






DRL-GAN: Dual-Stream Representation Learning GAN for Low-Resolution Image Classification in UAV Applications

Yue Xi , Wenjing Jia , *Member, IEEE*, Jiangbin Zheng , Xiaochen Fan, Yefan Xie, Jinchang Ren , *Senior Member, IEEE*, and Xiangjian He , *Senior Member, IEEE*

Abstract—Identifying tiny objects from extremely low-resolution (LR) unmanned-aerial-vehicle-based remote sensing images is generally considered as a very challenging task, because of very limited information in the object areas. In recent years, there have been very limited attempts to approach this problem. These attempts intend to deal with LR image classification by enhancing either the poor image quality or image representations. In this article, we argue that the performance improvement in LR image classification is affected by the inconsistency of the information loss and learning priority on low-frequency (LF) components and high-frequency (HF) components. To address this LF–HF inconsistency problem, we propose a dual-stream representation learning generative adversarial network (DRL-GAN). The core idea is to produce enhanced image representations optimal for LR recognition by simultaneously recovering the missing information in LF and HF components, respectively, under the guidance of high-resolution (HR) images. We evaluate the performance of DRL-GAN on the challenging task of LR image classification. A comparison of the experimental results on the LR benchmark, namely HRSC and CIFAR-10, and our newly collected ‘WIDER-SHIP’ dataset demonstrates the effectiveness of our DRL-GAN, which significantly improves the classification performance, with up to 10% gain on average.

Index Terms—Convolutional neural networks (CNNs), generative adversarial networks, low-resolution (LR) image

Manuscript received June 27, 2020; revised August 13, 2020, September 8, 2020, September 24, 2020, and October 16, 2020; accepted November 25, 2020. Date of publication December 8, 2020; date of current version January 15, 2021. This work was supported in part by the ONR-G N62909-18-1-2169 for Creating the Dataset, National Natural Science Foundation of China, under Project 61972321 and in part by the Research and Development Plan of Shaanxi Province under Grant 2017ZDXM-GY-094 and Grant 2015KTZDGY04-01. (Yue Xi and Wenjing Jia contributed equally to this work.) (Corresponding authors: Jiangbin Zheng and Xiangjian He.)

Yue Xi is with the School of Computer Science, Northwestern Polytechnical University, Shaanxi 710072, China, and also with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: yuexi@mail.nwpu.edu.cn).

Wenjing Jia, Xiaochen Fan, and Xiangjian He are with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: wenjing.jia@uts.edu.au; xiaochen.fan@student.uts.edu.au; xiangjian.he@uts.edu.au).

Jiangbin Zheng and Yefan Xie are with the School of Computer Science, Northwestern Polytechnical University, Shaanxi 710072, China (e-mail: zhengjb@nwpu.edu.cn; hadhe145@mail.nwpu.edu.cn).

Jinchang Ren is with the Department of Electronic and Electrical Engineering, University of Strathclyde, G1 1XW Glasgow, U.K. with the National Subsea Centre, School of Computing, Robert Gordon University, AB10 7AQ Aberdeen, U.K. and also with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou 510665, China (e-mail: jinchang.ren@strath.ac.uk).

Digital Object Identifier 10.1109/JSTARS.2020.3043109

classification, representation learning, unmanned aerial vehicle (UAV)-based remote sensing.

I. INTRODUCTION

IDENTIFYING objects from low-resolution (LR) images plays critical roles for a wide range of computer vision applications, such as unmanned aerial vehicles (UAVs)-based video surveillance [1], remote sensing for Earth vision [2], and privacy-preserving video analysis [3]. With the advancement of convolutional neural networks (CNNs), many CNN-based object recognition algorithms have been presented, e.g., [4]–[8], to name a few. These algorithms attempt to extract discriminative representations, namely features, from Regions of Interests (RoIs) and to classify image into different types. Although working well on high-resolution (HR) images with sufficient details, they perform poorly on extremely LR images. Fig. 6 shows some examples of such LR images.

Image classification is the task of predicting what an image represents by assigning a class label to the image. In this work, the input image is the RoI cropped from the original LR image containing one object to be classified. The information provided inside the LR RoIs is so limited that CNNs-based object recognition algorithms have difficulty in extracting discriminative representations from the RoIs. For example, a minimum ship resolution of 16×16 in UAV remote sensing images is required for an independent recognition algorithm to perform satisfactorily in Fig. 6. It is of great difficulty to directly extract the discriminative representations of LR objects for final classification [4] [see Fig. 1(a)]. As a consequence, there is much room for improvement in the performance of recognition algorithms on LR objects, and effective solutions are still rare so far.

An intuitive solution to LR image classification is to reconstruct superresolution (SR) images from LR images [see Fig. 1(b)], and then simply apply standard classification algorithms designed for objects of high or normal resolution [11], [12]. However, the purpose of image classification is very different from that of image SR. The goal of image classification is to achieve high classification performance, whereas the goal of image SR is to improve the visual quality for human viewing by generating HR images from LR images. The generated SR images usually include distorted information and even severe artifacts, which significantly degrades the image classification performance.

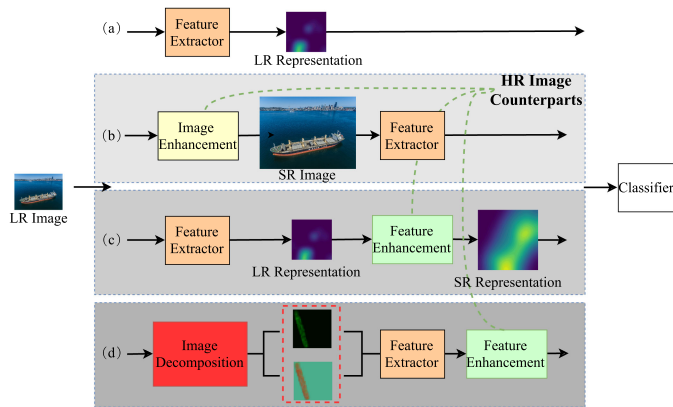


Fig. 1. Illustration of the major existing ideas attempting to address the LR image classification problem. (a) is a standard image classifier [4]. (b) is an image-enhancement based idea [9]. (c) is a representation-enhancement based idea [10]. (d) is our proposed simultaneous representation-enhancement idea.

Recently, some other efforts have been made to improve the capacity of discriminative representations with low-quality images by feature SR techniques [see Fig. 1(c)]. However, the existing representation-learning-based approaches have not considered that the information loss during the image-capturing process for different types of regions is inconsistent. Let us take remote sensing images as an example. The areas containing objects, e.g., ships, with sharp intensity changes, suffer from more severe information loss than the background areas, e.g., sea, with gentle changes in intensity. Those methods fail to effectively recover the missing information in the regions with sharper intensity changes.

Very recent studies [13], [14] found the learning bias of deep neural networks, namely, *spectral bias*, showing that a CNN with common settings would first quickly capture their dominant LF components, and then capture HF components in a relatively slow manner. Those SR-feature-based approaches are proved to be effective on recovering the missing information of the LF components. However, they are unable to recover sufficient HF-component information, which contains subtle details for differentiating images and critical for image classification. Therefore, those approaches suffer from poor classification performance.

In this article, taking into consideration of spectral bias, we formulate this problem as an HF–LF inconsistency problem, to address the inconsistent information loss and learning in HF and LF components. We show that this inconsistency has significantly degraded the final performance of LR image classification. To approach this HF–LF inconsistency problem and generate a superrepresentation effective for image classification, we, for the first time, simultaneously recover the missing information separately in both HF and LF frequency domains [see Fig. 1(d)] so that both channels can be enhanced effectively. To achieve the high classification accuracy, we propose a dual-stream representation learning generative adversarial network (*DRL-GAN*) by generating enhanced image representations optimal for LR classification. As far as we know, there has not been any previous work that considers the HF–LF inconsistency problem in representation-learning-based GANs.

Fig. 2 shows the overall view of our proposed *DRL-GAN*. In order to simultaneously recover the LF and HF components of the missing information, the input image is first decomposed into two channels, which carry LF and HF information, respectively. The representations, learned from the two channels of the HR images, are utilized to simultaneously improve the discriminative ability of the representations extracted from the corresponding channels of the LR image. The enhanced LR representations are finally fused as an enhanced image representation for recognition. Such simultaneous enhancement is essentially to superresolve the LR representations, to recover the LF and HF missing information and to make the representations more discriminative for better classification.

