Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

See Clearly in the Distance: Representation Learning GAN for Low Resolution Object Recognition

YUE XI[1,2], JIANGBIN ZHENG[1], WENJING JIA[2], XIANGJIAN HE[2](SENIOR MEMBER, IEEE), HANHUI LI[3], ZHUQIANG REN[2], AND KIN-MAN LAM[4](SENIOR MEMBER, IEEE)

[1]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
[2]School of Electrical and Data Engineering, University of Technology Sydney, NSW 2007, Australia
[3]Institute for Media Innovation, Nanyang Technological University, Singapore
[4]Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong

Corresponding author: Jiangbin Zheng (e-mail: zhengjb@nwpu.edu.cn) and Xiangjian He (e-mail: Xiangjian.He@uts.edu.au)

This work was supported in part by ONR-G for creating the dataset, National Natural Science Foundation of China (Project ID 61972321), and Research and Development Plan of Shaanxi Province under Grant 2017ZDXM-GY-094 and Grant 2015KTZDGY04-01.

ABSTRACT Identifying tiny objects with extremely low resolution is generally considered a very challenging task even for human vision, due to limited information presented inside the object areas. There have been very limited attempts in recent years to deal with low-resolution recognition. The existing solutions rely on either generating super-resolution images or learning multi-scale features. However, their performance improvement becomes very limited, especially when the resolution becomes very low. In this paper, we propose a Representation Learning Generative Adversarial Network (*RL*-GAN) to generate super *image representation* that is optimized for recognition. Our solution deals with the classical vision task of object recognition in the distance. We evaluate our idea on the challenging task of low-resolution object recognition. Comparison of experimental results conducted on public and our newly created WIDER-SHIP datasets demonstrate the effectiveness of our *RL*-GAN, which improves the classification results significantly, with 10-15% gain on average, compared with benchmark solutions.

INDEX TERMS convolutional neural networks, generative adversarial networks, low resolution object recognition, representation learning

I. INTRODUCTION

Recent advances in object recognition are largely stimulated by deep learning techniques, such as ResNet [1], DenseNet [2] and SeNet [3], which learn deep representations from regions of interest (RoIs) and perform classification. Those models work well on regions with sufficient image details, but they perform poorly when dealing with objects with extremely low resolution (FIGURES 4 and 6 show some examples of such low-resolution images). However, identifying objects in the far distance is of great interest in many applications, such as remote sensing for Earth Vision [4], far-field video surveillance on Unmanned Aerial Vehicles (UAVs) [5], and privacy-preserving video analysis [6].

Low-resolution (LR) object recognition, *i.e.*, identifying tiny objects from extremely low resolution images is generally considered a very challenging task even for human vision, because the information presented inside the object areas is too little to allow vision algorithms to identify them. As pointed out in [7] where the very low resolution face recognition problem was defined for the first time, a minimum face resolution of 32×32 is required for stand-alone recognition algorithms. Therefore, contrary to its high-resolution (HR) counterpart, which can achieve high accuracy, the performance of LR object recognition is poor and functional solutions are still rare.

The last couple of years have seen increasing interest from the research community on LR face or activity recognition, *e.g.*, the discriminative learning approach [8], the knowledge distillation method [9], as well as [6], [10], [11] for LR activity recognition.

An intuitive solution is to super-resolve LR images and generate super-resolution (SR) images (*a.k.a*, 'Hallucination') and then simply apply techniques designed for rec-

ognizing objects of high or normal resolution [12], [13]. Recall that there is a fundamental difference between object recognition and image super-resolution (SR). Image SR aims to generate images of better visual quality for human viewing, but the goal of object recognition is to achieve high recognition accuracy. Although, intuitively, classification conducted on images of higher resolution produces higher accuracy in general, this is not always and necessarily true, especially when the generated super-resolution images contain distorted information or severe artifacts, which result in poor classification results. Moreover, the two steps, namely super-resolution and classification, are typically designed and optimized separately and the resultant SR images do not necessarily lead into optimal recognition performance. Last but not the least, this approach generally requires high computation load during both training and inference stages. Therefore, training the entire system end-to-end and optimising the networks also for the task of interest has become a recent trend.

Another major stream of solutions is to exploit the semantic similarity among all predicted candidate objects and cluster those candidates of the same category into one group to boost the recognition performance of the network when handling tiny objects [14], [15]. However, this approach cannot work effectively when the objects are not from the same scene or not crowded enough.

Recently, representation-transforming based methods [9], [16]–[18] have attempted to simultaneously transform LR images and their corresponding HR images into a common feature subspace while minimizing the distance between them, and have attracted much interest from the research community.

Li [19] designed a generator, which learned to transfer perceived poor representations of small objects to super-resolved ones that were similar enough to real large objects to fool a competing discriminator. That is to say, the method used the features of large size objects as supervision signals to guide the features of small size objects. The correspondence between the large size objects and their small size counterparts is the key, without which, the features of a different, large size object will mislead the features of another small size object and hence affect the classification. Instead, the pairing HR and LR images used in our approach ensures that the representation of an HR image is used to guide the generator to transform the representation of its LR counterpart to the high-quality one.

In our work, aiming at achieving high classification accuracy directly from LR images, we propose a Representation Learning Generative Adversarial Network (*RL*-GAN). In the proposed approach shown in FIGURE 1, the feature representations learned from HR images are used as a guide to enhance the discriminative ability of the feature representation extracted from LR images. Such enhancement is essentially to super-resolve the LR feature representation, so as to achieve similar attributes to the HR feature representation and make them more discriminative, for better classification.

As an application of our proposed *RL*-GAN for low-resolution object recognition, in this work we define a rarely attempted problem of 'low resolution ship classification' (LRSC) from satellite images, and demonstrate, with extensive experiments, how our proposed *RL*-GAN can see more clearly in the distance. We focus on the key step in object detection and recognition, and focus our experiments on low resolution object classification. The existing low-resolution ship datasets either are created for detection purpose (*e.g.*, DOTA [4]) and do not contain ground-truth ship type labels, or are captured from CCTV cameras mounted on harbors (*e.g.*, SeaShip [20]) instead of satellites, or contain only high-resolution images (*e.g.*, HRSC [21]). We have created a new dataset 'WIDER-SHIP' for low-resolution ship classification and evaluated our proposed approach on it. We have also tested our approach on other benchmark datasets, to show that our proposed solution can also be applied to other objects.

In summary, the main contributions of this work are: 1) We propose a ***RL*-GAN architecture** to enhance the discriminability of the LR image representation resulting in comparable classification performance with that conducted on HR images. 2) We propose a **Residual Representation based generator** to generate a more effective representation of LR images. The residual representation is adapted to fuel back the lost details in the representation space of LR images. 3) We produce a new dataset **WIDER-SHIP**, which provides paired images of multiple resolutions of ships in satellite images and can be used to evaluate not only LR image classification, but also LR object recognition.

II. RELATED WORK

Recently, there have seen increasing interest from the research community on various low-resolution vision problems. The existing solutions can be roughly grouped into three major streams, *i.e.*, the super-resolution based approaches, the resolution- or scale-invariant representation based approaches, the transfer learning based methods and the representation-transforming based approaches.

