IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 1, 2019, accepted February 26, 2019, date of publication March 6, 2019, date of current version March 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2903422

Object Detection in Aerial Images Using Feature Fusion Deep Networks

HAO LONG[1,2], YINUNG CHUNG[2], ZHENBAO LIU[3], (Senior Member, IEEE), AND SHUHUI BU[3], (Member, IEEE)

[1]College of Robotics, Beijing Union University, Beijing 100101, China
[2]Department of Electrical Engineering, National Changhua University of Education, Changhua 50007, Taiwan
[3]School of Aeronautics, Northwestern Polytechnical University, Xi'an 710072, China

Corresponding authors: Yinung Chung (ynchung@cc.ncue.edu.tw) and Zhenbao Liu (liuzhenbao@nwpu.edu.cn)

This work was supported in part by the Natural Science Foundation of China under Grant 61672430, in part by the Shaanxi Key Research and Development Program under Grant S2019-YF-ZDCXL-ZDLGY-0227, in part by the Aeronautical Science Fund under Grant BK1829-02-3009, in part by the NWPU Basic Research Fund under Grant 3102018jcc001, in part by the Science and Technique Program of Beijing Municipal Education Commission under Grant KM201711417009, and in part by the Ministry of Science and Technology under Grant MOST 105-2221-E-018-023.

ABSTRACT Object detection acts as an essential part in a wide range of measurement systems in traffic management, urban planning, defense, agriculture, and so on. Convolutional Neural Networks-based researches reach a great improvement on detection tasks in natural scene images enjoying from the strong ability of feature representations. However, because of the high density, the small size of objects, and the intricate background, the current methods achieve relatively low precision in aerial images. The intention of this work is to obtain better detection performance in aerial images by designing a novel deep neural network framework called Feature Fusion Deep Networks (FFDN). The novel architecture combines a designed structural learning layer based on a graphical model. As a result, the network not only provides powerful hierarchical representation but also strengthens the spatial relationship between the high-density objects. We demonstrate the great improvement of the proposed FFDN on the UAV123 data set and another novel challenging data set called UAVDT benchmark. The objects which appear with small size, partial occlusion and out of view, as well as in the dark background can be detected accurately.

INDEX TERMS Convolutional neural networks (CNNs), aerial images, feature fusion deep networks (FFDN), object detection.

I. INTRODUCTION

In Unmanned Aerial Vehicles (UAV) images, object detection has attracted significant attention worldwide and has received lots of significant applicable achievements [1]–[3]. However, this task still faces lots of challenges: First, aerial images are taken from top to bottom vertically or obliquely at high altitude, so the background is more cluttered than that in the images taken from the ground. For instance, when detecting vehicles in aerial images, some similar objects such as the roofing equipment and substation box possibly cause false positive detection. Second, objects in aerial images are much smaller and always with a higher density than that in the natural scene images, especially when shooting images at a wider view angle. Third, the lack of large-scale and well-annotated data sets limits the high performance of a trained network.

In previous years, the works that are mostly based on the sliding window search, as well as the shallow-learning-based features [4], [5] have been fully researched. Liu and Mattyus [6] propose a method for detecting vehicles in several categories and different orientation in aerial images. Nowadays, R-CNN based detection methods [7] have brought about the huge success in natural scene images. Although CNNs can learn powerful hierarchical features, they would result in signal down-sampling problem and relatively weak spatial description when they are utilized to object detection task in aerial images [8], [9]. This is because the network performs multiple operations of max-pooling and down-sampling, which are originally constructed for feature abstraction task. The feature dimension is largely reduced. This results in the weak spatial description with

The associate editor coordinating the review of this manuscript and approving it for publication was Naveed Akhtar.

drastically reduced spatial resolution. CNN cannot learn the spatial description of the structural relationship effectively. Moreover, due to the high and variable flight altitude and multi-angle rotate shooting of UAV views, the ground objects with the same semantics are usually with small size, multiple scales and high density. The convolutional mode is effective for large objects, but not so effective for small objects with high density. Thus, the feature extraction performance is not outstanding in aerial images. This inherently limits the spatial description ability of CNNs especially for aerial images. Conditional Random Fields (CRFs) [10], which can clearly infer the spatial dependencies between objects, have been applied for object detection problem for boosting the accuracy of detection [11]. Zhang *et al.* [12] build a CRF model to use the interaction between neighboring regions for overtaking object detection. Li *et al.* [13] fuse the LIDAR and monocular images into the CRF for detecting the road robustly in intricate scenarios.

In this paper, different from above-mentioned methods [12], [13], we design a frame of networks combining Convolutional Restricted Boltzmann Machines (CRBM) with CRF, in which CRF model is one layer of FFDN. By this way, the networks can obtain the structural learning ability. In short, there are three major layers in the proposed FFDN:

1) **Feature learning layer**: we adopt CRBM to learn deep hierarchical features (DHF), which can express pyramid representations of different scales of the input image. Therefore, the DHF can hierarchically express small objects and intricate background information in aerial images.
2) **Structural learning layer**: the CRF is adopted as a layer of the FFDN for expressly generating the spatial relationship between adjacent objects and background, then we generate spatially inferred features (SIF) by encoding region information using spatial relationships. By this way, we can improve the structural representative performance of the FFDN.
3) **Feature fusion layer**: to take both advantages of the DHF and the SIF mentioned above, we fuse DHF and SIF by Deep Sparse Autoencoder (DSA) [14], [15], as a result, the network has got stronger representative features for locating objects in intricate backgrounds.

This paper has three main contributions summarized as follows: First, in this work, the upper layer of the CRBM covers a larger region through the pooling and downsampling, therefore the CRBM can absolutely learn the structural relationship in small regions. Furthermore, the proposed FFDN includes the CRF layer, which means that the structural learning executes explicitly and thus enhance the capability of inference; Second, the SIF can boost the spatial inference by encoding the region information between objects. Thus the features of both the object itself and its region relationship information can be encoded to learn stronger representation. The network achieves significant improvements in object detection among adjacent small objects; Third, a new competitive feature learning algorithm is proposed. To take advantage of various features, we put forward a feature fusion layer based on DSA to fuse the multimodal features, and thus to learn the inherent nonlinear relationships of comprehensive local and global information. We consequently improve the ability to model complicated transformations of the FFDN.

Extensive experiments have been conducted on the UAV123 data set [16] and UAVDT data set [17]. Quantitative comparisons and analysis show that the proposed FFDN obtains promising performance.

II. RELATED WORKS

There are two main factors affecting the performance of the detectors: one is the partial occlusions caused by other things that would significantly increase detection errors, and the other is the illumination condition critical for detection task in aerial images [18]. Many effective research methods have been proposed to find suitable techniques and have achieved great performance. In the following content, we review the traditional detection methods which rely on shallow-learning-based feature extraction, then discuss the representative CNN-based and feature fusion based methods, respectively.