In summary, the article makes the main contributions as follows.

- 1) We identify the LF–HF inconsistency problem in LR image classification and propose a novel representation learning model for LF and HF domains to address this problem.
- 2) We are the first to propose a *DRL-GAN* architecture to generate enhanced image representations optimal for LR recognition.
- 3) We propose a dual-channel autoencoder (AE), which decomposes an input LR image into components in the LF and HF channels, and intends to recover the missing information in the decomposed components in LF and HF domains, respectively.
- 4) We conduct extensive experimental studies on different LR image datasets, and the results prove the effectiveness of the proposed *DRL-GAN* in improving the classification accuracy of LR image classification.

The rest of the article is organized as follows. In Section II, we first summarize the existing attempts for LR vision works. Then, we formally define the problem and the key components of our proposed model in Section III. In Section IV, we show experimental results and analysis. Finally, Section V concludes the article.

II. RELATED WORK

The existing solutions to LR vision problems can be categorized into two major streams, i.e., image and representation enhancement based methods. In the following, we first review the state-of-the-art works in the aforementioned two streams, and then introduce the emerging F-principle and spectral bias in deep learning for LR image processing.

A. Image-Enhancement-Based Methods

The image-enhancement-based methods reconstruct high-quality images from their LR counterparts so as to improve the recognition performance. In [9], [11], and [12], image SR modules were designed to convert LR images into photorealistic HR images for the final classification task. However, it is hard for the methods to obtain a solution optimized for the classification task because the SR module and the classification module are usually optimized separately. A multitask GAN-based framework was used for SR based on RoIs in [15]. Its generator was

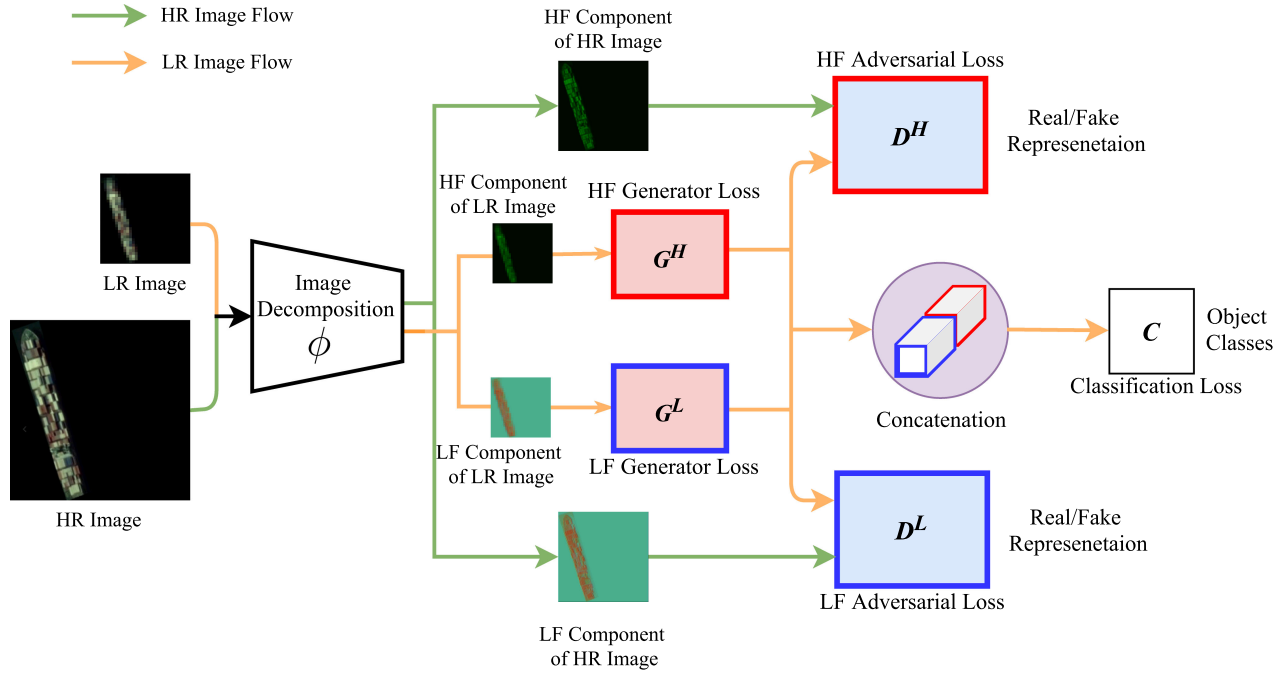


Fig. 2. Overview of the DRL-GAN. The whole framework consists of an image decomposition module ϕ , LF and HF generators G^L and G^H , the corresponding LF and HF discriminators D^L and D^H , and a classifier C . The image decomposition module is used to decompose the input LR and HR images into their low- and high-frequency components, respectively. G^L and G^H generate enhanced LF and HF components of the LR representations by recovering their missing information. D^L and D^H differentiate the enhanced LF and HF components of LR representations from their HR counterparts, respectively. The enhanced LF (blue rectangle) and HF (red rectangle) representations are concatenated to form the enhanced representation for final classification. In the HR image flow, the LF and HF components of the HR images are sent into D^L and D^H as real samples to guide the adversarial learning, respectively. In the LR image flow, the LF and HF components of the LR images are fed into G^L and G^H , respectively, to generate their corresponding enhanced representations. The enhanced LF and HF components are also fed into D^L and D^H as fake samples for adversarial learning, respectively.

an SR-based model, which was used for recovering the detailed information by converting RoIs of LR images into good-quality ones, and the discriminator was a multitask learning model used for bounding box classification and regression. They failed to integrate the context information into the generation of SR RoIs. A unified CNN architecture was proposed to bridge person reidentification (re-ID) and SR module optimization in [16]. A unified cascaded SR model was further proposed to connect several SR-based GAN models [17] for the LR person re-ID task in [18]. In [19], Yi *et al.* proposed a multitemporal ultradense memory network for video SR. They integrated long short-term memory (ConvLSTM) into ultradense residual block (UDRB) to construct an ultradense memory block (UDMB) for extracting and retaining spatiotemporal correlations. Shao *et al.* [20] presented a novel coupled sparse AE (CSAE), which leveraged the feature representation ability of both sparse decomposition and CSAE to accurately obtain the mapping relation between the LR and HR images. Nevertheless, the reconstructed images might include serious artifacts, which is the main drawback of image-enhancement-based methods. As a result, the severe information loss in LR images makes it daunting to learn sufficient discriminative representations from poor-quality RoIs.

B. Representation-Enhancement-Based Methods

The representation-enhancement-based approaches attempt to improve the discriminative capacity of representations

from LR images by introducing the resolution-invariant or representation-transforming mechanisms.

Resolution-invariant representations are crucial for cross-resolution recognition [21]–[25]. Image representations, which are invariant to the image resolution change, were proposed and used to compensate the missing details in LR images for boosting the performance of the person re-ID task in [23]. Mao *et al.* [21] jointly trained a foreground-focus SR-based module and a resolution-invariant representation extractor to obtain novel representations, which were robust to resolution variance. In addition, Chen *et al.* [22] presented a novel scale adaptation and re-ID network to perform feature alignment and extraction across different image resolutions.

The representation-transforming methods transform an HR representation and its corresponding LR representation into a shared representation space, minimizing the distance between the two representations [26]–[28]. The seminal work in [29] first attempts to address the very LR recognition problem utilizing CNNs models. Wei *et al.* [30] proposed a sparse image transformation algorithm, which coupled the sparse architecture for image pairs from an LR image and its HR counterpart. A deep-coupled ResNet was proposed to learn discriminative representations by facial images across different pixel resolutions [31]. A representation-level enhancement module was designed for LR image classification, which guided the feature enhancing procedure of the LR images by leveraging HR image representations in [32]. In [33], a multitask neural network was

presented, which introduced GANs for both face SR and facial landmark localization.

Enhancement-based approaches reduce the gap between HR representations and LR representations down to improve the performance of LR image classification. Desired representations are supposed to be selectively produced by a generator from HR data, which are guided by a discriminator. The proposed *DRL*-GAN is inspired by this key idea.