The **super-resolution based** approaches attempt to convert LR images or representations into their HR counterparts for improved recognition. In [22], Noh *et al.* proposed representation-level enhancement method for LR object recognition, which leveraged HR image features as supervision signals for guiding the enhancement of the LR ones. For example, in [12], [13], [23], photo-realistic HR images were generated from LR images for the task of classification. However, since SR and recognition are often optimized separately, it is hard to achieve a solution optimal for the recognition task with these SR based recognition models. Bai *et al.* [24] proposed a super-resolution RoIs based generative adversarial network, which consisted of two modules, *i.e.*, the generator, which was a super-resolution network to up-sample LR images into HR ones and recover the detailed information for more accurate detection, and the discriminator, which was a multi-task network for clas-

sification and bounding box regression. In [25] Jiao *et al.* proposed a unified CNN architecture capable of bridging SR and person re-identification (re-ID) model learning. Instead of in the image space, Tan *et al.* [18] proposed a Feature Super-Resolution GAN model that super-resolved the poor representations of LR images to highly discriminative ones. In [26], Wang *et al.* proposed a Cascaded Super-resolution GAN model, which cascaded multiple SR-GANs [27] in series for low resolution re-ID. However, one of the major drawbacks of the above-mentioned approaches is that the resultant SR images may contain serious artifacts, especially if the original LR images are of very low resolutions. In other words, the severe information loss in LR images makes it unlikely to extract sufficient recognizable features directly from LR subjects.

Resolution-invariant or scale-invariant representations have been proven to be very useful for cross-resolution recognition [16], [17], [28], [29]. Mao *et al.* [28] proposed a novel representation robust to resolution variance by jointly training a Foreground-Focus Super-Resolution module and Resolution-invariant Feature Extractor. Li *et al.* [16] proposed resolution-invariant image representations, which could recover the missing details in LR images for improving person re-ID performance. Inspired by this, Chen *et al.* [17] proposed a Resolution Adaptation and re-Identification Network, which could effectively align and extract feature representations across resolutions. In addition, Lin *et al.* [29] proposed Feature Pyramid Networks to be robust to resolution or scale variation.

The transfer learning based methods transfer external knowledge in high-resolution images to improve the performance for low-resolution object recognition. In micro-video classification, Nie *et al.* [30] presented a deep transfer model, which could transfer external sound knowledge to strengthen the low-quality acoustic modality in micro-videos. Luo *et al.* [31] proposed a significance-aware information bottlenecked transferring network for domain adaptive semantic segmentation. By transferring a significance-aware feature purified from the source domain, the method eased feature alignment, and thus significantly improved the feature-space adaptation performance. Inspired by this, Luo *et al.* [32] proposed the category-level transferring network for domain adaptive semantic segmentation. By transferring category-level data distribution from the source domain, the method adaptively weighted the adversarial loss for each feature according to how well their category-level alignment is, thus improving the feature-space adaptation performance.

Before the recognition step, [33]–[35] all designed an image super-resolution module, which super-resolved LR images into high quality images to improve the performance of recognition. Note that, image super-resolution focuses on increasing the resolution of a given image to provide better visual quality for human viewing. Instead of image super-resolution, we propose a feature enhancement module, which enhances the whole poor features of LR images by learning the discrepancy in feature space between HR and LR to narrow the gap between the two representations. [18] proposed a feature super-resolution model, which transformed the raw poor features of LR images to high quality features of their corresponding HR images. Instead of transforming the whole poor features of LR images to the high quality feature of its corresponding HR image, in our work we enhance the whole poor features of LR images by learning the discrepancy in feature space between HR and LR to narrow the gap between the two representations.

In [36], Han *et al.* proposed a Part-based Convolutional Neural Network for visual categorization, which consisted of Squeeze-and-Excitation(SE) block, Part Localization Network(PLN) and Part Classification Network(PCN), used for feature re-calibration, distinctive part localization, and image classification, respectively. In [37], Yao *et al.* proposed an efficient stacked discriminative sparse autoencoder, which learned high-level features on an auxiliary satellite image data set for the land-use classification task. In [38], Cheng *et al.* proposed a simple but effective method to learn discriminative CNNs to boost the performance of remote sensing image scene classification. In this work, we propose to explore low-resolution object recognition instead of just object recognition. The major challenge is how to recover the missing information in low-resolution images and significantly improve recognition performance for low-resolution images simultaneously.

A more direct approach is to simultaneously **transform** a LR feature map and the corresponding HR feature map into the feature maps in a common feature subspace where the distance between the two feature maps is minimized [39], [40]. Li *et al.* [41] proposed a joint multi-scale discriminant component analysis model by learning a shared representation across different scales to solve the LR person Re-ID problem. Wang *et al.* [42] made the first attempt to solve the very low resolution recognition problem using a deep learning approach. Lu *et al.* [43] presented a deep-coupled ResNet, which extracted discriminative features shared by face images of different resolutions in a trunk network. Wei *et al.* [44] presented an algorithm that learned a sparse image transformation by coupling the sparse structures of image pairs from both HR and LR spaces. Bulat *et al.* [45] proposed a multi-task deep model to simultaneously learn face super-resolution and facial landmark localization trained using a generative adversarial network (GAN).

The core idea of the representation-transforming based approaches is to narrow the gap between LR representation and HR representation. As a result, the performance in LR image classification is mainly influenced by what representation to learn and how to make use of it. In other words, the desired representation should be selectively generated with a generator from HR data, which is guided by a discriminator and an LR image classification process in a proper way. Our method proposed in this paper follows this core idea.

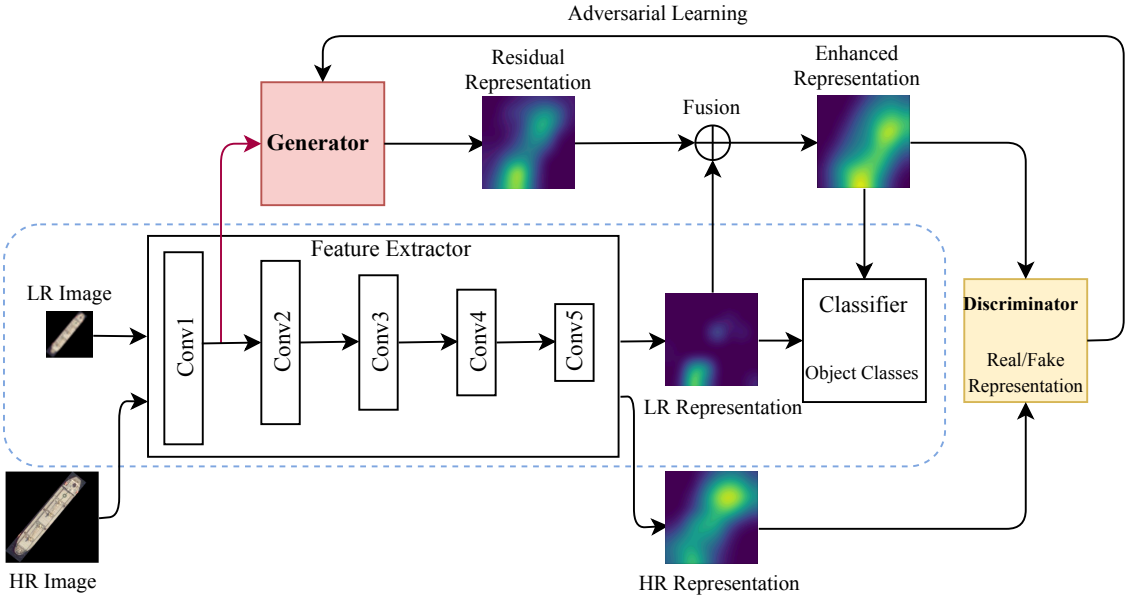


FIGURE 1: Details of the proposed *RL*-GAN. The context enclosed by the blue dotted lines is a standard CNN for object recognition. (a) The residual feature generator is a deep residual network, which takes the features from lower-level layers as input and learns the residual feature between HR and LR images in feature representation. Then, the enhanced representation is achieved through element-wise sum operation between residual and LR representation. (b) The feature discriminator takes the features of the enhanced representation (fake samples) of LR images and feature of HR images (real samples) as inputs and tries to differentiate them.