A. TRADITIONAL DETECTION METHODS

The traditional detection methods generally rely on handcrafted feature extraction. For instance, Moon *et al.* [19] focus on four elongated edges of cars in aerial images of parking lots, and they also discuss how much prior information improves detection performance. Hinz *et al.* [20] introduce a 3-D model describing the geometric features and the radiometric features (colors of vehicles and windshields, the intensity of car's shadow). Wang [21] propose a framework to fusion improved shadow features and shape matching of corner features. Moranduzzo [2], by using the scalar invariant feature which identifies a set of key points of cars and a support vector machine classifier, represent a ''one keypoint-one car'' method for car detection. Yamazaki *et al.* [22] suggest the parameters of gray values and sizes for object classification in aerial images and also derive the speed of moving cars by exploiting the shadows. Moranduzzo and Melgani [23] utilize several invariant features to discover the objects, in addition, to calculate the moving speed by analyzing the centroid position movement between two successive frames. Although these methods have made good performances, they still have limited application range and always cause inconsistent on other different tasks because of the trivial partial information of objects generated from images.

B. DEEP-LEARNING BASED DETECTION METHODS

Over the past several years, to overcome the disadvantages of those features, some methods [24]–[26]tend to simulate human vision mechanism. In recent researches, deep learning based methods [27]–[30] have become a hot topic in computer vision and have made great achievements. After the AlexNet [31] won the ILSVRC-2012 competition,

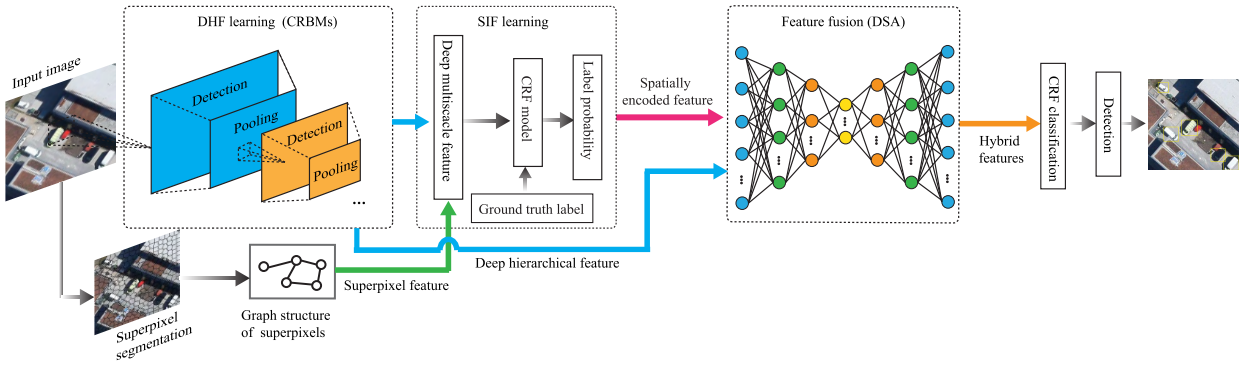

FIGURE 1. Overview of the proposed FFDN. Different categories of features generated from different layers are represented by different colors.

CNNs [32] enhance the feature extraction performance of various computer vision tasks (object detection and tracking [32], [33], image classification [31], [32], and fine-grained categorization [8]). There are typically two groups of methods in CNN-based detection methods. One is two-stage methods, in which the process of object location and object classification is finished in two steps. The representatives of this group are Faster R-CNN [34], Mask R-CNN [35] and so on. The other is one-stage methods such as YOLOv3 [36] and DSSD [37]. The former group outperforms the latter one in terms of accuracy. The latest Mask R-CNN expands the Faster R-CNN to an instance segmentation framework by generating the bounding boxes and predicts masks for them simultaneously in the second stage. Whereas, the latter group of methods, which remove the proposal generation stage, performs much faster than the former. YOLOv3 [36], which has a simpler pipeline and uses the strategy of multi-scale prediction, achieve excellent performance in detection speed. Although these two groups of detection methods achieve satisfactory performance in natural scene images, when applied to the aerial images, there still exist problems. It is very essential for the former group to ensure a reliable recall rate during the proposal generation process, if some objects are missed in the head network, they will not be recovered in the second stage. Focusing only on extracting features from a single feature map layer is not sufficient for object detection in challenging aerial images.

Except for the above methods, there are some other methods that focus more on feature fusion strategy and have received increasing attention in recent researches. In Li *et al.* [38], the authors propose a SingleNet for object detection. They apply the fully convolutional network as the base network to generate feature maps and construct a fusion network to fuse these feature maps in different layers. Finally, they merge these features by the element-wise sum. Guan *et al.* [39] propose multi-scale feature fusion based object detection method, which constructs a region object network and jointly fuses the high-abstracted semantic knowledge to learn a fine resolution feature maps. Jiang *et al.* [40] integrate the semantic segmentation feature layer by layer into the feature pyramid structure and predict the location on fusion feature maps of different layers independently.

Different from above-mentioned researches, in our work the graphical model (CRF) is considered as one layer of the proposed FFDN to enhance the definite inference ability, then we generate structural relationships by encoding the inference results. Therefore, deep learning based feature is combined with the graphical model to simultaneously exploit the advantages of both of them. Furthermore, DSA is used to fuse deep hierarchical features and spatially inferred features, and thus the nonlinear relationships of them would be obtained, which means that we can fully grasp the inherent feature of small objects.

III. FEATURE FUSION DEEP NETWORKS

The feature learning layer constructed by CRBMs is firstly introduced in this section, then we discuss the structural learning layer of getting the strong spatial relationship representation between objects and background. Finally, we briefly introduce the feature fusion layer to generate more powerful representative features. Fig. 1 gives a general view of the proposed networks.

In our framework, structural learning can be regarded as a processing layer and is trained separately. It is known that the CRF model has the advantage of overcoming lack of long-range spatial inference of traditional deep neural networks because it definitely generates a spatial relationship between objects and background. As a result, object detection performance can be improved. Moreover, training those three layers individually can make three separate modules to be trained repeatedly and simultaneously, as a consequence, it makes the training more quickly and then reduces the time consumption.

A. FEATURE LEARNING LAYER

It is well known that powerful representations are crucial for promising performance on computer vision task. Contemporary methods indicate that powerful internal feature representations are hierarchical, and the convolutional operation is invariant to tilt, translation, scaling, and other deformations, which makes object detection more accurate and convenient

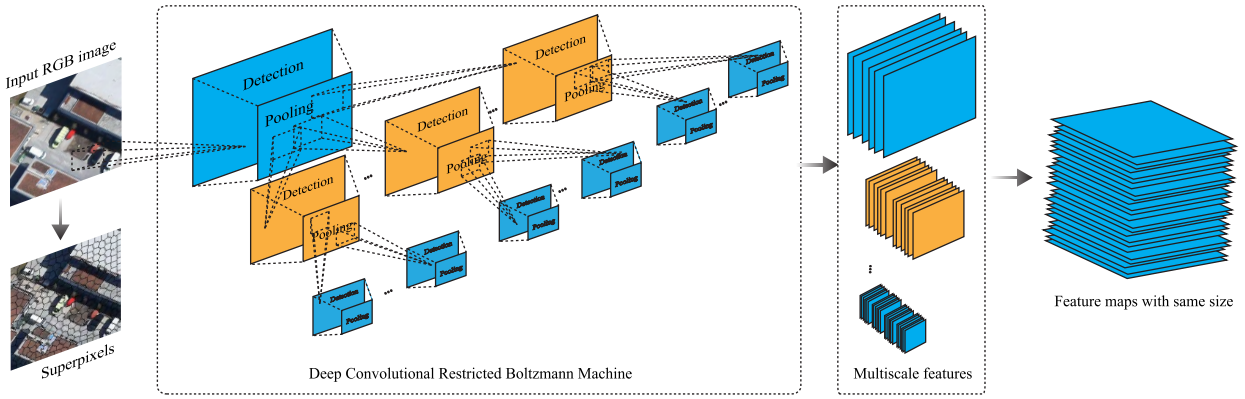

FIGURE 2. The principle of feature learning with CRBMs.

than other shallow feature based methods. Norouzi *et al.* [41] introduce the stacks of CRBM for shift-invariant feature learning and have proved their extracted features are effective for object detection. Here, we adopt multilayer CRBMs to generate the DHF. Fig. 2 shows the principle of feature learning with CRBMs.