C. Spectral Bias in Deep Learning

In [34], Xu *et al.* proposed the F-principle for deep learning and argue that deep neural networks usually learn a mapping function from low to high frequencies during the training. Here, the concept of frequency refers to response frequency, i.e., the frequency of a general input–output mapping f that measures the rate of change of mapping function in terms of the intensity of every pixel.

Xu *et al.* [35] developed a theoretical framework with Fourier analysis and demonstrated that CNNs endow LR components with higher priority during the training process. In [13], they also found that for real-world datasets, CNNs would start capturing fast their dominant low-frequency (LF) components, and then capture high-frequency (HF) components in a relatively slow manner. Similarly, Rahaman *et al.* [14] highlighted the learning bias of deep neural networks, i.e., *spectral bias*. Cao *et al.* [36] further presented a more comprehensive explanation for spectral bias by relating it with the kernel function. Based on the previous theoretical exploration of F-principle, different deep learning frameworks have been proposed, such as MuffNet for mobile deep learning [37], DeepXDE for physics informed neural networks, and the frequency-aware reconstruction of fluid simulations [38].

In this work, we take into consideration of the spectral bias and address the HF–LF inconsistency problem for LR image classification. Different from the previous studies, we mainly focus ourselves on improving the generated representations by simultaneously recovering information in both HF and LF domains.

D. Machine Learning for LR Image Classification in UAV Vision

The machine learning research on UAV vision is an evolving research area, which attempts to bring UAV into human capabilities for image sensing and image understanding. A typical approach to classify LR images is through learning the representations of all the objects at cross-resolution UAV images. This approach is, however, highly inefficient with limited performance gains. Wu *et al.* [39] designed nuisance disentangled feature transform and integrated it into an image classification framework for UAV vision. Cao *et al.* [40] proposed an SR algorithm using coupled dictionary learning to transfer the target region into an HR counterpart to ‘augment’ its visual appearance. Liu *et al.* [41] proposed to internally superresolve the feature maps of LR objects to make them resemble similar characteristics as HR objects.

III. METHODOLOGY

We start to define the identified LF–HF inconsistency problem in LR image classification. We then propose the *DRL*-GAN architecture for LR image classification. Finally, we present the details of the dual-channel AE, which is leveraged to decompose input images into LF and HF components.

A. LF–HF Inconsistency Problem

We focus on the LF–HF inconsistency problem in LR image classification. The goal is to learn the *DRL*-GAN classifier C , which predicts the labels for LR images. Our core idea is to generate enhanced image representations optimized for LR classification, by simultaneously recovering the LF and HF components of the missing information under the guidance of HR images. Towards this end, a dual-channel AE is first introduced to decompose the input image into the LF and HF channels. To facilitate the following discussion, we divide C into an AE, a generator G and a discriminator D .

Let $R = \{(x_{\text{hr}}^i, x_{\text{lr}}^i, y_i) | x_{\text{hr}}^i \in I_{\text{hr}}, x_{\text{lr}}^i \in I_{\text{lr}}, y_i \in Y, i = 1, 2, \dots, N\}$ be the training data, where $I_{\text{hr}} = \{x_{\text{hr}}^1, \dots, x_{\text{hr}}^N\}$ represents the set of N HR images used for training, $I_{\text{lr}} = \{x_{\text{lr}}^1, \dots, x_{\text{lr}}^N\}$ is the set of corresponding N LR images, and $Y = \{y_1, \dots, y_N\}$ is the set of corresponding ground truth labels of the training images.

Let $S = \{(x_{\text{lr}}^{j'}, y_j^{j'}) | x_{\text{lr}}^{j'} \in I_{\text{lr}}', y_j^{j'} \in Y', j = 1, 2, \dots, M\}$ be the testing data, where $I_{\text{lr}}' = \{x_{\text{lr}}^{1'}, \dots, x_{\text{lr}}^{M'}\}$ consists of M disjoint LR testing images, independent from the training data, and $Y' = \{y_1', \dots, y_M'\}$ is the set of corresponding class labels. There are only the LR images in the testing dataset S , but LR and HR images in the training set R .

Problem: Given R , LR image classification aims to train a classifier C , which minimizes the loss \mathcal{L} that measures the difference between the set of the learned labels $\{\hat{y}_i | i = 1, 2, \dots, N\}$ and the set of the corresponding ground truth labels $\{y_i | i = 1, 2, \dots, N\}$, i.e.

$$\hat{\theta}_C = \arg \min_{\theta_C} \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\hat{y}_i, y_i) \quad (1)$$

where θ_C represents various sets of parameter values of C , \mathcal{L} is the cross-entropy loss function, and $\hat{\theta}_C$ is the set of optimal parameter values of C .

Definition 1. (Dual-Channel AE): The Dual-Channel AE, denoted by *AE*, consists of an encoder and two decoders. It is a model learned with a parameter set θ_{AE} to decompose an input image into two images in LF and HF channels (or domains), Ch^L and Ch^H , respectively, using the encoder, and reconstruct the image by merging the images decoded from the two decomposed images by the two decoders. Let $\phi(x)$ represent the decomposition function (i.e., the encoder) that decomposes any input image $x \in I_{\text{hr}} \cup I_{\text{lr}}$ to an LF component $\text{Ch}^L(x)$ in Ch^L , and an HF component $\text{Ch}^H(x)$, in Ch^L and Ch^H , respectively, as shown in Fig. 2, i.e.

$$\phi(x) = [\text{Ch}^L(x), \text{Ch}^H(x)]. \quad (2)$$

We target on retaining most of the contents in the original image x after the decomposition.

Definition 2. (LF- and HF-based Generator): The LF- and HF-based generator, denoted by G , is a model learned with a parameter set θ_G to recover the missing information of $\phi(x_{lr}^i)$ for any $x_{lr}^i \in I_{lr}$, where $i = 1, 2, \dots, N$. Let $G^H(\cdot)$ and $G^L(\cdot)$, as shown in Fig. 2, represent the functions (namely, HF generator and LF generator) that recover the missing information and produce the enhanced images, $G^H(\text{Ch}^H(x_{lr}^i))$ and $G^L(\text{Ch}^L(x_{lr}^i))$ in the $G^H(\text{Ch}^H)$ and $G^L(\text{Ch}^L)$ domains, respectively, for $i = 1, 2, \dots, N$.

Definition 3. (Adversarial-Learning-Based Discriminator): The Adversarial-learning-based discriminator D is a model learned with a parameter set θ_D to tell the difference between the HR feature representation $\text{Ch}^H(x_{hr}^i)$ and the regenerated LR feature representation $G^H(\text{Ch}^H(x_{lr}^i))$ in the HF domain, and between the HR feature representation $\text{Ch}^L(x_{hr}^i)$ and the regenerated LR feature representation $G^L(\text{Ch}^L(x_{lr}^i))$ in the LF domain, where $i = 1, 2, \dots, N$. Let $D^H(\cdot)$ and $D^L(\cdot)$ represent the HF and LF components of D (namely, HF discriminator and LF discriminator), as shown in Fig. 2.

The learned $G^H(\text{Ch}^H)$ and $G^L(\text{Ch}^L)$ will then be concatenated (i.e., joined together as one 2-D feature map) for the final classification.

B. Dual-Stream Representation Learning GAN

The inspiration for our DRL-GAN is the observation that after the process of image downsampling, the HF discrepancy of images in Ch^H is larger than the LF discrepancy of the corresponding images in Ch^L . We explain the reason from the view of digital signal processing. The signal mainly loses much more severe information in the HF domain. In addition, according to the F-principle stated in the very recent studies on the training process of a CNN [13], [34], [35], a CNN starts to fast learn the LF components, but it relatively slowly learns their HF components. Our proposed DRL-GAN aims to address the LF–HF inconsistency problem by simultaneously recovering LF and HF components of missing information.

Inspired by DCGAN [42], we intend to perform representation learning in LF and HF domains with GAN and reuse it for representation enhancement.

As shown in Fig. 2, our DRL-GAN consists of an AE, a G network containing two representation generators G^L and G^H , and a D network containing two representation discriminators D^L and D^H . In the LF or HF channel, the network G discovers the underlying distribution relationships between the HR images and the LR images to convert the poor-quality representations of the LR images into the highly discriminative ones, which bridges the gap between the HR representations and the LR representations. The network D is used to calculate the probability that representations are sampled from the distribution of the fake data generated by G and the real data. D offers the guidance for updating G , so as to maximize the probability that resultant representations do not come from the HR images but from the real HR representations of the corresponding HR images. Moreover, we give the details of dual-stream representation learning generator as shown ahead.