III. PROBLEM DEFINITION

We first formally define the LR object recognition problem, and then introduce our proposed residual-learning based generator and discriminator.

Let $R = \{(x_{hr}, x_{lr}, y)|x_{hr} \in I_{hr}, x_{lr} \in I_{lr}, y \in C\}$ be the training data, where $I_{hr} = \{x_{hr}^1, ..., x_{hr}^N\}$ consisting of N images are the HR images used for training and $I_{lr} = \{x_{lr}^1, ..., x_{lr}^N\}$ are their LR counterparts used for training.

Denote $S = \{(x'_{lr}, y')|x'_{lr} \in I'_{lr}, y' \in C\}$ as the testing data, where $I'_{lr} = \{x_{lr}^{'1}, ..., x_{lr}^{'M}\}$ consists of M disjoint LR testing images. Note that, although the training dataset R contains both HR images I_{hr} and LR images I_{lr}, the testing dataset S only contains LR images I'_{lr}. In this paper, we focus on how to train a mapping function $\phi(\cdot)$ on R, but test $\phi(\cdot)$ on only S to inference its label $\hat{y}'$.

Let $F_i(x)$, for $x \in I_{hr} \cup I_{lr}$, $i \in \{1, 2, ..., Q\}$ and $i \in \mathbb{Z}$, be the feature map obtained after the i-th block convolution layer. Here, Q is the index of the last block convolution layer for feature extraction. Let $\phi_C(\cdot)$ be the classification module of $\phi(\cdot)$.

Problem. Given R and S, the task of LR object recognition is to train a classifier $\hat{y} = \phi(x_{hr}, x_{lr})$, which minimizes the loss L that measures the difference between $\hat{y}$ and its ground truth label y as:

$$\hat{\theta}_\phi = \arg\min_{\theta_\phi} \frac{1}{N} \sum_{i=1}^{N} L(\hat{y}, y), \tag{1}$$

where θ_ϕ is the set of parameters of ϕ and L represents the cross-entropy loss function.

Definition 1 (Residual-Learning based Generator). Given $F_i(x)$, the task of the generator is to generate the missing features in the representation space of x_{lr} by residual learning. The generator, essentially a mapping function $G(F_i(x_{lr}))$ with the set of parameters θ_g, is to learn the residual function:

$$G(F_i(x_{lr})) \approx F_Q(x_{hr}) - F_Q(x_{lr}), \tag{2}$$

where $F_Q(x_{hr})$ and $F_Q(x_{lr})$ are the feature maps obtained after the last convolution layer of the Q-th block from the HR and LR training images, respectively.

Thus, the representation, denoted as $E_i(x)$, $x \in I_{lr}$ after the missing details being generated with the Generator G, can be represented as:

$$E_i(x) = G(F_i(x)) + F_Q(x). \tag{3}$$

Definition 2 (Adversarial-Learning based Discriminator). Following the adversarial training scheme, the discriminator, denoted as $D(x, \theta_d)$, $x \in \{F_Q(x_{hr}), E_i(x_{lr})\}$ with the set of parameters θ_d, is to learn to differentiate between the HR feature representation $F_Q(x_{hr})$ and the regenerated LR feature representation $E_i(x_{lr})$.

IV. REPRESENTATION LEARNING GAN (*RL*-GAN)

In this section, we present the details of our *RL*-GAN to ensure that feature representations learned from LR images have comparable capability with those learned from HR

images in terms of image classification. We first give a brief overview of our proposed *RL*-GAN architecture. The details of the generator G used to generate an enhanced representation from LR representation are presented. Then, we give the details of the discriminator D, which is used to differentiate generated representations from the real HR representation. In the end, we describe the testing pipeline on LR image classification for object recognition.

A. OVERVIEW

Inspired by the DCGAN in [46], we propose to perform representation learning by GANs and then reuse parts of GANs as representation enhancement for classification.

Our *RL*-GAN consists of two subnetworks, *i.e.*, a representation generator network G and a representation discriminator network D. The G network aims to map the raw representations of LR images to highly discriminative ones by discovering the latent distribution correlations between LR and HR domains, so as to narrow the gap between the representations of LR and HR. The D network estimates the probability that a representation comes from the real data or from the fake data generated by G. While maximizing the probability that a real HR representation comes from the real HR images and a generated HR representation does not come from the HR images, it actually also provides guidance for updating G. Furthermore, we propose an effective residual-learning based generator.

Formally, the generator G and discriminator D in standard GANs [47] play the following minimax two-player game:

$$\begin{aligned}\min_G \max_D V(D,G) =& \mathbb{E}_{x\sim P_{data}(x)}[logD(x)]+ \\ & \mathbb{E}_{z\sim P_z(z)}[log(1-D(G(z)))],\end{aligned} \tag{4}$$

where G learns to map data $\mathbf{z}$ from the noisy distribution $P_z(z)$ to the real data distribution $P_{data}(x)$, and D estimates the probability of a sample coming from the data distribution $P_{data}(x)$ rather than that generated by G.

In our case, $\mathbf{x}$ and $\mathbf{z}$ represent HR and LR image representations, *i.e.*, $F_Q(x_{hr})$ and $F_Q(x_{lr})$, respectively. We design a generator G to map data $\mathbf{z}$ from the LR representation distribution to the HR representation distribution as

$$G(F_Q(x_{lr})) \approx F_Q(x_{hr}), \tag{5}$$

in the feature space rather than the pixel space. Therefore, our goal becomes optimizing the minimax objective:

$$\begin{aligned}&\min_G \max_D V(D,G) = \\ &\mathbb{E}_{F_Q(x_{hr})\sim P_{data}(F_Q(x_{hr}))}[logD(F_Q(x_{hr}))]+ \\ &\mathbb{E}_{F_Q(x_{lr})\sim P_G(F_Q(x_{lr}))}[log(1-D(G(F_Q(x_{lr}))))].\end{aligned} \tag{6}$$

In [1], the hypothesis for residual learning was that, it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. Inspired by this idea, we propose a residual-learning based generator $G(F_i(x_{lr}),\theta_g)$. Rather than hoping that G directly generates the original mapping in Equation (5), we optimize G so as to fit a residual mapping. Also, considering that low-level features contain more details than high-level features, we take Equation (2), so our goal further becomes optimizing the following minimax objective:

$$\begin{aligned}&\min_G \max_D V(D,G) = \\ &\mathbb{E}_{F_Q(x_{hr})\sim P_{data}(F_Q(x_{hr}))}[logD(F_Q(x_{hr}))]+ \\ &\mathbb{E}_{F_Q(x_{(lr)})\sim P_G(F_Q(x_{lr}))}[log(1-D(\underbrace{G(F_i(x_{lr}))+F_Q(x_{lr})}_{ResidualLearning}))].\end{aligned} \tag{7}$$

B. RESIDUAL-LEARNING BASED GENERATOR

As mentioned above, the generator $G(F_i(x_{lr}),\theta_g)$ in the proposed network aims to *recover* the missing details in the representation space of x_{lr}. We obtain θ_g by optimizing the following loss function:

Adversarial Loss L_{adv}, defined by

$$L_{adv} = log(1-D(G(F_i(x_{lr}))+F_Q(x_{lr}))). \tag{8}$$

That is to say, G tries to confuse D with the generated representation by residual learning, and the L_{adv} is introduced to encourage G to produce the super-resolved representation for x_{lr} as that of x_{hr}.