1) CONVOLUTIONAL RESTRICTED BOLTZMANN MACHINE

So as to obtain local rotation invariant characterization, Lee *et al.* [42] put forward CRBM and Convolutional Deep Belief Networks (CDBN). The basic CRBM is a two-layer structure (visible layer $\mathbf{V}$ and detection layer $\mathbf{H}$) similar to RBM. Suppose $\mathbf{H}$ includes k ''groups'' of units. Max-pooling layers $\mathbf{P}$ reduce the dimensions of layer $\mathbf{H}$ by using the pooling window B_α(with the width of G pixels). Due to the same meaningful features might appear anywhere of the image, the convolutional kernel $W^k(k \in [1, k])$ are shared among all regions of an image between layer $\mathbf{V}$ and layer $\mathbf{H}$. By stacking multiple CRBMs, we can construct, similar to DBN, a CDBN. Suppose $\mathbf{V}$ is a binary unit of N_v dimension. The detection layer uses K convolution filters, each of which is a matrix of N_w dimensions. The k-th convolutional kernel W^k is used to acquire the K ''groups'' of $\mathbf{H}^k$ in the layer $\mathbf{H}$ by convolving the entire image. The $\mathbf{H}^k$ is a N_h-dimensional unit matrix, in which each unit h_{ij}^k shares the same weight W^k and bias b^k, and i, j represent the vertical or horizontal indexes. The single bias c is shared with all of the units v_{ij} in visible layer $\mathbf{V}$. The joint probability over visible unit $\mathbf{v}$ and detection unit $\mathbf{h}$ is given by

$$p(\mathbf{v}, \mathbf{h}) = \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})), \tag{1}$$

where $Z = \sum_{\mathbf{v}} \sum_{\mathbf{h}} exp(-E(\mathbf{v}, \mathbf{h}))$ is defined as the normalized parameter of the separation function. The energy function of CRBM is defined as

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{k=1}^{K} h^k \cdot (\widetilde{W}^k * v) - \sum_{k=1}^{K} b^k \sum_{ij} h_{ij}^k - c \sum_{ij} \mathbf{v}_{ij}, \tag{2}$$

where $*$ denotes the two-dimensional convolution, $\cdot$ represents element-wise multiplication, as well as the tilde above W^k denotes flipping W^k horizontally and vertically. To learn high-level representations, similar to DBNs, stacking multiple CRBMs can obtain deeper CDBNs. Lee *et al.* [42] further combine the probabilistic max-pooling with CRBM, so the structure of the CDBN is generally based on probabilistic max-pooling. In other words, the units calculate (by probability) the maximum activation in small areas of the detection layer $\mathbf{H}$. The energy function of this probabilistic max-pooling-CRBM is given by

$$E(\mathbf{v}, \mathbf{h}) = -\sum_{K} \sum_{ij} (h_{ij}^k (\widetilde{W}^k * v)_{ij} + b_k h_{ij}^k) - c \sum_{ij} v_{ij}$$
$$s.t. \sum_{(ij) \in B_\alpha} h_{ij}^k \leq 1, \forall k, \alpha, \tag{3}$$

where B_α (with the width of G pixels) is pooling window of detection layer. The stochastic gradient descent (SGD) is performed to optimize the parameters of the CRBM [43]. However, it is unrealistic to calculate the exact gradient accurately, instead, we use the contrast divergence (CD) [44] approximation, which has been confirmed to work well in practice. Similar to [43], we train the CDBN in a greedy way.

2) SUPERPIXELS

In our proposed FFDN, superpixel segmentation is a vital step. We use the algorithm of simple linear iterative clustering (SLIC) [45] to obtain superpixels, which are regarded as elementary units of similar color, textual, and category to eliminate some aberrant pixels. On the other hand, using superpixel can boost the total computational speed significantly because there are much fewer superpixels than pixels in an image. As we all know, superpixels can explicitly keep the boundaries between objects, which can help us get a precise distinguishment of the adjacent objects. It can especially promote the detection performance of small objects [9]. For each superpixel, we calculate the average feature denoted as $S_p \in \mathbb{R}^N$ in the detected region.

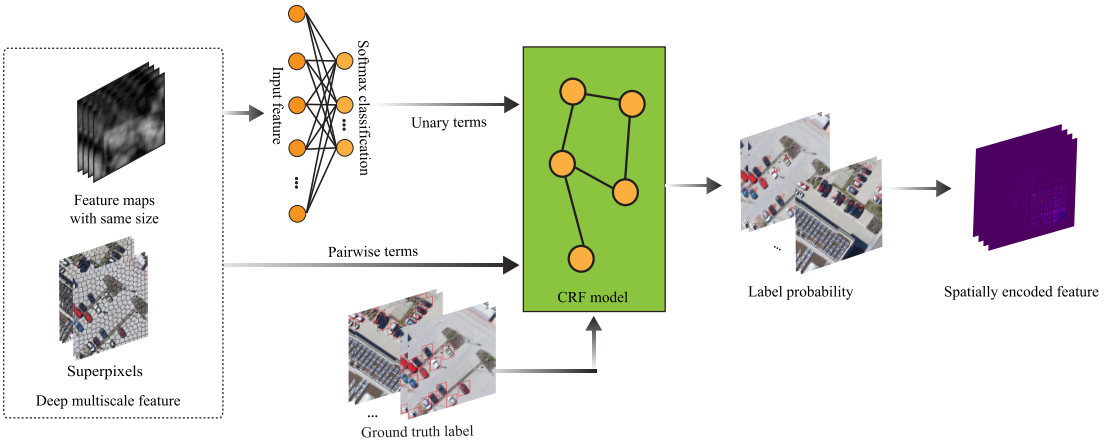

FIGURE 3. The flowchart of structural representation learning.

B. STRUCTURAL LEARNING LAYER

Although CNNs can generate powerful hierarchical features, the hierarchical features are still short of strong spatial relationship representations (without neighborhood information) among objects, thus CNNs is not particularly suitable for accurate small-sized objects detection alone. For these reasons, we use the CRF model based on superpixels to learn the SIF [9]. We illustrate how to learn the spatially inferred feature in Fig. 3.