Mathematically, a discriminator D and a generator G of a standard GAN would adopt the scheme of the following mini-max two-player game [43], i.e.

$$\min_G \max_D V(D, G) := \mathbb{E}_{\mathbf{x} \sim P_{\text{data}}(\mathbf{x})} [\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim P_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G(\mathbf{z})))] \quad (3)$$

where G is trained to learn the underlying relationship between data \mathbf{z} from the noisy distribution $P_{\mathbf{z}}(\mathbf{z})$ and the real data distribution $P_{\text{data}}(\mathbf{x})$, and D is used to calculate the probability of a sample coming from the data distribution $P_{\text{data}}(\mathbf{x})$ rather than that generated by G .

In this article, let \mathbf{x}^L and \mathbf{x}^H represent the LF and HF component of an HR image $x_{hr} \in I_{hr}$, respectively. Let \mathbf{z}^L and \mathbf{z}^H represent the LF component and HF component of the corresponding LR image $x_{lr} \in I_{lr}$, respectively. Following the ideas of producing a generator in GANs, the generators G^L and G^H are designed to map \mathbf{z}^L and \mathbf{z}^H to \mathbf{x}^L and \mathbf{x}^H , respectively, i.e.

$$\begin{aligned} G^L(\mathbf{z}^L) &\approx \mathbf{x}^L \\ G^H(\mathbf{z}^H) &\approx \mathbf{x}^H \end{aligned} \quad (4)$$

in the feature space rather than the pixel space. Finally, we aim to optimize the following mini-max objectives:

$$\begin{aligned} &\min_{G^L} \max_{D^L} V^L(D^L, G^L) \\ &:= \mathbb{E}_{\mathbf{x}^L \sim P_{\text{data}}(\mathbf{x}^L)} [\log D^L(\mathbf{x}^L)] \\ &+ \mathbb{E}_{\mathbf{z}^L \sim P_{\mathbf{z}^L}(\mathbf{z}^L)} [\log(1 - D^L(G^L(\mathbf{z}^L)))] \end{aligned} \quad (5)$$

and

$$\begin{aligned} &\min_{G^H} \max_{D^H} V^H(D^H, G^H) \\ &:= \mathbb{E}_{\mathbf{x}^H \sim P_{\text{data}}(\mathbf{x}^H)} [\log D^H(\mathbf{x}^H)] \\ &+ \mathbb{E}_{\mathbf{z}^H \sim P_{\mathbf{z}^H}(\mathbf{z}^H)} [\log(1 - D^H(G^H(\mathbf{z}^H)))] \end{aligned} \quad (6)$$

C. Dual-Channel Image Decomposition

To recover simultaneously the HF and LF information, we first decompose input images to LF and HF components. Traditional approaches, such as wavelet transformation [44] can be used for this purpose. However, the LF and HF channels decomposed using the wavelet will lose some information from the input image, since there is not a reconstruction stage of the decomposed image. The missing information significantly degrades the classification performance, as shown in Table I. In addition, wavelet decomposition and its subsequent feature enhancement module are optimized separately, so the decomposed channels do not necessarily result in the best classification performance.

The section describes the details of a dual-channel AE. The classical AE [45] and its recent variants are pairs of encoders and decoders for learning efficient representations, as described ahead.

The pipeline of a standard AE is the branch on the left-hand side with black arrows in Fig. 3. The encoder ϕ and decoder φ are trained to obtain efficient representations by minimizing a reconstruction loss, which is formalized as an unsupervised

TABLE I
WIDER-SHIP: COMPARISON OF CLASSIFICATION ACCURACIES (%) OF BENCHMARK MODELS WITH *DRL*-GAN WITH VARIOUS IMAGE RESOLUTIONS (IN TERMS OF METRES PER PIXEL)

Method	0.60m	1.19m	2.39m	4.78m
ResNet [4]	91.00	83.33	75.33	57.67
DenseNet [5]	91.50	84.00	76.67	58.33
Wavelet + CNN [44]	86.33	81.67	74.53	57.00
Wavelet + CNN [44]	86.33	81.67	74.53	57.00
Nearest train + ResNet	-	84.67	77.33	67.67
ESRGAN [9]	-	88.33	63.33	43.33
EEGAN [54]	-	86.67	60.67	42.67
<i>RL</i> -GAN [47]	-	89.52	87.53	68.33
Our <i>DRL</i>-GAN	-	90.00	89.67	76.33

The bold values indicate the performance of proposed method.

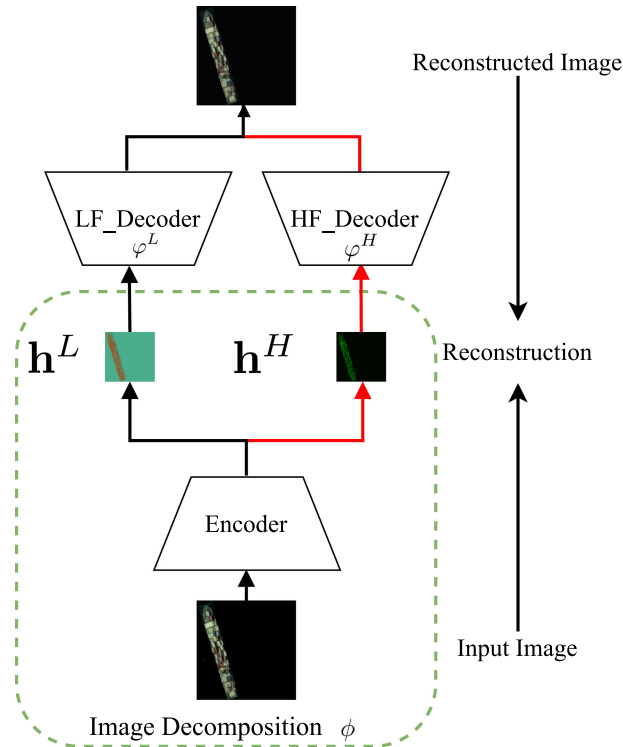


Fig. 3. Dual-channel AE for image decomposition. We use an AE module to decompose an input image into two channels, which carry LF and HF information, respectively, and to reconstruct the input image using the LF and HF decoders. Black arrows in left subfigure indicates the process of an standard AE.

problem

$$\begin{aligned} \phi &: \mathcal{X} \rightarrow \mathcal{F} \\ \varphi &: \mathcal{F} \rightarrow \mathcal{X} \\ \phi, \varphi &= \arg \min_{\phi, \varphi} \mathcal{L}_{\text{rec}}(\phi, \varphi). \end{aligned} \quad (7)$$

In (7), the encoder ϕ takes the input $\mathbf{x} \in \mathbb{R}^d = \mathcal{X}$ and maps it to $\mathbf{h} \in \mathbb{R}^p = \mathcal{F}$, \mathcal{X} is the input space, and \mathcal{F} is the representation space. Similarly, the decoder φ maps \mathbf{h} to \mathbf{x}' , which is the reconstruction of \mathbf{x} . The AE is trained to minimize the reconstruction loss

$$\mathcal{L}_{\text{rec}} = \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}'\|_2 = \sum_{\mathbf{x}} \|\mathbf{x} - (\phi \circ \varphi)\mathbf{x}\|_2$$

where $\|\cdot\|_2$ represents the L_2 norm.

We design a dual-channel AE, which adopts the information constraints on the two channels as follows. First, an input image is encoded in LF and HF channels to carry low- and high-frequency information (the lower part of Fig. 3), respectively, and is enabled to be reconstructed through a decoding process (the upper part of Fig. 3) when needed. Second, the HF channel should carry minimum information.