Classification Loss, denoted as L_{cla}, is to guarantee that the generated representation $E_i(x_{lr})$ works well for training an image classifier, and is defined by

$$L_{cla} = \frac{1}{N}\sum_{i=1}^{N} L(\phi_C(E_i(x_{lr})), y). \tag{9}$$

MSE Loss, inspired by [45], is a strong pixel-wise constraint and is added to G, to help guide G to generate a representation, which converges to the data representation better and more efficiently. The MSE Loss is defined as

$$L_{MSE} = \frac{1}{WH}\sum_{j=1}^{W}\sum_{k=1}^{H}(F_{Q,j,k}(x_{hr}) - E_{i,j,k}(x_{lr}))^2, \tag{10}$$

where W and H are the dimensions of x_{hr}.

Thus, the overall loss function used for training G is:

$$\theta_g = \arg\min_{\theta_g}(\alpha \times L_{adv} + \beta \times L_{cla} + \gamma \times L_{MSE}), \tag{11}$$

where α, β and γ are the hyper-parameters used to control the relative importance of the corresponding losses. In our work, we set the hyper-parameters in all experiments as $\alpha = 1$, $\beta = 1$ and $\gamma = 0.5$ to emphasize the contribution of adversarial and classification losses and lower down the relative importance of the pixel-wise MSE loss.

FIGURE 2 shows the architecture of the generator G in our *RL*-GAN, which takes the features $F_i(x_{lr})$ output from the last convolutional layer of the i-th block as its input. The input $F_i(x_{lr})$ is first passed into the 9×9 convolutional filters. Its output is then fed into the 1×1 convolutional filters so that its dimension is aligned with $F_Q(x_{lr})$. Note that here we employ a large kernel to exploit more global contextual information in $F_i(x_{lr})$. Also, the core of our

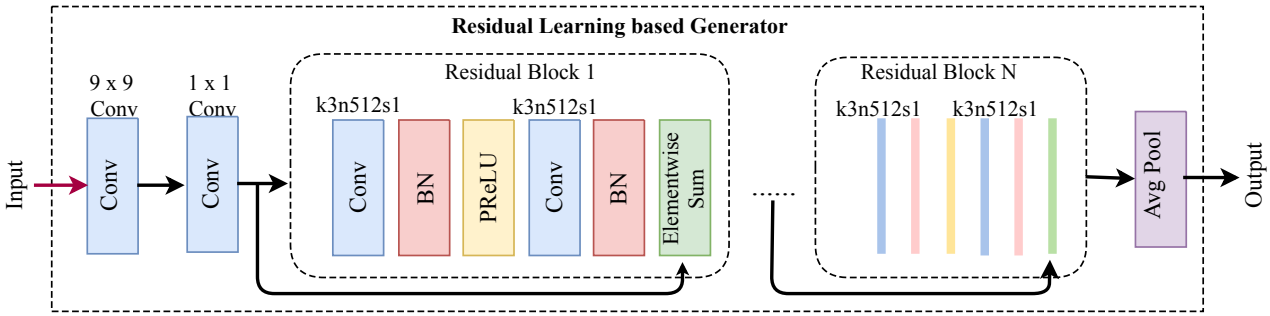


FIGURE 2: The architecture of residual-learning based generator. The input is the feature map $F_i(x_{lr})$, and its output is the residual representation $G(F_i(x_{lr}))$.

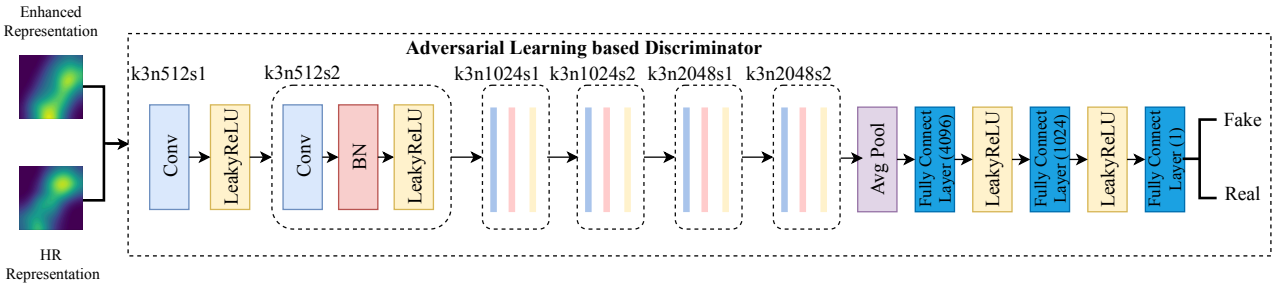


FIGURE 3: The architecture of our adversarial-learning based discriminator architecture. It attempts to differentiate between the high-resolution feature representation $F_Q(x_{hr})$ and the regenerated low-resolution feature representation $E_i(x_{lr})$.

generator G includes several cascaded residual blocks, each of which consists of two convolutional layers with small 3×3 kernels and 512 feature maps followed by batch-normalization layers and PReLU as the activation function. Then, the adaptive average pooling layer is used to resize the width and height of representation $F_Q(x_{lr})$ to be the same as $F_Q(x_{hr})$. Thus, the learned residual representation is enhanced from the representation $F_Q(x_{lr})$ for LR image classification by element-wise sum operation. Moreover, we use the features from the bottom layer of the feature extractor, because they preserve many low-level details in the feature space between $F_Q(x_{lr})$ and $F_Q(x_{hr})$.

C. ADVERSARIAL LEARNING BASED DISCRIMINATOR

The discriminator $D(x, \theta_d)$ aims to differentiate between $E_i(x_{lr})$ and $F_Q(x_{hr})$ to guide the G to produce a more realistic representation.

We obtain θ_d by optimizing the following loss function:

$$L_d = -[log(D(F_Q(x_{hr}))) + log(1 - D(E_i(x_{lr})))]. \quad (12)$$

FIGURE 3 shows the architecture of the Discriminator D, which contains seven convolutional layers with an increasing number of 3×3 filter kernels. Similar to the architecture in [27], we use LeakyReLU activation throughout the network. Strided convolutions are used to reduce the representation resolution each time the number of features is doubled. The resultant $2,048$ feature maps are followed by two dense layers and a final sigmoid activation to obtain a probability for representation classification.