1) CONDITIONAL RANDOM FIELDS

We generate the superpixels of an input image by performing the SLIC [45] algorithm. For an image, we define a graph model $G = (V, E)$, in which the vertexe $v \in V$ and the edge $e \in E \in \mathbb{R}^{V\times V}$, respectively. Under this definition, each superpixel can be considered as a vertex unit, as well as the edge can be regarded as the connection among the neighboring unit pairs. Specifically, the symbol e_{ij} represents the edge consisting vertex v_i and v_j, while the observation of units are expressed as $\mathbf{x}$, whose corresponding states are $\mathbf{y} =< \mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n >$. The conditional distribution of every vertex and edge is decomposed into the potentials of the units $\phi_N(\mathbf{x}_i, \mathbf{y}_i)$ (unary) and edges $\psi_E(\mathbf{x}_{ije}, \mathbf{y}_i, \mathbf{y}_j)$ (pairwise). Considering the possibility of having a weight $\mathbf{w}$ in the training data, the conditional probability distribution is given by:

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}) = \frac{1}{Z(\mathbf{x}, \mathbf{w})} \prod_{i\in V} \phi_N(\mathbf{x}_i, \mathbf{y}_i) \prod_{e_{ij}\in E} \psi_E(\mathbf{x}_{ije}, \mathbf{y}_i, \mathbf{y}_j), \tag{4}$$

where $Z(\mathbf{x}, \mathbf{w})$ denotes a partition function with unary or pairwise potentials on the constructed graph model. Defining the feature functions of the unit $\phi_N(\mathbf{x}_i, \mathbf{y}_i)$ and $\psi_E(\mathbf{x}_{ije}, \mathbf{y}_i, \mathbf{y}_j)$ as $\mathbf{f}_N$ and $\mathbf{f}_E$, the potentials is expressed as the log-linear combination of $\mathbf{f}_N$ and $\mathbf{f}_E$. we reform the training process as:

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 - \sum_{n=1}^{M} \Big(\sum_{i\in V} \mathbf{w}_N^{\mathrm{T}} \mathbf{f}_N(\mathbf{x}_i^n, \mathbf{y}_i^n) + \sum_{e_{ij}\in E} \mathbf{w}_E^{\mathrm{T}} \mathbf{f}_E(\mathbf{x}_{ije}^n, \mathbf{y}_i^n, \mathbf{y}_j^n) \Big) + \sum_{n=1}^{M} \log Z(\mathbf{x}^n, \mathbf{w}), \tag{5}$$

where λ denotes the non-negative L2-regularizer parameter, $\mathbf{w} = [\mathbf{w}_N, \mathbf{w}_E]$ is the weight of unary and pairwise terms, and $(\mathbf{x}_i^n, \mathbf{y}_i^n)$ is defined as the training examples of a sample of the graphical model. When $p(\mathbf{y}|\mathbf{x}, \mathbf{w}^*)$ reaches the maximum, the conditional probability distribution over the class variable and the most likely assignment of labels $\mathbf{y}$ can be acquired simultaneously by solving (5).

2) SPATIALLY INFERRED FEATURE

It can be noted that in the recent research literature, CRF is usually considered to be the last procedure for refining the classified labels in different vision tasks. In our work, we improve the structure by adjusting this procedure. It is well known that inference label probabilities for each superpixel generated by CRBM learning features lack the powerful capability to learn spatial relationships. Even though the graphical model can partially make up those disadvantages, to further improve performance, SIF is put forward to indicate both the feature of superpixel and the spatial relationships. We define the connection graph as $G_\mu = (V_\mu, E_\mu)$, which is generated by the superpixel μ and its local region, thus the SIF $\odot(\mu)$ can be expressed as

$$\odot(\mu) = \lambda \sum_{i\in V_\mu} \sum_{j\in V_\mu} \theta_i \theta_j^T \exp\left(-k_d \frac{d(v_i, v_j)}{\sigma_d}\right). \tag{6}$$

In (6), $\odot$ is $n \times n$ matrix, which indicates the frequency of occurrence of the nearby probability of vertices i and j. $d(v_i, v_j)$ denotes the distance of superpixel i and j, while k_d, σ_d, and λ denote the distance decay rate, the maximal distance of the vertices in the graph G_μ, and the normalized parameter, respectively [9].

C. FEATURE FUSION LAYER

In this section, we briefly discuss DSA used in our proposed framework, which is a neural network with multilayer sparse autoencoder (SAE). DSA can discriminatively learn hierarchical features by finding out similarities between training samples. We show how to fuse the various features in Fig. 4. When the networks accomplish the courses of

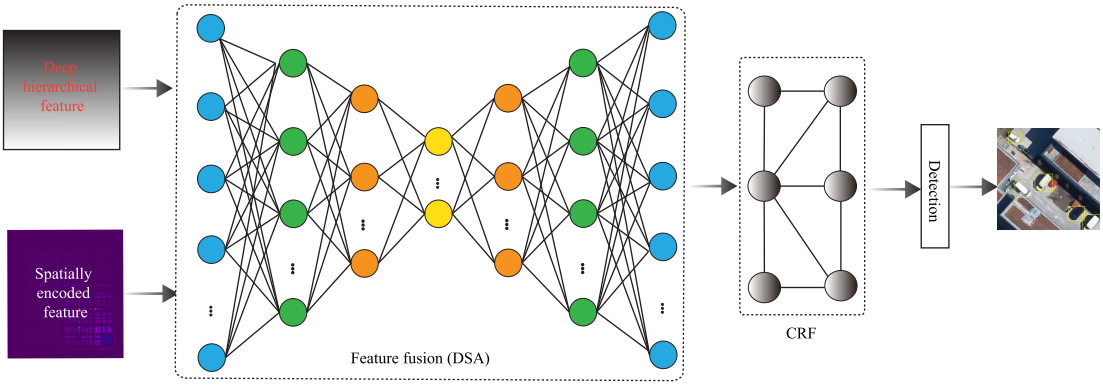

FIGURE 4. The illustration of hybrid feature learning.

feature learning and structural learning, we get two types of descriptors: DHF S_p and SIF $\odot$, which are concatenated into $[S_p, \odot] \in \mathbb{R}^{N+n\times n}$ in our method, and then we use DSA [14] to fuse the connected features, at the same time, the comprehensive nonlinear relationships between different dimensional features are also explored.

DSA application results in recent years [14], [15] have demonstrated its capability to learn multi-layer nonlinear features with less labeled data, which are beneficial for object detection. The features are generated layer by layer with greedy learning strategy [15] through contrastive divergence (CD) algorithm [44]. When the unsupervised pre-training phase is completed, supervised backpropagation with less labeled data is carried out to fine-tune the network for optimal parameters, as a result, the DSA can output the highly representative feature that encodes its input data.

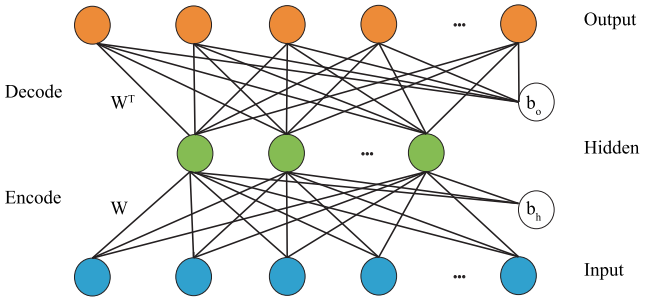

FIGURE 5. The architecture of Sparse Auto Encoder.