As shown in Fig. 3, given the LF decoder φ^L acting on LF channel CH^L and the HF decoder φ^H acting on the HF channel CH^H , the standard AE as shown in (7) is modified to

$$\begin{aligned} \phi &: \mathcal{X} \rightarrow \mathcal{F} \\ \varphi^L &: \mathcal{F} \rightarrow \mathcal{X} \\ \varphi^H &: \mathcal{F} \rightarrow \mathcal{X} \\ \phi, \varphi^L, \varphi^H &= \arg \min_{\phi, \varphi^L, \varphi^H} \mathcal{L}_{\text{rec}}(\phi, \varphi^L, \varphi^H) \end{aligned} \quad (8)$$

where the new AE is trained to minimize the reconstruction loss redefined by

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \sum_{\mathbf{x}} \|\mathbf{x} - \mathbf{x}'\|_2 \\ &= \sum_{\mathbf{x}} \|\mathbf{x} - ((\phi \circ \varphi^L)\mathbf{x} + (\phi \circ \varphi^H)\mathbf{x})\|_2 \end{aligned} \quad (9)$$

to ensure that \mathbf{h}^L and \mathbf{h}^H , representing the encoded results of \mathbf{x} in CH^L and CH^H , respectively, are expected to possess minimum information loss and retain almost all contents of the input image as it is invisible to the subsequent classifier. In addition, to ensure \mathbf{h}^H to carry minimum HF information, we adopt an energy constraint (or loss) not only to minimize its energy but also to push other information to \mathbf{h}^L . The energy loss can be formulated by

$$\mathcal{L}_e = \sum_{\mathbf{x}} \|\mathbf{h}^H\|_2^2. \quad (10)$$

We combine the above two losses on the decomposed and reconstructed results, to form the loss function of the AE by

$$\mathcal{L}_t = \mathcal{L}_{\text{rec}} + \lambda \mathcal{L}_e \quad (11)$$

where λ is set to 1 in our experiments.

More details of the dual-channel AE architecture are provided as follows. Inspired by the waveletlike AE [46], the encoder ϕ consists of three convolutional layers with their strides of 1, which are followed by two branches of convolutional layers with strides of 1 to generate, respectively, \mathbf{h}^L and \mathbf{h}^H . For an efficient computation, the small kernels with sizes of 3×3 are utilized and the numbers of all intermediate layers channels are set to 16. The decoding module consists of two branches, which share the same architecture, and each branch performs one transform. We use a deconvolutional layer with a stride of 1, followed by three stacked convolutional layers with strides of 1 in each branch. The decoding module further refines the upsampled feature maps so as to produce the final reconstructed image.

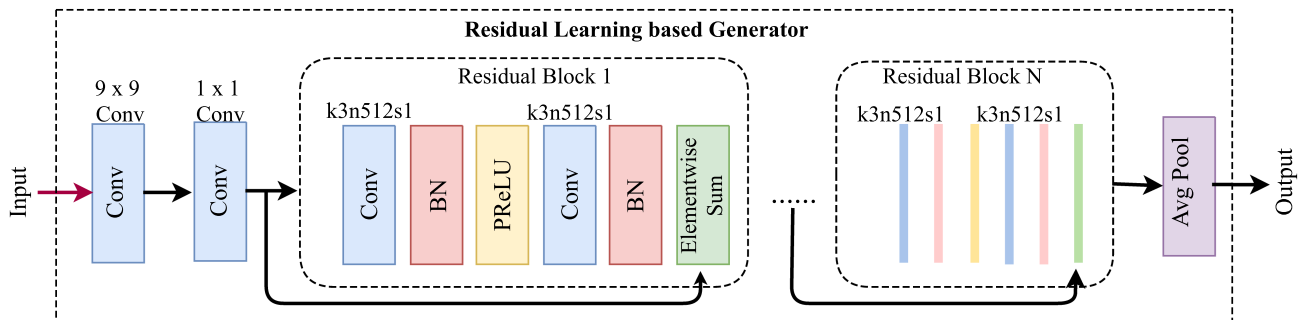


Fig. 4. Architecture of residual-learning-based generator. The input is the feature map $Ch^H(x_{lr}^i)$ or $Ch^L(x_{lr}^i)$, and its output is the residual representation $G^H(Ch^H(x_{lr}^i))$ or $G^L(Ch^L(x_{lr}^i))$.

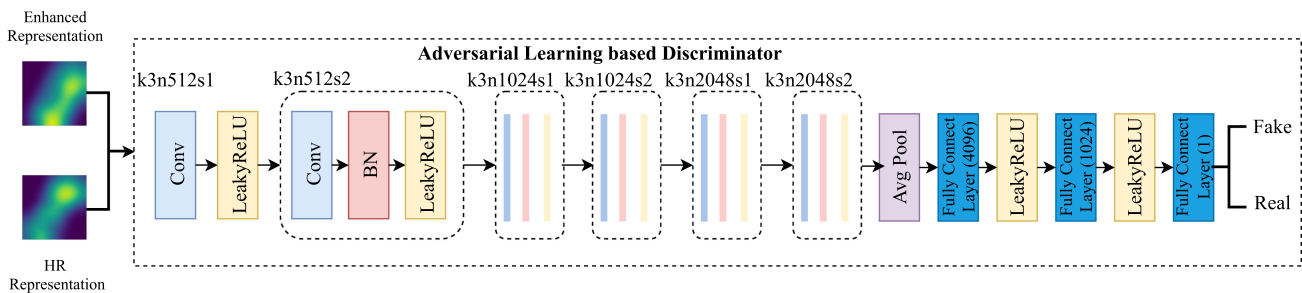


Fig. 5. Architecture of our adversarial-learning-based discriminator architecture. It attempts to differentiate between the HR feature representation and the regenerated LR feature representation.

D. Generators and Discriminators

The architecture of the generator G in our RL -GAN is shown in Fig. 4, which takes the decomposed representations $Ch^H(x_{lr}^i)$ and $Ch^L(x_{lr}^i)$ from the last convolutional layer of the i th block as its input. The input is first passed into the 9×9 convolutional filters. Its output is then fed into the 1×1 convolutional filters. Note that here we employ a large kernel to exploit more global contextual information in the input representation. Also, the core of our 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, an adaptive average pooling layer is used to resize the width and height of the output of residual blocks to be the same as those of the input. Thus, the input representation is enhanced by the Generators for LR image classification.

The architecture of the discriminator D is presented in Fig. 5, which contains seven convolutional layers with an increasing number of 3×3 filter kernels. Similar to the architecture in [17], 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 2048 feature maps are followed by two dense layers and a final sigmoid activation to obtain a probability for representation classification.

IV. EXPERIMENTS

In order to demonstrate that the DRL -GAN can effectively classify LR ships into different types, we conduct extensive

experiments on a newly created dataset WIDER-SHIP [47] and the benchmark dataset HRSC [48]. Moreover, to further verify the scalability for general object classification tasks, we conduct more experiments on the CIFAR-10 dataset [49] and achieve better performance than the recent state-of-the-art methods.

A. Datasets

The WIDER-SHIP benchmark [47] consists of 590 satellite images, with 3077 ships annotated with oriented bounding boxes. The dataset contains three most common ship subcategories, namely Tanker, Container, and Bulker. The dataset consists of four levels of image pixel resolutions, namely 0.60, 1.19, 2.39, and 4.78 m. Fig. 6 shows some ship image samples of different categories and pixel resolutions. The pixel resolutions of the testing set are 1.19, 2.39, and 4.78 m. At each resolution, we randomly select 80% and 20% data for training and testing, respectively. We choose the common evaluation metrics, which are adopted on the PASCAL VOC benchmark for a fair comparison in our experiments. In the article, we mainly focus on image classification. Specifically, the input image to an image classifier is cropped by our proposed ship detector [50] and contains only one ship.

