	96	94	94	93
Zhu *et al.* [45] (A)	80	89	87	93	94	88	89	97	100	90	90	89	84	87	93
Ours (R)	**93**	**96**	**100**	**93**	**78**	**88**	**85**	**96**	**98**	**96**	**93**	**96**	**92**	**96**	**91**
Ours (A)	**92**	**90**	**83**	**93**	**97**	**68**	**69**	**97**	**98**	**92**	**91**	**90**	**86**	**87**	**92**

Class	pl70	pl80	pm20	pm30	pm55	pn	pne	po	pr40	w13	w32	w55	w57	w59	wo
Fast R-CNN [11] (R)	2	34	43	19	58	87	90	46	95	32	41	43	73	74	16
Fast R-CNN [11] (A)	100	84	70	67	76	85	87	66	78	40	100	57	66	64	55
Faster R-CNN [32] (R)	68	68	63	63	79	77	83	63	98	71	59	63	79	78	50
Faster R-CNN [32] (A)	61	52	61	67	61	37	47	37	75	33	54	39	48	39	37
Zhu *et al.* [45] (R)	93	95	88	91	95	91	93	67	98	65	71	72	79	82	45
Zhu *et al.* [45] (A)	95	94	91	81	60	92	93	84	76	65	89	86	95	75	52
Ours (R)	**91**	**99**	**88**	**94**	**100**	**96**	**97**	**83**	**97**	**94**	**85**	**95**	**94**	**95**	**53**
Ours (A)	**97**	**86**	**90**	**77**	**81**	**89**	**93**	**78**	**92**	**66**	**83**	**88**	**93**	**71**	**54**

Table 3. Comparisons of detection performance with several variants of Perceptual GAN on Tsinghua-Tencent 100K. (R): Recall, (A): Accuracy. (In %)

Object size	Small	All
Skip Pooling (R)	76	87
Skip Pooling (A)	82	86
Large Scale Images (R)	85	92
Large Scale Images (A)	81	86
Multi-scale Input (R)	89	93
Multi-scale Input (A)	77	83
Ours (R)	**89**	**93**
Ours (A)	**84**	**88**

resolved features, as shown in Figure 6. The second and the last column show the original features pooled from the top convolutional layer for proposals of small objects and large objects respectively. The learned residual representation and the generated super-resolved features by the generator for small objects are shown in the third and the fourth column respectively. One can observe that the generator successfully learns to transfer the poor representations of small objects to super-resolved ones similar to those of large objects, validating the effectiveness of the Perceptual GAN.

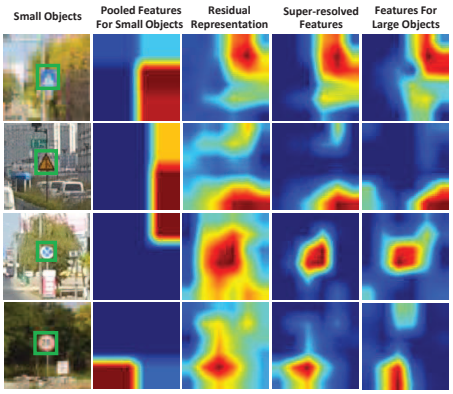

Figure 6. Visualization of the super-resolved features.

4.4.2 The Effectiveness of Adversarial Training

The proposed Perceptual GAN trains the generator and the discriminator through alternative optimization. To demonstrate the necessity of adversarial training, we report the performance of our model with or without alternative optimization during training stage in Table 4. "Ours_Baseline" indicates the model of training the proposed detection pipeline with the generator network end-to-end without any alterna-