Suppose each training sample of SAE as $\mathbf{x} = (x_1, x_2, ..., x_N)^T$ and there are N_h hidden units in layer l, the learned features of $\mathbf{x}$ from the hidden units in layer l is expressed as $\mathbf{h}^l = (h_1^l, h_2^l, ..., h_{N_h}^l)$. SAE is a symmetric network including the encoding and decoding stages, the architecture of SAE is shown as Fig.5. In the encoding stage, we define the linear mapping and nonlinear *sigmoid* activation function as

$$\mathbf{h}^l = \text{sigmoid}(\mathbf{W}\mathbf{x} + \mathbf{b}_h), \tag{7}$$

where $\mathbf{h}^l$ denotes the relevant representation changed by the encoder from the input $\mathbf{x}$. $\text{sigmoid}(z) = (1 + \exp(-z))^{-1}$, $\mathbf{W} \in R^{N\times N_h}$, and bias $\mathbf{b}_h \in R^{N_h\times 1}$. At the same time, an approximation $\tilde{\mathbf{x}}$ can be written as (8).

$$\tilde{\mathbf{x}} = \text{sigmoid}(\mathbf{W}^T\mathbf{h}^l + \mathbf{b}_o), \mathbf{b}_o \in R^{N\times 1}. \tag{8}$$

In (8), $\mathbf{W}$ and $\mathbf{b}_o$ is the weight and the bias. We define (9) to minimize the error between $\mathbf{x}$ and $\tilde{\mathbf{x}}$.

$$O = \frac{1}{N_s}\sum_{i=1}^{N_s}\left\|\tilde{x}^i - x^i\right\|^2 + \beta\left\|\mathbf{W}\right\|_2^2 + \alpha\sum_{j=1}^{N_h} KL(\rho||\tilde{\rho}_j), \tag{9}$$

where $\tilde{\rho}_j$ denotes the average activation of the *j-th* hidden unit, while ρ is the desired activation, which can be set by users. N_s, α, and β denote the number of training samples, the sparse penalty, and weight penalty term, respectively. $KL(\rho||\tilde{\rho}_j)$ is the sparse term expressing the Kullback-Leibler (KL) divergence between the actual average and the desired activation of the hidden unit.

D. TRAINING PROCEDURE

The stacked CRBMs in feature learning layer is trained in a greedy layerwise [41]. The 1-*th* convolutional feature extraction layer is followed by a pooling layer aggregating features over local regions of images. We define the pooling layer as the deterministic max-pooling (DM) layer, which can learn features invariant to slight distortions and shifts. Subsequently, we froze the parameters of the lower layer, and use the conditional probability of N_h to generate features for training the 3-*th* layer. Again, another DM layer is stacked above the feature detectors. In our method, the procedure stops after the 6-*th* layer. In the structural learning layer, the graph-cut algorithm [46], [47] is adopted to obtain the optimal weight with the CRBM features, so we can optimize the CRF energy function. For feature fusion, deep sparse autoencoder is broken down into lots of SAE trained through CD algorithm [44]. After the pre-training procedure, the backpropagation is performed to fine-tune the parameters.

E. BOUNDING BOX PREDICTION

Through the steps introduced in previous sections, we have performed high-performance pixel-level object segmentation, and then we predict candidate bounding boxes of different aspect ratios based on the height and width distance of the semantic segmentation location boundaries. To ensure that the object can be assigned to at least one candidate predicted box, we adopt the following strategy: First, we match the ground truth with the candidates that have got maximum Jaccard overlap. Then, we match these candidates with the ground truths with Jaccard overlap higher than 0.5. We assign positive labels to the boxes in which the Jaccard overlap is over 0.7, negative labels to those below 0.3, and unrecognizable or no labels to boxes with overlap between 0.7 and 0.3.

IV. EXPERIMENTS

In this section, we demonstrate the great improvement of the proposed FFDN on the UAV123 data set [16] and another novel challenging data set called UAVDT benchmark [17]. The former dataset has various target categories while the latter one is constructed in an unconstrained complex scene. We show a comprehensive analysis of the experimental results. The experiments are conducted on a computer with Intel Core i9-7900 3.3-GHz CPU, a NVIDIA GTX-1080Ti GPU.

A. DATA SETS

The UAV123 data set [16] is a recent data set constructed in 2016, which comprises 123 videos recorded by UAV cameras. We choose 33 challenging videos to generate 48,770 frames which cover all kinds of scenarios of the data set. We generate 13,871 frames from the videos and manually produce the ground truth. The main objects varieties considered for the experiments mainly focus on bikes, boats, buildings, people, cars, and so on.

To evaluate the effectiveness of the FFDN in more unconstrained complex scenes, another data set we use for experimental comparative analysis is the UAVDT benchmark [17], which is a new data set constructed in 2018 and has more complex scenarios and higher challenges. The UAVDT data set is captured in over 6 different urban areas and defines 6 attributes (i.e., weather condition, vehicle occlusion, flying altitude, out of view, camera view, and vehicle category) [17]. On consideration with the low resolution, the authors declared the ''Ignored'' regions where cover too small vehicles. ''Ignored'' regions are labeled as pink regions as shown in Fig.8. There are over 2,700 vehicles annotated in this data set. We choose 50 videos sequences with all above-mentioned challenges to generate 40,735 frames, which cover all kinds of scenarios and attributes of the data set.

B. IMPLEMENTATION DETAILS

In both feature learning layer and feature fusion layer, when the unsupervised pre-training phase is completed, supervised backpropagation with labeled data is carried out to fine-tune the network. For UAV123 data set, we randomly divide the labeled data into the training set and the test set with a ratio of 1:1. We choose 30 videos sequences for training, while 20 sequences for testing on UAVDT data set. They share similar scenes and attributes but have different shooting location, which would help to avoid overfitting to some extent. For the stacked CRBMs, the 2-*th*, 4-*th*, and 6-*th* layers of this hierarchy are DM layers that only have the parameters of the subsampling window size. The 1-*th*, 3-*th*, and 5-*th* layers are the convolutional layers adjusted by CD learning of individual CRBMs. During the CD learning procedure, we update the batch gradient using the additional momentum of the previous step gradient [41]. Since some high and low learning rate suppress some of CRBM's feature maps and always makes them inactive, in fact, some meaningful features are dismissed. Several learning rates are tested to select the most proper one that can activate most of the features. On UAV123 data set, there are 15 filters of 7×7 pixels at the 1-*th* layer, and both 30 filters of 5×5 pixels at the 3-*th* and 5-*th* layers. The DM layer is with 4×4 and both with 2×2 subsampling windows at the 2-*th*, 4-*th*, and 6-*th* layers, respectively. The learning rate is set to be 0.01; On UAVDT data set, we learned 20 filters of 5×5 pixels at the 1-*th* layer, and both 30 filters of 3×3 pixels at the 3-*th* and 5-*th* layers. The sub-sampling windows in DM layer remain constant. The number of hidden units in three layers is 500 on both data sets in the experiment, and 0.05 as the learning rate on UAVDT data set. The RBMs are tuned with 3000 epochs of pre-training and 5000 epochs of fine-tuning. The initial biases are set to be 0.

In the structural learning layer, CRF is utilized as a processing layer, because it is trained with DHF without back propagation optimization. When generating superpixel some experiments are conducted to select a proper region size to ensure the good experimental performance and high computation efficiency simultaneously. We choose the region size 15 of each superpixel. We use 0.2 as the non-negative L2-regularizer parameter λ and 0.1 as the distance rate k_d. The CRF computes the spatial relationships between superpixels, thus the unreasonable and incorrect labels are efficiently rectified.