The HRSC dataset [48] is collected from the Google Earth and consists of 1061 images with 2976 ships annotations of four categories. The spatial resolution of HRSC is 1.19 m, which is a relatively HR but not covering LR. HR images are downsampled by a factor of $s \in \{1, 0.5, 0.25, 0.125\}$ and normalized to $p \times p$, $p \in \{32, 64, 128\}$, to produce LR images for training. The pixel resolutions of the testing set are 1.19, 2.39, and 4.78 m. Similar

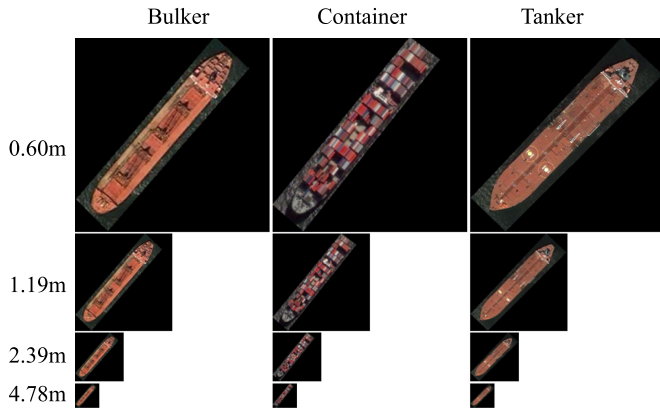


Fig. 6. Some samples of ship images in the WIDER-SHIP benchmark, which contains images with four levels of pixel resolutions (meters per pixel), namely 0.60, 1.19, 2.39, and 4.78 m.

to the WIDER-SHIP dataset, at each resolution, we randomly select 80% and 20% data for training and testing, respectively. The downsampled LR images are then upsampled to the size of the original HR images with the nearest neighbor (NN) interpolation algorithm to ensure sufficient space supporting for the pooling layers.

The CIFAR-10 benchmark [49] contains 60 000 32×32 RGB images of 10 categories including airplane, automobile, bird, etc. There are 6000 images in each category with and 1000 images for testing and 5000 images for training. To make a fair comparison with the work in [29] focusing on LR image classification, we adopt the common experimental settings in [29], which first downsample the original HR images by a factor of $s = 0.25$ to 8×8 . The down-scaled images are then up-scaled back to the 32×32 images with NN interpolation to the LR version.

B. Implementation Details

The training procedure of *DRL-GAN* is composed of three stages as follows. We start to train the image decomposition ϕ with the reconstruction loss \mathcal{L}_{rec} on the final generated images and the energy loss \mathcal{L}_e defined in (10) on the HF channel. First, we initialize the learning rate with 2×10^{-4} and decay it by a factor of 2 in 2×10^5 iterations. Second, we fix parameters of the well trained ϕ and train the generator and discriminator using (5) and (6). Finally, the pretrained ResNet34 model [4] is used for extraction and classification in the LF and HF channels. The pretrained model is available at: <https://download.pytorch.org/models/resnet34-333f7ec4.pth>. The minibatch size is set to 16 with 16 HR images and 16 LR ones in each minibatch. We use Adam [51] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ to optimize the *DRL-GAN* by alternately updating the generator and discriminator modules until the model converges.

C. Performance Comparison and Discussion

In this work, our proposed *DRL-GAN* is compared with several state-of-the-art LR image classifiers [4], [5], [9], [29], [47], [52], [53] on the benchmark datasets WIDER-SHIP [47], HRSC [48], and CIFAR-10 [49]. Among these, both ResNet [4] and DenseNet [5] are highly cited models widely used for

image classification tasks. The ESRGAN [9] and EEGAN [54] approaches are used to superresolve the LR images and produce HR images for classification. The RL-GAN [47] is the first time to design a feature enhancement model using GAN architecture for LR image classification. In our experiments, all hyperparameters are adopted from the settings of their original publications. In particular, the total number of weighted layers in ResNet is 34, and the generator of ESRGAN contains 16 residual blocks.

We implement a baseline method, which utilizes the decomposed channels as input to the ResNet for classification. Wavelet + CNN represents that discrete wavelet transforms (DWTs) decompose an input image to four half-resolution channels, i.e., cA, cH, cV, and cD. Here, cA is an approximation to the input image, and cH, cV, and cD preserve image details. The original image can be reconstructed based on cA, cH, cV, and cD using an inverse transform. Then, cA is fed into the standard network, and cH, cV, and cD are concatenated for final classification. Here, we use the widely used implementation in [44].

1) *Ship Classification on WIDER-SHIP*: Table I provides a comparison of benchmark models for ship classification with our *DRL-GAN* on the WIDER-SHIP dataset. The table shows that our *DRL-GAN* outperforms all baseline models significantly by 6%–18%. The improvement is especially significant when the pixel resolution further deteriorates to 2.39 and 4.78 m, where there is an improvement over 10% on classification accuracy to both the ResNet [4] and DenseNet [5] (89.67% and 76.33% versus 75.33% and 57.67% compared with the ResNet, and 89.67% and 76.33% versus 76.67% and 58.33%) compared with the DenseNet. More remarkably, our *DRL-GAN* has achieved a classification accuracy of 90% at 1.19-m resolution, which is comparable to the classification accuracy obtained with ResNet (91%) and with DenseNet (91.5%) but at much higher resolution (i.e., 0.6-m resolution).

To validate the effectiveness of the proposed *DRL-GAN*, we also compare with “Wavelet + CNN.” The results show that our model outperforms the baseline method by a large margin on the classification performance. Besides, we further compare our solution with the SR-based methods, i.e., training the ESRGAN [9] and EEGAN [54] to superresolve the LR ship images to produce HR ship images for recognition. We first use these methods to transfer the original LR images (with resolutions of 1.19 m and 128×128 , 2.39 m and 64×64 , and 4.78 m and 32×32) to HR images (256×256), and then, we use the trained ResNet34 as the base model to test on the new images. Table I shows the comparison results, where there are 3%–15% gains achieved with our approach. The earlier experimental results demonstrate the effectiveness of *DRL-GAN* in accurately classifying LR images.

2) *Ship Classification of HRSC*: Similarly, Table II compares the performance of our proposed *DRL-GAN* method and benchmark methods in ship classification accuracy on the HRSC dataset. The table shows that our proposed *DRL-GAN* outperforms all other models and improves the classification performance on all four spatial resolutions, namely 1.19, 2.39, and 4.78 m, remarkably by 20%–30%. In particular, at pixel resolutions of 2.39 and 4.78 m, our *DRL-GAN* outperforms the ResNet [4] and DenseNet [5] by over 25% (87.96% and 75.48% versus 56.67% and 47.54% compared to ResNet, and 87.96% and 75.48% versus 58.56% and 49.63% compared to DenseNet).

TABLE II
COMPARISON OF CLASSIFICATION ACCURACIES (%) OF BENCHMARK MODELS ON THE HRSC DATASET WITH OUR DRL-GAN AT VARIOUS IMAGE RESOLUTIONS (I.E., METRES PER PIXEL)

Method	0.60m	1.19m	2.39m	4.78m
ResNet [4]	88.76	66.82	56.67	47.54
DenseNet [5]	89.15	68.95	58.56	49.63
Nearest Neighbor	-	87.54	79	66.61
ESRGAN [9]	-	87.68	76.80	67
EEGAN [54]	-	85.33	74.86	64.62
RL-GAN [47]	-	87.75	85.56	67.15
Our DRL-GAN	-	88.53	87.96	75.48

The bold values indicate the performance of proposed method.

TABLE III
COMPARISON OF CLASSIFICATION ERROR RATES ON THE CIFAR-10 TESTING SET

Method	Error (%)
Partially Coupled Nets [29]	18.77
DenseNet [5]	22.36
MobileNetV2 [52]	22.28
EfficientNet [53]	26.12
RL-GAN [47]	11.89
Our DRL-GAN	9.48

The bold values indicate the performance of proposed method.

It is also worth noting that the result obtained with our DRL-GAN at the 1.19-m resolution is comparable with that of ResNet and DenseNet obtained at the 0.6-m HR (88.53% versus 88.76% and 89.15%). While the SR-based approach [9] could provide a classification performance improvement when the resolution of the input images is not very low (1.19 m), the improvement drops significantly when the resolutions are much lower (e.g., 2.39 m and 4.78 m). However, our proposed method performs much better for very LR images.

3) *LR Classification of CIFAR-10*: The DRL-GAN could be used for other types of objects. The existing works for LR vision either do not supply codes for comparison, or are created for different applications (namely person re-ID and image retrieval, activity recognition, or face). Therefore, our proposed method is compared with the benchmarks of representation-enhancement-based approaches focusing on LR classification, namely partially coupled nets [29] and RL-GAN [47], and three state-of-the-art classifiers, i.e., EfficientNet [53], MobileNetV2 [52], and DenseNet [5] with regard to classification error rate. Table III provides the comparison results that the DRL-GAN greatly decreases the classification error rate by 2.41%, compared with RL-GAN.

D. Ablation Study

To investigate the effectiveness of different components on the overall performance, we conduct a comprehensive ablation study by disabling each component and compare the performance.