D. *RL*-GAN FOR LOW-RESOLUTION IMAGE CLASSIFICATION

FIGURE 1 illustrates the pipeline of using our proposed *RL*-GAN for low-resolution image classification. Firstly, an LR image x_{lr} is fed into the Feature Extractor, which yields $F_Q(x_{lr})$ and $F_i(x_{lr})$. Then, $F_i(x_{lr})$ is passed through G, which outputs the residual representation $G(F_i(x_{lr}))$. After that, the enhanced representation $E_i(x)$ is achieved through element-wise sum operation between $G(F_i(x_{lr}))$ and $F_Q(x_{lr})$. Finally, we apply the Classifier $\phi_C(\cdot)$ on $E_i(x)$ as the final predicted label $\hat{y}$ as:

$$\hat{y} = \phi_C(F_Q(x_{lr}) + G(F_i(x_{lr}))). \quad (13)$$

V. WIDER-SHIP DATASET

As an application of our proposed *RL*-GAN for low-resolution object recognition, we aim at a rarely attempted problem of low resolution ship classification from satellite images. However, among the existing ship databases collected from various channels [4], [20], [21], [50], the DOTA dataset, a large-scale dataset for object detection in aerial images published in [4], does not provide fine-grained category information of ships; The HRSC dataset [21] identifies 16 categories of ships, but their pixel resolution is high at around

TABLE 1: Comparison of the existing ship datasets. Resolution refers to metres per pixel.

Ship Dataset	Resolution	#Category	Source
CIFAR-10 [48]	-	1	WEB
VOC2007 [49]	-	1	WEB
SeaShip [20]	-	6	CCTV
NWPU VHR [50]	2m	1	Satellite
HRSC [21]	1.1m	16	Satellite
DOTA [4]	2.5m	1	Satellite
WIDER-SHIP	0.6m $\sim$ 4.9m	3	Satellite

1m; The NWPU VHR dataset [50] contains only a limited number of ship instances; The SeaShip dataset [20] consists of a large number of ships labelled with fine-grained categories, but their images were captured at harbors rather than from satellites. Similarly, the CIFAR-10 [48] and VOC2007 datasets [49] contain CCTV images, where cameras were fixed in harbors, so there is a domain gap from our goal to test on satellite images. Furthermore, the pixel resolutions of ship instances in the aforementioned datasets are relatively high, and each pixel covers at most a $3m \times 3m$ area. Thus, for low-resolution ship classification, creating a dataset consisting of a reasonable number of ship instances and fine-grained category annotations, has become one of the main obstacles to such research.

Therefore, we create a ship dataset for LRSC, which is named "WIDER-SHIP" to highlight the large dynamic range of the pixel resolutions of images, ranging from $0.6m$ to $4.9m$ in this dataset, for ship classification. To the best of our knowledge, the WIDER-SHIP dataset is currently the first dataset for LRSC, containing a large number of ship instances and fine-grained category annotations.

To be specific, we collect 590 satellite images and fully annotated 3,077 ships using oriented bounding boxes, with three most popular ship categories, *i.e.*, Bulker, Container and Tanker. There are four levels of pixel resolutions, *i.e.*, 0.60m, 1.19m, 2.39m and 4.78m, in the dataset. Some samples of the images from this dataset are shown in FIGURE 4. FIGURE 5 presents the statistics of the spatial resolutions and orientations of the three types of ships.

In our experiments, at each resolution, we conduct 5-fold cross-validation and report the average and standard deviation of the accuracy, with 80% and 20% data for training and testing, respectively, in each round. Moreover, for fair comparison, we adopt the same evaluation metrics employed in the PASCAL VOC.

VI. EXPERIMENTS

To demonstrate the effectiveness of our proposed *RL*-GAN for ship-type classification, extensive experiments are conducted on the benchmark datasets HRSC, as well as our newly created dataset WIDER-SHIP. Moreover, to further show that our proposed approach can also be applied to general object classification, more experiments are conducted on the CIFAR-10 dataset [48] and compared with other LR image classification approaches.

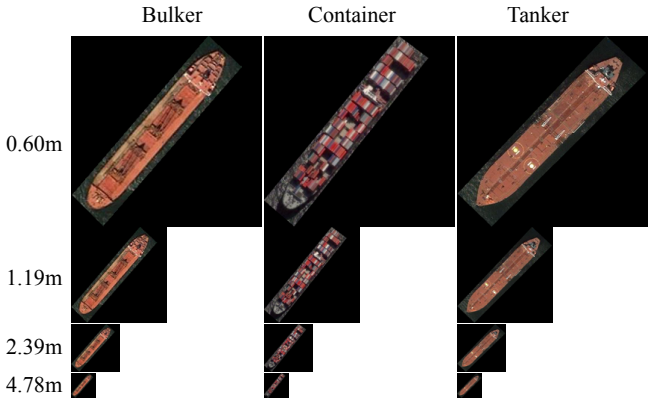
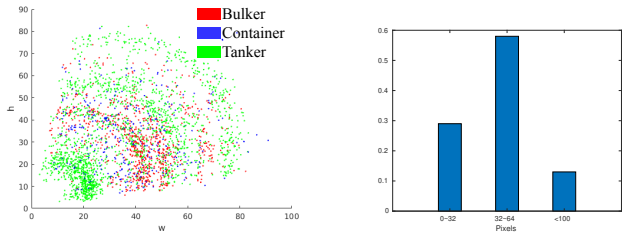


FIGURE 4: Examples of ship images in the WIDER-SHIP dataset, which consists of four levels of pixel resolution (meters per pixel), *i.e.*, 0.60m, 1.19m, 2.39m and 4.78m.



(a) Distribution of bounding boxes' width and height.

(b) Histogram of instances with respect to size.

FIGURE 5: Statistics of instances in WIDER-SHIP.

A. DATASETS AND EVALUATION

The HRSC dataset [21] contains $1,061$ images with $2,976$ ships of four categories, which are collected from Google Earth images. The resolution of HRSC is only 1.19m, not covering low resolution. To generate LR images for both training and testing, we down-sample HR images by a factor of $s = \{1, 0.5, 0.25, 0.125\}$, and normalize them into dimensions of $p \times p$, where $p \in \{128, 64, 32\}$. They are then up-scaled back to the original resolution using Nearest Neighbor (NN) interpolation to ensure sufficiently large spatial supports for the pooling layers.

The CIFAR-10 dataset [48] consists of $60,000$ 32×32 color images of 10 classes of objects, with 6,000 images per class. For each class, there are 5,000 images for training and 1,000 for testing. To compare fairly with [42] who focused themselves on low resolution object recognition, we follow the same settings as their experiments, where the original HR images are first down-scaled by $s = 0.25$ into 8×8. They are then up-scaled back to 32×32 by NN interpolation, becoming the LR images. As shown in FIGURE 6, images in the first row are high resolution, and images in the second row are their corresponding LR ones.