In the stage of feature fusion, we stacked three SAEs for constructing DSA. There are 1100, 800 and 500 hidden units in each hidden layer. We set the sparse penalty term α to 2, 0.1, and 0.05, respectively. The weight penalty term β is set to 0.001, the activation ρ 0.05, and learning rate 0.1. We use 1000 as the batch size and 2000 as the epoch. These parameters remain unchanged in all experiments for training convergence and avoiding overfitting.

C. EVALUATION METRICS

We use four metrics (i.e., precision (P), recall (R), F1-score (F1) and mean intersection over union (Mean IoU))

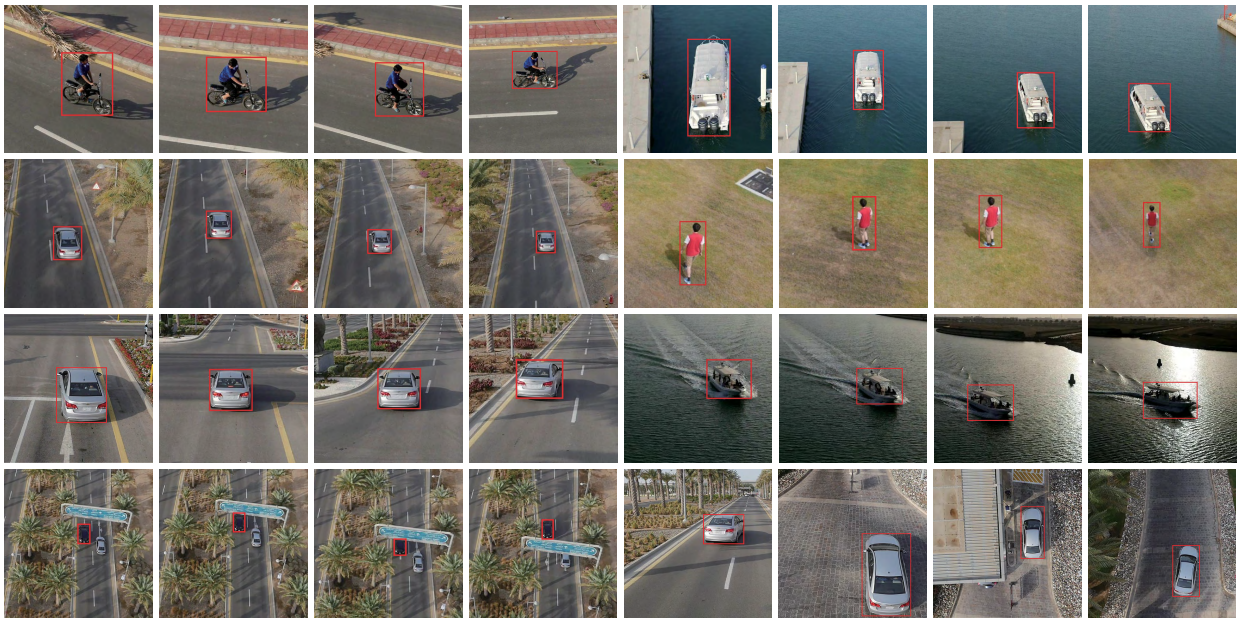

FIGURE 6. Some detection results in several scenarios of the UAV123 data set include Bike1, Boat1, Car1, Person5, Car5, Boat4, Car4, and Car6.

expressed in (11) - (13) to compare the performance quantitatively.

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{10}$$

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{11}$$

$$\mathrm{F1} = \frac{2 \times \mathrm{R} \times \mathrm{P}}{\mathrm{R} + \mathrm{P}}, \tag{12}$$

where TP (true positive) expresses the number of positive prediction which should be positive, FP (false positive) denotes the number of positive prediction which should be negative, and FN (false negative) indicates the number of negative prediction which should be positive.

$$\mathrm{IoU} = \frac{\mathrm{target} \cap \mathrm{prediction}}{\mathrm{target} \cup \mathrm{prediction}} \tag{13}$$

The IoU metrics is virtually a way to quantify the accuracy of the predicted bounding box. As shown in (13), it measures the percentage of overlap between the intersection and the union of the target and prediction by the number of pixels. Mean IoU indicates the average value of all the categories of IoU.

D. EXPERIMENTAL RESULTS

The quantitative comparison, some images of detection results, as well as comprehensive analysis of the proposed FFDN are illustrated on two latest challenging data sets. Finally, we summarize the reasons behind performance improvement and some failure detection samples.

1) UAV123 DATA SET

In order to illustrate the great performance, Table 1 reports the comparison results of five recent methods. Accurate Vehicle Proposal Network(AVPN) [48] integrates heretical feature maps for detecting small-sized objects. Hyper Region Proposal Network with cascade classifier (HRPN with CC) [49] is used to improve the recall rate by adopting a technique similar to [48]. Then, the authors use the cascade classifier to replace the one after region proposal network to reduce the false alarm. It is obvious that Faster R-CNN, AVPN, and HRPN with CC perform better than ACF, Especially, our method FFDN achieves the highest recall of 88.31%, precision of 89.82%, as well as F1-Score of 0.89 atop the leaderboard. Compared to the second in the leaderboard, our method achieves comparatively improves the recall by 10.01%, F1-score by 0.06, respectively.

TABLE 1. Comparison of performance with different methods on UAV123 data set.

Method	R	P	F1
ACF 2015 [6]	50.36%	41.58%	0.45
Faster R-CNN 2017 [34]	67.75%	88.95%	0.77
AVPN 2017 [48]	75.02%	86.83%	0.81
HRPN with CC 2017 [49]	78.30%	89.30%	0.83
FFDN	88.31%	89.82%	0.89

From Fig.6, we can find that even small objects that are partially occluded and the objects in the dark background

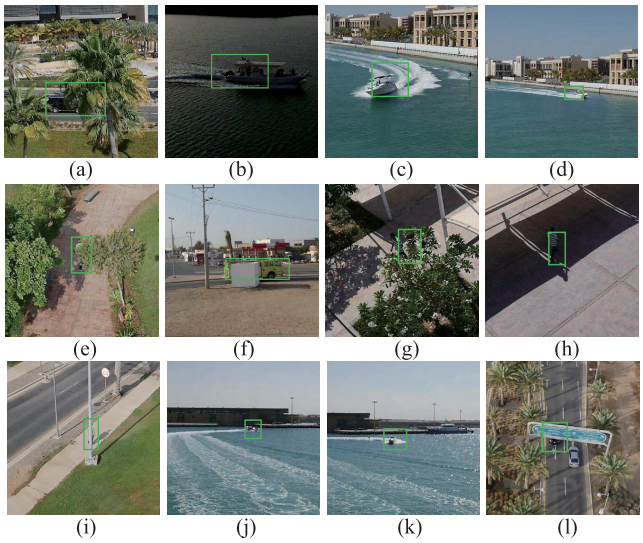

FIGURE 7. Some bad detection results of the large occlusion and small size. Green boxes denote missing and inaccurate detection.

can be located accurately. Except for the good results illustrated above, several bad detection examples are shown in Fig.7, in which the green boxes denote missing or inaccurate detection. The serious problem is that some small-sized objects which have the color similar to the intricate background or objects with the large occlusion are difficult to recognize and locate accurately. As shown in Fig.6, despite the objects appear in small scale or with a small and medium occlusion, the proposed networks has accurately detected the objects. These results indicate that the FFDN has promising detection ability in the UAV images, but it still has some unsatisfactory performance. In detail, bad detection results mostly come from the objects which are occluded heavily (see Fig. 7a, e, f, g, i, and l). The missing detection occurs when the very small-sized objects have the color similar to the intricate background (see Fig. 7d, j, and k), the detector would consider objects as parts of the backgrounds. This may be because the FFDN loses some details of objects in Fig. 7 when obtaining their superpixels.