1) *Effectiveness of Image Decomposition ϕ : Visualization of the decomposed two channels*. We show the generated representations in Fig. 7. The second column is the reconstructed images,

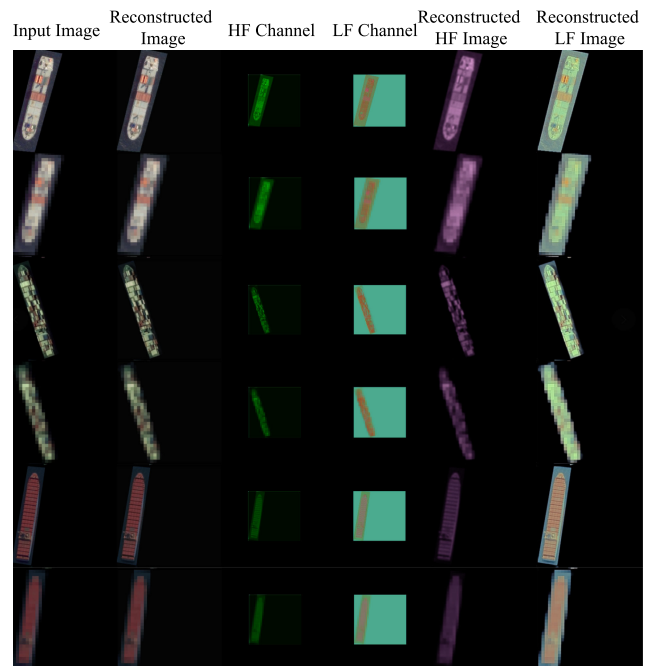


Fig. 7. Samples of original images, reconstructed images, and decomposed and reconstructed images in HF and LF channels, respectively.

which show excellent reconstruction results. The results, in the HF and LF channels, decomposed by the image decomposition module are shown in the two middle columns. The images at the last two columns are reconstructed from the images in the HF and LF channels, respectively. Moreover, the results in the odd rows are from HR images, and the results in the even rows are from LR images.

Image classification performance of the image decomposition module. We aim to verify that the images decomposed by ϕ can be effectively used for feature extraction and classification. We train three classifiers with the decomposed images in the CH^H , CH^L , and $CH^H \cup CH^L$ domains, respectively, and present the classification performance. The classification accuracies of the three classifiers with various resolutions are depicted in columns 2, 4, and 6 in Table IV. We can observe that the performance of the network using either the images in CH^H or CH^L is lower than that using the images in both CH^H and CH^L . This is not surprising because the images in the two channels lose some image details after the image decomposition. Therefore, the images in either CH^H or CH^L are all useful and can provide auxiliary information for classification.

It is worth to note that the results (column 6 in Table IV) using the images in CH^H and CH^L are still worse than the DRL-GAN results shown in Table I. This again proves the superiority of DRL-GAN.

2) *Effectiveness of LF and HF Generator G^L and G^H : Visualization of the channels reconstructed by G* . To verify the ability of the generator to produce high-qualified channels for classification, we show some of the generated channels in Fig. 8. Rows 1 and 3 show the images decomposed from HR images to the channels containing HF and LF information, respectively.

TABLE IV

COMPARISON OF THE CLASSIFICATION ACCURACIES (%) ON THE WIDER-SHIP USING IMAGES IN THE CH^H , $G^H(CH^H)$, CH^L , $G^L(CH^L)$, $CH^H \cup CH^L$, AND $G^H(CH^H) \cup G^L(CH^L)$ DOMAINS, RESPECTIVELY

Resolution	CH^H	$G^H(CH^H)$	CH^L	$G^L(CH^L)$	$CH^H \cup CH^L$	$G^H(CH^H) \cup G^L(CH^L)$
0.60m	87.66	N/A	85.67	N/A	90	N/A
1.19m	87.33	87.67	83.33	84	88.33	90.00
2.39m	81.67	84.67	78.67	81.33	84	89.67
4.78m	68.67	75.33	67.67	73.67	69.67	76.33

Resolution refers to metres per pixel.

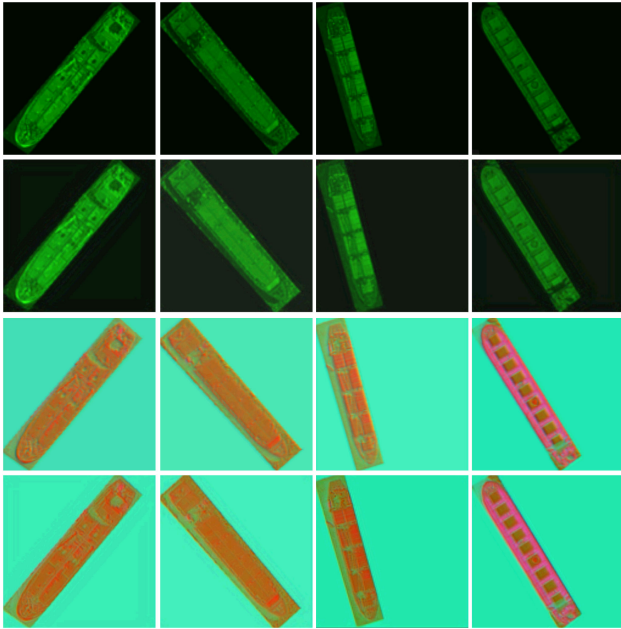


Fig. 8. Visualization of decomposed HR images and the corresponding LR images reconstructed by G .

The images enhanced by generators G^H and G^L from the downsampled (i.e., LR) images of those in rows 1 and 3 are shown in rows 2 and 4. We can observe that the images in rows 1 and 3 look very similar to the images in rows 2 and 4, respectively. Therefore, the generators are successfully learned to transfer the LR images to ones similar to those of HR. This again validates the effectiveness of G in DRL -GAN.

Image classification performance of G^L and G^H . To verify the advantages of employing the two generators on the channels decomposed by ϕ , we train another two classifiers with the decomposed images in the $G^H(CH^H)$ and $G^L(CH^L)$ domains, respectively, and compare them with the proposed DRL -GAN that uses the images in $G^H(CH^H) \cup G^L(CH^L)$. The results are also shown in Table IV. We can observe that both G^H and G^L enhance the performance of classification with higher accuracy results as shown in columns 3 and 5 in comparison with the results in columns 2 and 4 with using the GAN generators. Finally, the classification accuracies of DRL -GAN with the two generators G^H and G^L are depicted in column 7 of Table IV. Column 7 shows that DRL -GAN further improve the performance in classifying LR images by reconstructing the images using the GAN generators, compared the results in column 6.

V. CONCLUSION

In this article, we have formulated an LF–HF inconsistency problem, which significantly degrades the performance of CNN-based methods in the applications based on LR images. We have proposed DRL -GAN to address the problem. DRL -GAN generates enhanced image representations optimized from LR image representations for LR object recognition by simultaneously recovering the missing information of the LF and HF components. We have also demonstrated that our proposed dual-channel AE can effectively decompose input images into LF and HF components. Extensive experiments on three datasets have shown the effectiveness of our proposed solution in terms of quantitative and visual results.

Our method can be used to deal with the RoIs produced by any small object detector for challenging applications, especially for UAV and even satellite remote sensing. When using the proposed DRL -GAN for identifying the types of the LR targets, the low-quality RoI representation could be improved to the high-quality one for the final object recognition.

REFERENCES

- [1] D. Du *et al.*, “The unmanned aerial vehicle benchmark: Object detection and tracking,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 375–391.
- [2] G.-S. Xia *et al.*, “DOTA: A large-scale dataset for object detection in aerial images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.
- [3] M. S. Ryoo, K. Kim, and H. J. Yang, “Extreme low resolution activity recognition with multi-siamese embedding learning,” in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7315–7322.
- [4] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [5] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.
- [6] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
- [7] J. Tschannerl *et al.*, “MIMR-DGSA: Unsupervised hyperspectral band selection based on information theory and a modified discrete gravitational search algorithm,” *Inf. Fusion*, vol. 51, pp. 189–200, 2019.
- [8] J. Ren, “ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging,” *Knowl.-Based Syst.*, vol. 26, pp. 144–153, 2012.
- [9] X. Wang *et al.*, “ESRGAN: Enhanced super-resolution generative adversarial networks,” in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 63–79.
- [10] W. Tan, B. Yan, and B. Bare, “Feature super-resolution: Make machine see more clearly,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3994–4002.
- [11] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 2472–2481.
- [12] X. Hu, H. Mu, X. Zhang, Z. Wang, T. Tan, and J. Sun, “Meta-SR: A magnification-arbitrary network for super-resolution,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1575–1584.