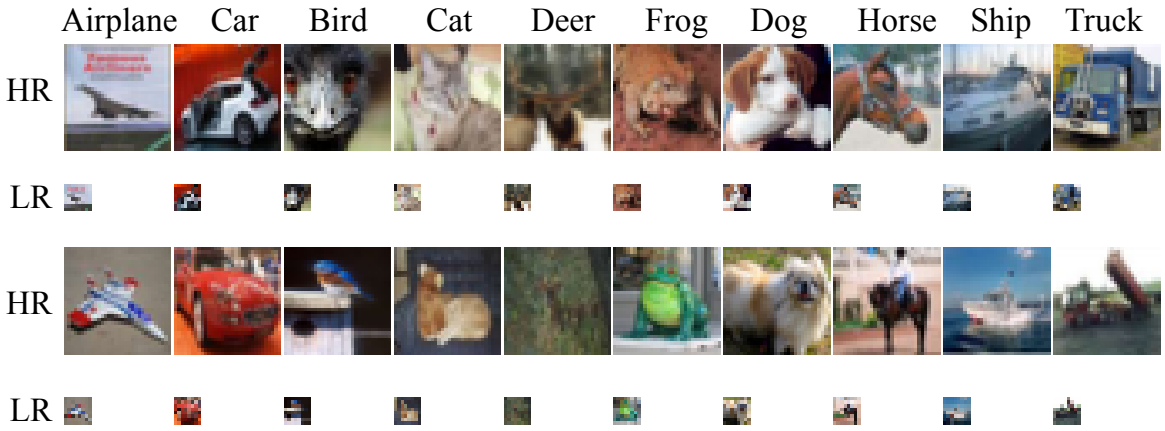


FIGURE 6: Examples of original and down-sampled images in CIFAR-10.

B. IMPLEMENTATION DETAILS

The training process is divided into two stages. First, ResNet34 is trained with the loss function in Equation (1). The learning rate is initialized as 1×10^{-4}. Secondly, we employ the trained ResNet34 model for HR and LR image feature extraction and classification. The mini-batch size is set to 16, and each mini-batch consists of 16 HR images and 16 LR ones. During this stage, to easily fit the distribution of the representation of HR images, we fix the well trained ResNet34 and optimize *RL*-GAN. For the baseline models, the total number of weighted layers of ResNet is 34, and the generator of ESRGAN contains 16 residual blocks.

For optimization, we use Adam with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We alternately update the generator and discriminator networks until the model converges. We implement our model with the PyTorch framework and train it on a single NVIDIA QUADRO P5000 GPU with 16GB RAM.

C. PERFORMANCE COMPARISON

1) WIDER-SHIP Classification

Table 2 provides the comparison of ship classification with or without our *RL*-GAN on our newly created WIDER-SHIP dataset. It can be seen from the table that, with our proposed *RL*-GAN, image classification performed on 1.19m, 2.39m and 4.78m resolutions outperform the ResNet on the same low resolution images by more than 10% (81.67%, 80% and 66% vs 71%, 65% and 55%) and even outperforms the performance of the 0.6m high-resolution images. This demonstrates the effectiveness of our *RL*-GAN in accurately classifying low-resolution images.

We further compare our solution with the SR-based methods, *i.e.*, using the ESRGAN [23] to super-resolve the LR images, to produce HR images for recognition. We first use these methods to transfer the original LR images (with resolution of 1.19m and 128×128, 2.39m and 64×64, and 4.78m and 32×32) to high-resolution images (256×256), and then we use the trained ResNet34 as the base model to test on the new images. Table 2 shows the comparison results, where there are 3-15% gains achieved with our approach.

2) HRSC Ship Classification

Similarly, Table 2 compares the ship classification accuracy with or without our proposed *RL*-GAN on the HRSC dataset. With our approach, image classification performed on the images with the resolutions of 1.19m, 2.39m and 4.78m has improved by 20-30% and the results are comparable with those on the 0.6m high-resolution images. Also note that, the SR-based approaches [23] can improve the recognition when the resolution of the input images is not very low (1.19m). However, when the original images' resolutions become too low (2.39m and 4.78m), the improvement drops significantly. On the contrary, our proposed *RL*-GAN has performed much better, especially for very low resolution images.

3) CIFAR-10 for Low Resolution Classification

Our proposed approach can also be applied to other types of objects. Note that, the existing works for LR vision are created either for different applications (*e.g.*, face or activity recognition, image retrieval, person re-ID), or do not provide codes for evaluation. Thus, we compare our approach with a benchmark representation-transforming based approach which is for LR object classification, *i.e.*, Partially Coupled Nets [42].

Table 3 provides the comparison of our approach with the Partially Coupled Nets [42], as well as three state-of-the-art classifiers, *i.e.*, DenseNet [2], MobileNetV2 [51], EfficientNet [52] in terms of classification error rate on the CIFAR-10 dataset. It can be observed that our proposed *RL*-GAN significantly reduces the classification error rate by 6.88 percentage points.

D. EFFECTIVENESS OF THE RESIDUAL-LEARNING BASED G

We compare our method with several other feature enhancement solutions, which combine low-level features, or im-

TABLE 2: Comparison of classification accuracy (%) (average of the 5-fold cross-validation ± standard variation) with or without *RL*-GAN at different image resolutions (in terms of metres per pixel) on WIDER-SHIP and HRSC datasets.

DataSet	Method	0.60m	1.19m	2.39m	4.78m
WIDER-SHIP	ResNet [1]	79.33 ± 0.59	71.67 ± 1.03	65.50 ± 1.33	55.25 ± 2.54
	Skip Connection + ResNet	-	77.50 ± 0.68	70.72 ±2.05	61.83 ±1.50
	Nearest train + ResNet	-	76.33 ± 1.66	70.25 ± 1.96	61.33 ± 1.88
	ESRGAN [23]	-	78.33 ± 1.59	70.67 ± 0.76	47± 2.05
	***RL*-GAN**	-	**82.17 ± 0.57**	**80.33 ± 0.55**	**67.33 ± 0.85**
HRSC	ResNet [1]	84.33 ± 0.88	63.42± 0.67	52.57±0.88	43.14±0.46
	ESRGAN [23]	-	83.50±0.35	74±0.95	66±2.06
	***RL*-GAN**	-	**84.29±0.24**	**83.43±0.74**	**73.14±0.98**

TABLE 3: Classification error rates obtained on the CIFAR-10 testing set.

Method	Error (%)
Partially Coupled Nets [42]	18.77
DenseNet [2]	22.36
MobileNetV2 [51]	22.28
EfficientNet [52]	26.12
RL-GAN	**11.89**

prove the image resolution by simply increasing the input scales, as shown in Table 2. In this table, "Skip Connection" indicates that the model is trained by directly combining the output of the first convolution layer to the end of the ResNet without the Residual blocks, just by using Adaptive Average Pooling Layers and 1×1 convolutional filter to ensure the same size as the end of the ResNet. "Train Nearest" represents the model trained with the interpolated images with NN algorithm of 256×256.

As shown in Table 2, at resolution of 2.39m, our generator outperforms the "Skip Connection" approach by around 10%. This shows that our method can effectively incorporate fine-grained details from low-level layers to improve image classification. Also, at the pixel resolution of 2.39m, our generator outperforms "Train Nearest" by around 9%. This shows that our method is more effective than simply increasing the scale of the input image.

We further visualize some of the generated representations as shown in FIGURE 7. The representation enhanced by our *RL*-GAN for low-resolution images are shown in the middle column. As seen from these examples, the generated enhanced representations are very similar to the representations for high-resolution ships in the fourth column. The first and the last columns are the low-resolution and their corresponding high-resolution images. The second and the fourth columns are their representations generated by the Feature Extractor. We can observe that the proposed generator successfully learns to transfer the poor representations of low-resolution images to enhanced ones similar to those of high-resolution images, validating the effectiveness of the proposed *RL*-GAN.