2) UAVDT DATA SET

We compare our work with the other three methods on UAVDT data set. They are Mask R-CNN [35], YOLOv3 [36], as well as another feature fusion based method called SingleNet [38]. SingleNet applies the fully convolutional network as the base network to generate feature map and construct a fusion network to fuse these feature maps in each layer. Finally, it merges these features by element-wise sum. Table 2 reports the comparison results on UAVDT data set. Analyzing the detection performance reported in Table 2, we can conclude that all the comparative methods achieve unfavorable performance on the UAVDT data set. This may be because the UAVDT data set pays more attention to the UAV based unconstrained real scenes [17]. There are numerous small targets with high density, and the background

TABLE 2. Comparison of performance on UAVDT data set.

Method	R	P	F1	Mean IoU
Mask R-CNN 2017 [35]	33.34%	68.74%	0.45	0.68
YOLOv3 2018 [36]	31.26%	65.17%	0.42	0.62
SingleNet 2017 [38]	33.49%	52.33 %	0.41	0.67
FFDN	35.44%	64.91%	0.46	0.71

becomes more cluttered in UAVDT data set. Moreover, there are many challenging weathers such as fog and night in it. These factors bring new challenges to the detection task in aerial images.

SingleNet [38] is inferior to our method, since the SingleNet only uses the fully convolutional network as a base network to generate feature map and fuse semantic information from each layer. The CNN has the shortcoming of weak spatial description, so the reverse fusion cannot capture the spatial relationship between objects effectively, which is very crucial for detecting small objects with high density in the cluttered background. Mask R-CNN obtains the highest precision which benefits from the region proposal process and segmentation mask prediction for each instance. Our method achieves the highest recall rate, F1-score and mean IoU of 35.44 percent, 0.46, and 0.71, respectively. Besides, our method makes an improvement of 1.95% recall rate over SingleNet. The experimental results validate the competitiveness of our method in unconstrained real scenes. We show detection results on UAVDT data set in Fig. 8. According to different weather conditions and attributes illustrated in the black box at the top right of the images, our method predicts the exact boxes that fit the different categories of vehicles. The bottom-left image shows a false negative case.

3) RESULTS ANALYSIS

Overall, the improvements benefit from two aspects: (1) we embed the CRF based structural learning in the framework to capture the spatial relationship features, and simultaneously to remedy the boundaries between objects at the pixel level. The spatially encoded features, including powerful spatial constraints between objects, can boost the performance of object locating, specifically for small objects with intricate backgrounds. (2) DSA is used to further calculate the nonlinear relationships between various low-level features. By fusing the spatial and structural features, the FFDN abstracts more representative features for detection, especially for differentiate ambiguous ground objects in large and aerial images. On the other hand, The inaccurate or missing detection reveals that the FFDN loses some key information of objects. These problems might come from two aspects: First, we just construct six layers for extracting feature maps in consideration of the calculation consumption factors, therefore some basic information is dropped accidentally. Second, the superpixels instead of pixels are used to accelerate the computational speed, which will also result in the loss of details of objects. In the future study, we will consider more

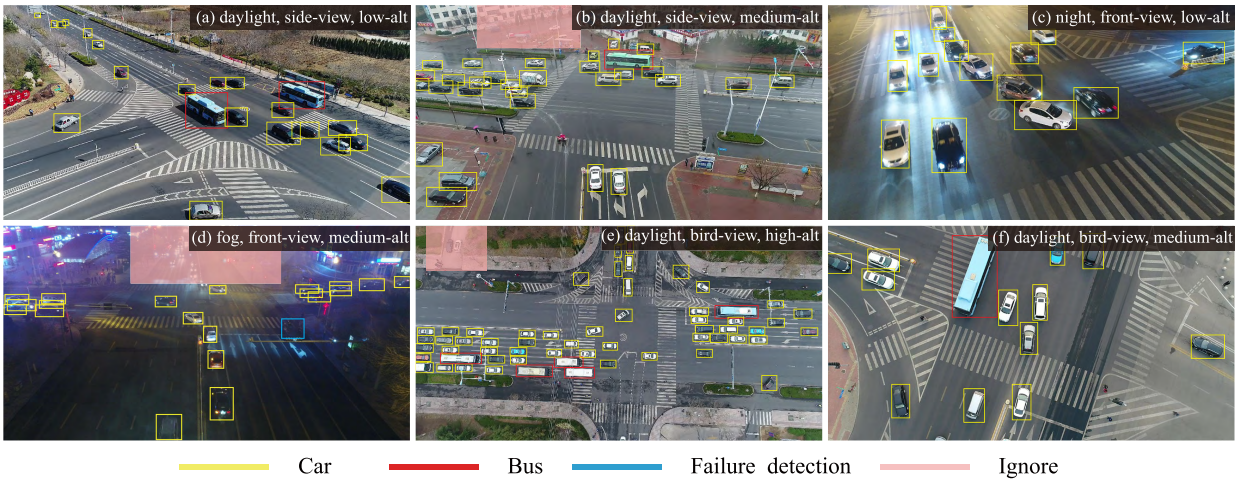

FIGURE 8. **Some detection results on UAVDT data set. The bottom-left image shows detection failure in the blue box.**

powerful structural learning models and multi-model feature fusion methods to enhance the detection ability of the deep networks.

V. CONCLUSIONS

Feature fusion based novel network framework is proposed for object detection in UAV-based aerial images. There are three main types of layers in the proposed FFDN. Different from other common methods, the structural learning in our model is embedded into the network for the purpose of providing more robust spatial information. The unsupervised deep learning methods (CDBN and DSA) are used to extract deep features and spatial information simultaneously with less labeled data. The experimental results verify the remarkable and powerful performance of the proposed FFDN on both UAV123 data set and UAVDT data set. Furthermore, the proposed FFDN is confidently suited for detection application to differentiate ambiguous ground objects in large and aerial images.

REFERENCES

[1] T. Moranduzzo and F. Melgani, "Detecting cars in UAV images with a catalog-based approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6356–6367, Oct. 2014.

[2] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 3, pp. 1635–1647, Mar. 2014.

[3] Z. Chen *et al.*, "Vehicle detection in high-resolution aerial images based on fast sparse representation classification and multiorder feature," *IEEE Trans. Intell. Transp. Syst.*, vol. 17, no. 8, pp. 2296–2309, Aug. 2016.

[4] A. Kembhavi, D. Harwood, and L. S. Davis, "Vehicle detection using partial least squares," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1250–1265, Jun. 2011.

[5] X. Chen, S. Xiang, C.-L. Liu, and C.-H. Pan, "Vehicle detection in satellite images by hybrid deep convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 10, pp. 1797–1801, Oct. 2014.

[6] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.

[7] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[9] S. Bu, P. Han, Z. Liu, and J. Han, "Scene parsing using inference embedded deep networks," *Pattern Recognit.*, vol. 59, pp. 188–198, Nov. 2016.