[13] Z.-Q. J. Xu, Y. Zhang, and Y. Xiao, "Training behavior of deep neural network in frequency domain," in *Proc. Int. Conf. Neural Inf. Process.*, 2019, pp. 264–274.

[14] N. Rahaman *et al.*, "On the spectral bias of neural networks," in *Proc. 36th Int. Conf. Mach. Learn.*, 2018, pp. 5301–5310.

[15] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "SOD-MTGAN: Small object detection via multi-task generative adversarial network," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2018, pp. 206–221.

[16] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6967–6974.

[17] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 105–114.

[18] Z. Wang, M. Ye, F. Yang, X. Bai, and S. Satoh, "Cascaded SR-GAN for scale-adaptive low resolution person re-identification," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 3891–3897.

[19] P. Yi, Z. Wang, K. Jiang, Z. Shao, and J. Ma, "Multi-temporal ultra dense memory network for video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2503–2516, Aug. 2020.

[20] Z. Shao, L. Wang, Z. Wang, and J. Deng, "Remote sensing image super-resolution using sparse representation and coupled sparse autoencoder," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 12, no. 8, pp. 2663–2674, Aug. 2019.

[21] S. Mao, S. Zhang, and M. Yang, "Resolution-invariant person re-identification," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 883–889.

[22] Y.-C. Chen, Y.-J. Li, X. Du, and Y.-C. F. Wang, "Learning resolution-invariant deep representations for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8215–8222.

[23] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8090–8099.

[24] N. Padfield, J. Zabalza, H. Zhao, V. Masero, and J. Ren, "EEG-based brain-computer interfaces using motor-imagery: Techniques and challenges," *Sensors*, vol. 19, no. 6, 2019, Art. no. 1423.

[25] J. Zabalza *et al.*, "Novel folded-PCA for improved feature extraction and data reduction with hyperspectral imaging and SAR in remote sensing," *ISPRS J. Photogrammetry Remote Sens.*, vol. 93, pp. 112–122, 2014.

[26] Y. Huang, J. Xu, Q. Wu, Z. Zheng, Z. Zhang, and J. Zhang, "Multi-pseudo regularized label for generated data in person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 3, pp. 1391–1403, Mar. 2019.

[27] J. Zabalza *et al.*, "Novel two-dimensional singular spectrum analysis for effective feature extraction and data classification in hyperspectral imaging," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4418–4433, Aug. 2015.

[28] Y. Yan *et al.*, "Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognit.*, vol. 79, pp. 65–78, 2018.

[29] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4792–4800.

[30] 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 Recognit.*, vol. 66, pp. 412–424, 2017.

[31] Z. Lu, X. Jiang, and A. Kot, "Deep coupled ResNet for low-resolution face recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 4, pp. 526–530, Apr. 2018.

[32] 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 *Proc. Int. Conf. Comput. Vis.*, 2019, pp. 9725–9734.

[33] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 109–117.

[34] Z.-Q. J. Xu, Y. Zhang, T. Luo, Y. Xiao, and Z. Ma, "Frequency principle: Fourier analysis sheds light on deep neural networks," 2019, *arXiv:1901.06523*.

[35] Z. J. Xu, "Understanding training and generalization in deep learning by Fourier analysis," 2018, *arXiv:1808.04295*.

[36] Y. Cao, Z. Fang, Y. Wu, D.-X. Zhou, and Q. Gu, "Towards understanding the spectral bias of deep learning," 2019, *arXiv:1912.01198*.

[37] H. Chen, M. Lin, X. Sun, Q. Qi, H. Li, and R. Jin, "MuffNet: Multi-layer feature federation for mobile deep learning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2019, pp. 2943–2952.

[38] S. Biland, V. C. Azevedo, B. Kim, and B. Solenthaler, "Frequency-aware reconstruction of fluid simulations with generative networks," 2019, *arXiv:1912.08776*.

[39] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon, and Z. Wang, "Delving into robust object detection from unmanned aerial vehicles: A deep nuisance disentanglement approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 1201–1210.

[40] L. Cao, R. Ji, C. Wang, and J. Li, "Towards domain adaptive vehicle detection in satellite image by supervised super-resolution transfer," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2016, pp. 1138–1144.

[41] D. Liu, B. Cheng, Z. Wang, H. Zhang, and T. S. Huang, "Enhance visual recognition under adverse conditions via deep networks," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4401–4412, Sep. 2019.

[42] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Proc. Int. Conf. Learn. Representations*, 2016.

[43] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Conf. Neural Inf. Proc. Syst.*, 2014, pp. 2672–2680.

[44] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[45] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[46] T. Chen, L. Lin, W. Zuo, X. Luo, and L. Zhang, "Learning a wavelet-like auto-encoder to accelerate deep neural networks," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6722–6729.

[47] Y. Xi *et al.*, "See clearly in the distance: Representation learning GAN for low resolution object recognition," *IEEE Access*, vol. 8, pp. 53 203–53 214, 2020.

[48] 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 Geosci. Remote Sens. Lett.*, vol. 13, no. 8, pp. 1074–1078, Aug. 2016.

[49] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep. TR-2009, Univ. Toronto, Toronto, ON, Canada, 2009.

[50] Y. Xi *et al.*, "Beyond context: Exploring semantic similarity for small object detection in crowded scenes," *Pattern Recognit. Lett.*, vol. 137, pp. 53–60, 2020.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

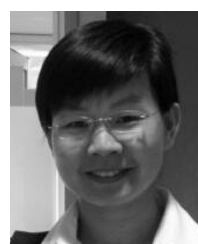
[52] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[53] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[54] K. Jiang, Z. Wang, P. Yi, G. Wang, T. Lu, and J. Jiang, "Edge-enhanced GAN for remote sensing image superresolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Aug. 2019.



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



Wenjing Jia (Member, IEEE) received the Ph.D. degree in computing science from the University of Technology Sydney (UTS), Ultimo, NSW, Australia, 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 more than 100 quality journal articles and conference papers. Her research interests include image/video analysis, computer vision, and pattern recognition.



Jiangbin Zheng received the B.S., M.S., and Ph.D. degrees in computer science from Northwestern Polytechnical University, Xi'an, China, 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, NSW, Australia. Since 2009, he has been a Professor and a Ph.D. Supervisor with the School of Computer Science, Northwestern Polytechnical University. He has authored/coauthored more than 100 peer-reviewed journal/conference papers covering a wide range of topics in image/video analytics, pattern recognition, machine learning, and big data analytics. His research interests include intelligent information processing, visual computing, multimedia signal processing, big data, and soft engineering.



Jinchang Ren (Senior Member, IEEE) received the B.E. degree in computer software, the M.Eng. degree in image processing, and the D.Eng. degree in computer vision from Northwestern Polytechnical University (NWPU), Xi'an, China, and the Ph.D. degree in electronic imaging media communication from the University of Bradford, Bradford, U.K.

He is currently a Professor of Computing with the Robert Gordon University, Aberdeen, U.K. He has authored/coauthored more than 300 peer-reviewed journal/conferences papers. His research interests include hyperspectral imaging, image processing, computer vision, big data analytics, and machine learning.

Dr. Ren is an Associate Editor for several international journals including the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING and *Journal of the Franklin Institute*.



Xiaochen Fan received the B.S. degree in computer science from the Beijing Institute of Technology, Beijing, China, in 2013. He is currently working toward the Ph.D. degree in computer science with the School of Electrical and Data Engineering, University of Technology Sydney, Ultimo, NSW, Australia.

His research interests include mobile/pervasive computing, pattern recognition, deep learning, and IoT.



Xiangjian He (Senior Member, IEEE) received the Ph.D. degree in computer science from the University of Technology Sydney (UTS), Ultimo, NSW, 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.



Yefan Xie received the M.S. degree in computer science and technology from Northwestern Polytechnical University, Xi'an, China, in 2020. He is currently working toward the Ph.D. degree in computer science.

His research interests include machine learning, computer vision, and lightweight network.