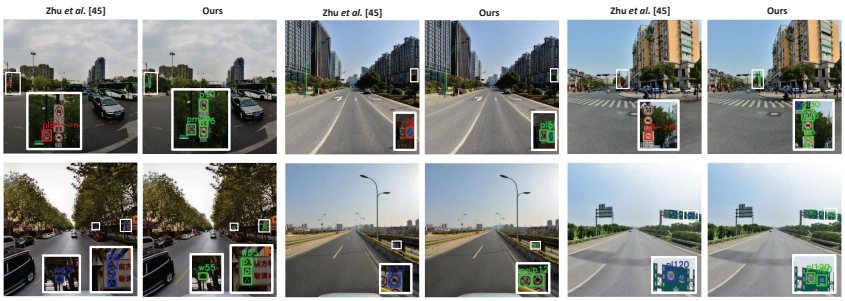

Figure 7. Detection results of Zhu *et al.* [45] and the proposed method on Tsinghua-Tencent 100K. The green, red, and blue rectangle denote the true positive, false positive and false negative respectively. The proposed Perceptual GAN can successfully detect most small-size traffic signs which the method of Zhu *et al.* [45] has missed or detected incorrectly. Best viewed in color.

Table 4. Comparisons of detection performance by Perceptual GAN with or without alternative optimization on Tsinghua-Tencent 100K. (R): Recall, (A): Accuracy. (In %)

Object size	Small	All
Ours_Baseline (R)	80	89
Ours_Baseline (A)	80	85
Ours_Alt (R)	**89**	**93**
Ours_Alt (A)	**84**	**88**

Table 5. Comparisons of detection performance for introducing fine-grained details from different lower-level layers on Tsinghua-Tencent 100K. (R): Recall, (A): Accuracy. (In %)

Object size	Small	All
Ours_Conv3 (R)	74	86
Ours_Conv3 (A)	78	85
Ours_Conv2 (R)	87	92
Ours_Conv2 (A)	80	86
Ours_Conv1 (R)	**89**	**93**
Ours_Conv1 (A)	**84**	**98**

tive optimization step. "Ours_Alt" indicates the model of alternatively training the generator and the discriminator. By comparing "Ours_Alt" with "Ours_Baseline", one can observe that considerable improvements in the recall and accuracy on small-size object detection can be obtained when using alternative optimization. This shows that Perceptual GAN can improve its performance in detecting small objects by recursively improving the ability of the generator and the discriminator through adversarial training.

4.4.3 Different Lower Layers for Learning Generator

The proposed generator learns fine-grained details of small objects from representations of lower-level layers. In particular, we employ the features from "Conv1" as the inputs for learning the generator. To validate the effectiveness of this setting, we conduct additional experiments using features from "Conv2" and "Conv3" for learning the generator, respectively. As shown in Table 5, we can observe that performance consistently decreases by employing the representations from higher layers. The reason is that lower layers can capture more details of small objects. Therefore, using low-level features from "Conv1" for learning the generator gives the best performance.

4.5. Discussion on General Small Object Detection

To evaluate the generalization capability of the proposed generator on more general and diverse object categories, we train the proposed detection pipeline with the generator network end-to-end on the union of the trainval set of PASCAL VOC 2007 and VOC 2012 [10], and evaluate it on the test set of VOC 2007 on the most challenging classes (*i.e.*, boat, bottle, chair and plant) in which small instances are most common. Our method achieves 69.4%, 60.2%, 57.9% and 41.8% in Average Precision (AP) for boat, bottle, chair, and plant, respectively. It significantly outperforms those of the Fast R-CNN [11] baseline, *i.e.*, 59.4%, 38.3%, 42.8% and 31.8%, well demonstrating the generalization capability of the proposed generator for general small object detection.

5. Conclusion

In this paper, we proposed a novel generative adversarial network to address the challenging problem of small object detection. Perceptual GAN generates super-resolved representations for small objects to boost detection performance by leveraging the repeatedly updated generator network and the discriminator network. The generator learns a residual representation from the fine-grained details from lower-level layers, and enhances the representations for small objects to approach those for large objects by trying to fool the discriminator which is trained to well differentiate between both representations. Competition in the alternative optimization of both networks encourages the Perceptual GAN to generate super-resolved large-object like representations for small objects, thus improving detection performance. Extensive experiments have demonstrated the superiority of the proposed Perceptual GAN in detecting small objects.