[10] J. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, San Mateo, CA, USA: Morgan Kaufmann, Jun. 2001, pp. 282–289.

[11] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[12] X. Zhang, P. Jiang, and F. Wang, "Overtaking vehicle detection using a spatio-temporal CRF," in *Proc. IEEE Intell. Veh. Symp. Proc.*, Jun. 2014, pp. 338–343.

[13] L. Xiao, B. Dai, D. Liu, T. Hu, and T. Wu, "CRF based road detection with multi-sensor fusion," in *Proc. IEEE Intell. Veh. Symp.*, Jun./Jul. 2015, pp. 192–198.

[14] Y. Bengio, "Learning deep architectures for AI," *Found. trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

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

[16] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. Eur. Conf. Comput. Vis.*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland, Springer, Oct. 2016, pp. 445–461.

[17] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking,"in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 370–386.

[18] Z. Chen *et al.*, "Vehicle detection in high-resolution aerial images via sparse representation and superpixels," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 1, pp. 103–116, Jan. 2016.

[19] H. Moon, R. Chellappa, and A. Rosenfeld, "Performance analysis of a simple vehicle detection algorithm," *Image Vis. Comput.*, vol. 20, no. 1, pp. 1–13, Jan. 2002.

[20] S. Hinz, C. Schlosser, and J. Reitberger, "Automatic car detection in high resolution urban scenes based on an adaptive 3D-model," in *Proc. 2nd GRSS/ISPRS Joint Workshop Remote Sens. Data Fusion Over Urban Areas*, May 2003, pp. 167–171.

[21] S. Wang, "Vehicle detection on aerial images by extracting corner features for rotational invariant shape matching," in *Proc. IEEE 11th Int. Conf. Comput. Inf. Technol.*, Aug./Sep. 2011, pp. 171–175.

[22] F. Yamazaki, W. Liu, and T. T. Vu, "Vehicle extraction and speed detection from digital aerial images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, vol. 3, Jul. 2008, pp. III-1334–III-1337.

[23] T. Moranduzzo and F. Melgani, "Car speed estimation method for UAV images," in *Proc. IEEE Geosci. Remote Sens. Symp.*, Jul. 2014, pp. 4942–4945.

[24] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li, "Two-stage learning to predict human eye fixations via SDAE$_S$," *IEEE Trans. Cybern.*, vol. 46, no. 2, pp. 487–498, Feb. 2015.
[25] J. Han, C. Chen, L. Shao, X. Hu, J. Han, and T. Liu, "Learning computational models of video memorability from FMRI brain imaging," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1692–1703, Aug. 2015.
[26] Z. Liu, J. Huang, J. Han, S. Bu, and J. Lv, "Human motion tracking by multiple RGBD cameras," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 9, pp. 2014–2027, Sep. 2017.
[27] S. Bu, P. Han, Z. Liu, J. Han, and H. Lin, "Local deep feature learning framework for 3D shape," *Comput. Graph.*, vol. 46, pp. 117–129, Feb. 2015.
[28] S. Bu, Z. Liu, J. Han, J. Wu, and R. Ji, "Learning high-level feature by deep belief networks for 3-D model retrieval and recognition," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2154–2167, Dec. 2014.
[29] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Proc. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 3361–3368.
[30] Z. Liu, Z. Jia, C.-M. Vong, S. Bu, J. Han, and X. Tang, "Capturing high-discriminative fault features for electronics-rich analog system via deep learning," *IEEE Trans. Ind. Informat.*, vol. 13, no. 3, pp. 1213–1226, Jun. 2017.
[31] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[32] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Apr. 2014, pp. 1–16.
[33] Y. Li, A. Møgelmose, and M. M. Trivedi, "Pushing the 'Speed Limit': High-accuracy us traffic sign recognition with convolutional neural networks," *IEEE Trans. Intell. Veh.*, vol. 1, no. 2, pp. 167–176, Jun. 2016.
[34] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
[35] K. He, G. Gkioxari, P. Dollar, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
[36] J. Redmon and A. Farhadi. (Apr. 2018). *YOLOv3: An Incremental Improvement*. [Online]. Available: https://arxiv.org/abs/1804.02767
[37] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," *CoRR*, vol. abs/1701.06659, pp. 1–11, Jan. 2017.
[38] J. Li, J. Qian, and J. Yang, "Object detection via feature fusion based single network," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2017, pp. 3390–3394.
[39] W. Guan, Y. Zou, and X. Zhou, "Multi-scale object detection with feature fusion and region objectness network,"in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2596–2600.
[40] H. Jiang, C. Zhang, and M. Wu, "Pedestrian detection based on multi-scale fusion features," in *Proc. Int. Conf. Netw. Infrastruct. Digit. Content*, Aug. 2018, pp. 329–333.
[41] M. Norouzi, M. Ranjbar, and G. Mori, "Stacks of convolutional restricted Boltzmann machines for shift-invariant feature learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2735–2742.
[42] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proc. Annu. Int. Conf. Mach. Learn.*, New York, NY, USA, Jun. 2009, pp. 609–616.
[43] Z. Han, Z. Liu, J. Han, C.-M. Vong, S. Bu, and X. Li, "Unsupervised 3D local feature learning by circle convolutional restricted Boltzmann machine," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5331–5344, Nov. 2016.
[44] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, Aug. 2002.
[45] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[46] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.
[47] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max- flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124–1137, Sep. 2004.
[48] Z. Deng, H. Sun, S. Zhou, J. Zhao, and H. Zou, "Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 8, pp. 3652–3664, Aug. 2017.
[49] T. Tang, S. Zhou, Z. Deng, H. Zou, and L. Lei, "Vehicle detection in aerial images based on region convolutional neural networks and hard negative example mining," *Sensors*, vol. 17, no. 2, p. 336, Feb. 2017.

HAO LONG received the M.S. degree in control theory and control engineering from Northwestern Polytechnical University, Xi'an, China, in 2005. She is currently pursuing the Ph.D. degree in means of electrical engineering with the National Changhua University of Education, Changhua, Taiwan. Her main research interests include flight control, UAV, and computer vision.

YINUNG CHUNG received the Ph.D. degrees from Texas Tech University, Lubbock, TX, USA, in 1990. He is currently a Full Professor with the National Changhua University of Education, Changhua, Taiwan. He is also the Head of the Department of Electrical Engineering. His research interests include image processing and computer vision.

ZHENBAO LIU (M'11–SM'18) received the B.S. and M.S. degrees in electrical engineering and automation from Northwestern Polytechnical University, Xi'an, China, in 2001 and 2004, respectively, and the Ph.D. degree in system and information engineering from the University of Tsukuba, Tsukuba, Japan, in 2009. He was a Visiting Scholar with Simon Fraser University, Canada, in 2012. He is currently a Professor with Northwestern Polytechnical University. His research interests include UAV, flight control, prognostics and health management, and computer vision. He is an Associate Editor of the IEEE Access.

SHUHUI BU received the master's and Ph.D. degrees from the College of Systems and Information Engineering, University of Tsukuba, Japan, in 2006 and 2009. From 2009 to 2011, he was an Assistant Professor with Kyoto University, Japan. He is currently an Associate Professor with Northwestern Polytechnical University, China. His research interests are concentrated on robotics, UAV, and computer vision.