E. INPUTTING FEATURES FROM LOWER LAYERS TO G

The proposed G leverages fine-grained details of LR images from the representations of lower-level convolution layers. In particular, we employ the representation from "Conv1" as the inputs for learning G.

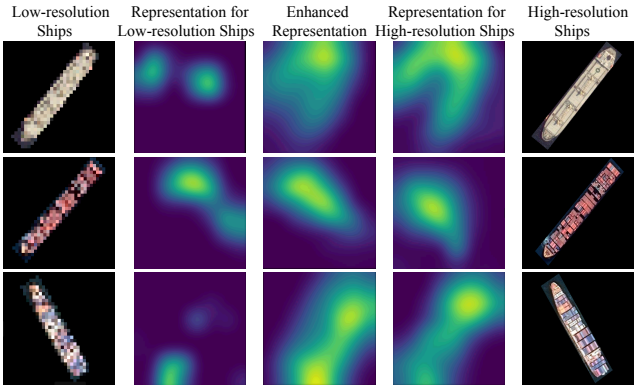


FIGURE 7: Visualization of the representations generated by *RL*-GAN and the original images.

TABLE 4: Comparisons of classification accuracy (%) of using features from different layers.

i-th convolutional layer	Accuracy (%)
Ours_conv5	68.71
Ours_conv4	71.28
Ours_conv3	74.43
Ours_conv2	77.52
Ours_conv1	80

To validate the effectiveness of this setting, we conduct extra experiments using features extracted from the "Conv2" and "Conv5" layer for learning G, respectively. As shown in Table 4, the performance decreases consistently by employing the representations output from the higher convolutional layers. The reason is that lower convolutional layers can capture more details of LR images than higher convolutional layers. Therefore, using low-level features from "Conv1" for learning the generator gives the best performance.

In general, deep features in standard CNNs evolve from general to specific along the network, and the transferability of features and classifiers decreases when the cross-domain discrepancy increases [53]. In other words, using low-level features from "Conv1" provides the best performance among all convolutional layers.

F. EFFECTIVENESS OF EACH LOSS

To analyze the effectiveness of each loss in our proposed loss function, we conduct an ablation study on the WIDER-

TABLE 5: Performance of the proposed *RL*-GAN trained with and without MSE Loss (see Eq. 10), Classification Loss (see Eq. 9) and Adversarial Loss (see Eq. 8) on WIDER-SHIP dataset.

Method	1.19m	2.39m	4.78m
***RL*-GAN**	81.67	80	66
w/o MSE Loss	79.33	77.67	63.33
w/o Classification Loss	71.67	65.33	55
w/o Adversarial Loss	76	75.67	60.33

SHIP dataset. Table 5 presents the performance of image classification at different resolutions with and without each of the proposed losses.

It can be seen that, without the MSE loss, there is no explicit supervision to guide the *RL*-GAN to perform image representation recovery. Therefore, the performance of low resolution ship-type recognition drops obviously. Once the Classification Loss is excluded, the proposed model cannot learn discriminative representation for low resolution ship-type recognition since ship-type labels are not used during training. Thus, the performance drops significantly by over 10%. When the Adversarial Loss is removed, our model does not encourage the poor-quality representations of LR images to produce realistic HR images representations any more, which results in a performance drop of about 5% in terms of classification accuracy.

Therefore, the experimental results show that the MSE loss, Classification Loss and Adversarial Loss are crucial to the whole method.

VII. CONCLUSION

In this paper, we have proposed a Representation Learning GAN to generate super image representation which is optimized for LR object recognition. By learning the latent distribution correlations between LR and HR domains, the HR feature representation becomes a guide to enhance the discriminative ability of the LR feature representation. Towards this end, we have proposed a residual-learning based generator that considers both adversarial and classification loss so as to narrow the gap between the two representations. We have also demonstrated that inputting features extracted from lower layers to the generator is most effective. Extensive experiments have demonstrated the superiority of our proposed solution over the state of the arts. The proposed method can be used to process RoIs extracted by any small object detector for more challenging applications, such as small object recognition in satellite or aerial images.

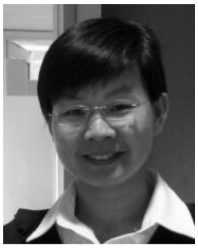
REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778.
[2] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in CVPR, 2017, pp. 4700–4708.
[3] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in CVPR, 2018, pp. 7132–7141.
[4] G.-S. Xia, X. Bai, J. Ding, Z. Zhu, S. Belongie, J. Luo, M. Datcu, M. Pelillo, and L. Zhang, "Dota: A large-scale dataset for object detection in aerial images," in CVPR, 2018.
[5] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian, "The unmanned aerial vehicle benchmark: Object detection and tracking," in ECCV, 2018.
[6] M. S. Ryoo, K. Kim, and H. J. Yang, "Extreme low resolution activity recognition with multi-siamese embedding learning," in AAAI, 2018.
[7] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," TIP, vol. 21, no. 1, pp. 327–340, 2011.
[8] P. Li, L. Prieto, D. Mery, and P. J. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," IEEE Transactions on Information Forensics and Security, vol. 14, no. 8, pp. 2000–2012, 2019.
[9] S. Ge, S. Zhao, C. Li, and J. Li, "Low-resolution face recognition in the wild via selective knowledge distillation," TIP, vol. 28, no. 4, pp. 2051–2062, 2018.
[10] M. Xu, A. Sharghi, X. Chen, and D. J. Crandall, "Fully-coupled two-stream spatiotemporal networks for extremely low resolution action recognition," in IEEE Winter Conference on Applications of Computer Vision, 2018.
[11] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in AAAI, 2017.
[12] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, "Residual dense network for image super-resolution," in CVPR, 2018, pp. 2472–2481.
[13] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, "Meta-sr: A magnification-arbitrary network for super-resolution," in CVPR, 2019, pp. 1575–1584.
[14] Y. Xi, J. Zheng, X. He, W. Jia, H. Li, Y. Xie, M. Feng, and X. Li, "Beyond context: Exploring semantic similarity for small object detection in crowded scenes," Pattern Recognition Letters, 2019.
[15] Y. Xi, J. Zheng, X. He, W. Jia, and H. Li, "Beyond context: Exploring semantic similarity for tiny face detection," in IEEE International Conference on Image Processing. IEEE, 2018, pp. 1907–1911.
[16] Y.-J. Li, Y.-C. Chen, Y.-Y. Lin, X. Du, and Y.-C. F. Wang, "Recover and identify: A generative dual model for cross-resolution person re-identification," in CVPR, 2019, pp. 8090–8099.
[17] Y.-C. Chen, Y.-J. Li, X. Du, and Y.-C. F. Wang, "Learning resolution-invariant deep representations for person re-identification," in AAAI, vol. 33, 2019, pp. 8215–8222.
[18] W. Tan, B. Yan, and B. Bare, "Feature super-resolution: Make machine see more clearly," in CVPR, 2018.
[19] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in CVPR, 2017, pp. 1222–1230.
[20] Z. Shao, W. Wu, Z. Wang, W. Du, and C. Li, "Seaships: A large-scale precisely annotated dataset for ship detection," IEEE Transactions on Multimedia, vol. 20, no. 10, 2018.
[21] Z. Liu, H. Wang, L. Weng, and Y. Yang, "Ship rotated bounding box space for ship extraction from high-resolution optical satellite images with complex backgrounds," IEEE Geoscience and Remote Sensing Letters, vol. 13, no. 8, pp. 1074–1078, 2016.
[22] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection," in ICCV, 2019, pp. 9725–9734.
[23] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in ECCV, 2018.
[24] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: Small object detection via multi-task generative adversarial network," in ECCV, 2018, pp. 206–221.
[25] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in Thirty-Second AAAI Conference on Artificial Intelligence, 2018.
[26] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded sr-gan for scale-adaptive low resolution person re-identification." in IJCAI, 2018, pp. 3891–3897.
[27] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., "Photo-realistic single image super-resolution using a generative adversarial network," in CVPR, 2017.
[28] S. Mao, S. Zhang, and M. Yang, "Resolution-invariant person re-identification," IJCAI, 2019.
[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in CVPR, 2017, pp. 2117–2125.
[30] L. Nie, X. Wang, J. Zhang, X. He, H. Zhang, R. Hong, and Q. Tian, "Enhancing micro-video understanding by harnessing external sounds,"

in Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1192–1200.
[31] Y. Luo, P. Liu, T. Guan, J. Yu, and Y. Yang, "Significance-aware information bottleneck for domain adaptive semantic segmentation," in ICCV, 2019, pp. 6778–6787.
[32] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in CVPR, 2019, pp. 2507–2516.
[33] J. Seo and H. Park, "Object recognition in very low resolution images using deep collaborative learning," IEEE Access, vol. 7, pp. 134 071–134 082, 2019.
[34] Y. Lee, J. Yun, Y. Hong, J. Lee, and M. Jeon, "Accurate license plate recognition and super-resolution using a generative adversarial networks on traffic surveillance video," in 2018 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia). IEEE, 2018, pp. 1–4.
[35] T. K. Lai, A. F. Abbas, A. M. Abdu, U. U. Sheikh, M. Mokji, and K. Khalil, "Super resolution of car plate images using generative adversarial networks," in 2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA). IEEE, 2019, pp. 80–85.
[36] J. Han, X. Yao, G. Cheng, X. Feng, and D. Xu, "P-cnn: Part-based convolutional neural networks for fine-grained visual categorization," TPAMI, 2019.
[37] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 6, pp. 3660–3671, 2016.
[38] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," IEEE transactions on geoscience and remote sensing, vol. 56, no. 5, pp. 2811–2821, 2018.
[39] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated data in person re-identification," TIP, vol. 28, no. 3, pp. 1391–1403, 2018.
[40] Y. Huang, Q. Wu, J. Xu, and Y. Zhong, "Sbsgan: Suppression of inter-domain background shift for person re-identification," in ICCV, 2019, pp. 9527–9536.
[41] X. Li, W.-S. Zheng, X. Wang, T. Xiang, and S. Gong, "Multi-scale learning for low-resolution person re-identification," in ICCV, 2015, pp. 3765–3773.
[42] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in CVPR, 2016, pp. 4792–4800.
[43] Z. Lu, X. Jiang, and A. Kot, "Deep coupled resnet for low-resolution face recognition," IEEE Signal Processing Letters, vol. 25, no. 4, pp. 526–530, 2018.
[44] X. Wei, Y. Li, H. Shen, W. Xiang, and Y. L. Murphey, "Joint learning sparsifying linear transformation for low-resolution image synthesis and recognition," Pattern Recognition, vol. 66, pp. 412–424, 2017.
[45] A. Bulat and G. Tzimiropoulos, "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans," in CVPR, 2018, pp. 109–117.
[46] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," ICLR, 2016.
[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in NIPS, 2014, pp. 2672–2680.
[48] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.
[49] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," IJCV, vol. 88, no. 2, pp. 303–338, 2010.
[50] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images," IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 12, pp. 7405–7415, 2016.
[51] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in CVPR, 2018, pp. 4510–4520.
[52] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," ICML, 2019.
[53] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in NIPS, 2014, pp. 3320–3328.

YUE XI received the B.S. degree from the Qingdao University of Technology, China, in 2011, the M.S.degree from Guizhou University, China, in 2014. He is currently pursuing the dual Ph.D. degrees in computer science with the University of Technology Sydney, Australia, and Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, image processing, machine learning, and deep learning.

JIANGBIN ZHENG received the B.S., M.S., and Ph.D. degrees in computer science from Northwestern Polytechnical University, in 1993, 1996, and 2002, respectively. From 2000 to 2002, he was a Research Assistant with The Hong Kong Polytechnic University, Hong Kong. From 2004 to 2005, he was a Research Assistant with The University of Sydney, Sydney, Australia. Since 2009, he has been a Professor and Ph.D. Supervisor with the School of Com-puter Science, Northwestern Polytechnical University. His research interestsinclude intelligent information processing, visual computing, multimediasignal processing, big data, and soft engineering. He has published over 100 peer reviewed journal/conference papers covering a wide range of topicsin image/video analytics, pattern recognition, machine learning, and big data analytics.



WENJING JIA received the Ph.D. degree in computing science from the University of Technology Sydney (UTS) in 2007. She is currently a Senior Lecturer with the Faculty of Engineering and IT and a Core Research Member with the Global Big Data Technologies Centre, UTS. She has authored over 100 quality journal articles and conference papers. Her research interests include image/video analysis, computer vision, and pattern recognition.

XIANGJIAN HE received the Ph.D. degree in computer science from the University of Technology Sydney (UTS), Australia,in 1999. He is currently a Full Professor and the Director of the Computer Vision and Pattern Recognition Laboratory, Global Big Data Technologies Centre, UTS.

HANHUI LI received the Ph.D. degree in computer software and Theory from Sun Yat-sen University, Guangzhou, China, in 2018, where he also received the B.S. degree in Computer Science and Technology in 2012. His research interests include computer vision and image processing.

ZHUQIANG REN received the B.S. degree from Xi'an Polytechnic University in 2013. He is currently pursuing the M.S. degree of Information Technology from University of Technology Sydney. His research interests include machine learning, image processing, and signal processing.

KIN-MAN LAM received the Associate ship (Hons.) in electronic engineering from The Hong Kong Polytechnic University (formerly called Hong Kong Polytechnic), in 1986, the M.Sc.degree in communication engineering from the Department of Electrical Engineering, Imperial College of Science, Technology and Medicine, London, U.K., in 1987, and the Ph.D. degree from the Department of Electrical Engineering,The University of Sydney, Sydney, Australia,in August 1996. From 1990 to 1993, he was a Lecturer with the Department of Electronic Engineering, The Hong Kong Polytechnic University. In October 1996, he joined the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, as an ssistant Professor, became an Associate Professor, in 1999, and has been a Professor since2010. His current research interests include human face recognition, image and video processing, and computer vision. He was a Member of the organizing committee and program committee of many international conferences.He was also a BoG Member of the Asia-Pacific Signal and Information Processing Association (APSIPA) and the Director-Student Services of the IEEE Signal Processing Society. He is currently a General Co-Chair of the IEEE International Conference on Signal Processing, Communications, and Computing (ICSPCC2012). He also serves as an Associate Editor of the IEEE TRANSACTIONS ON IMAGEPROCESSING, APSIPA Transactions on Signal and Information Processing, and the EURASIP International Journal on Image and Video Processing.