Acknowledgement

This work was partially supported by China Scholarship Council (Grant No. 201506030045). The work of Jiashi Feng was partially supported by National University of Singapore startup grant R-263-000-C08-133 and Ministry of Education of Singapore AcRF Tier One grant R-263-000-C21-112.

References

[1] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. *arXiv preprint arXiv:1512.04143*, 2015. 1, 6

[2] R. Benenson, M. Omran, J. Hosang, and B. Schiele. Ten years of pedestrian detection, what have we learned? In *ECCV*, pages 613–627, 2014. 6

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*, 2014. 5

[4] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun. 3d object proposals for accurate object class detection. In *NIPS*, pages 424–432, 2015. 1

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 6

[6] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *NIPS*, pages 1486–1494, 2015. 2

[7] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *TPAMI*, 36(8):1532–1545, 2014. 2, 5

[8] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, volume 2, page 5, 2009. 2

[9] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *TPAMI*, 34(4):743–761, 2012. 1, 2, 5, 6

[10] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. 88(2):303–338, 2010. 8

[11] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 1, 4, 5, 6, 7, 8

[12] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pages 580–587, 2014. 5

[13] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Aistats*, volume 9, pages 249–256, 2010. 5

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2, 3

[15] M. Haloi. A novel plsa based traffic signs classification system. *arXiv preprint arXiv:1503.06643*, 2015. 2

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 5

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678, 2014. 5

[18] H. Jiang and S. Wang. Object detection and counting with low quality videos. In *Technical Report*, 2016. 1

[19] J. Jin, K. Fu, and C. Zhang. Traffic sign recognition with hinge loss trained convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):1991–2000, 2014. 2

[20] T. T. Le, S. T. Tran, S. Mita, and T. D. Nguyen. Real time traffic sign detection using color and shape-based features. In *Asian Conference on Intelligent Information and Database Systems*, pages 268–278. Springer, 2010. 2

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 2

[22] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. *arXiv preprint arXiv:1601.04589*, 2016. 2

[23] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua. A convolutional neural network cascade for face detection. In *CVPR*, pages 5325–5334, 2015. 1

[24] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, pages 2167–2175, 2016. 5

[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015. 1

[26] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 2

[27] W. Nam, P. Dollár, and J. H. Han. Local decorrelation for improved pedestrian detection. In *NIPS*, pages 424–432, 2014. 5, 6

[28] W. Ouyang and X. Wang. A discriminative deep model for pedestrian detection with occlusion handling. In *CVPR*, pages 3258–3265, 2012. 2

[29] W. Ouyang and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, pages 2056–2063, 2013. 2

[30] S. Paisitkriangkrai, C. Shen, and A. van den Hengel. Strengthening the effectiveness of pedestrian detection with spatially pooled features. In *ECCV*, pages 546–561. 2014. 6

[31] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 1, 5, 6, 7

[33] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *CVPR*, pages 3626–3633, 2013. 2

[34] P. Sermanet and Y. LeCun. Traffic sign recognition with multi-scale convolutional networks. In *IJCNN*, pages 2809–2813. IEEE, 2011. 2

[35] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[36] Y. Tian, P. Luo, X. Wang, and X. Tang. Pedestrian detection aided by deep learning semantic tasks. In *CVPR*, 2015. 2, 6

[37] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004. 6

[38] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu. Traffic sign detection based on convolutional neural networks. In *IJCNN*, pages 1–7. IEEE, 2013. 2

[39] F. Yang, W. Choi, and Y. Lin. Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers. In *CVPR*, pages 2129–2137, 2016. 1

[40] R. Yeh, C. Chen, T. Y. Lim, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2016. 2

[41] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, pages 443–457. Springer, 2016. 2, 5, 6

[42] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele. How far are we from solving pedestrian detection? *arXiv preprint arXiv:1602.01237*, 2016. 5

[43] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, pages 1751–1760. IEEE, 2015. 6

[44] M. S. Zhaowei Cai and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015. 5, 6

[45] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu. Traffic-sign detection and classification in the wild. In *CVPR*, pages 2110–2118, 2016. 1, 2, 5, 6, 7, 8