IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 24, 2018, accepted January 14, 2019, date of publication April 1, 2019, date of current version April 15, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2896293

Real-World ISAR Object Recognition and Relation Discovery Using Deep Relation Graph Learning

BIN XUE, (Student Member, IEEE), AND NINGNING TONG
Air Force Engineering University, Xi'an 710051, China

Corresponding author: Bin Xue (xxbbxl@sina.com)

This work was supported in part by the National NSFC under Grant 61571459, Grant 61631019, and Grant 61701526.

ABSTRACT Real-world inverse synthetic aperture radar (ISAR) object recognition is the most critical and challenging problem in computer vision tasks. In this paper, an efficient real-world ISAR object recognition and relation discovery method are proposed, based on deep relation graph learning. It not only handles the real-world object recognition problem efficiently, but also exploits the inter-modal relationships among features, attributes, and classes with semantic knowledge. First, dilated deformable convolutional neural network, including dilated deformable convolution and dilated deformable location-aware RoI pooling, is introduced to greatly improve CNNs' sampling and transformation ability, and increase the output feature maps' resolutions significantly. And a related multi-modal regions ranking strategy is proposed. Second, deep graph attribute-association learning is proposed to jointly estimate a large number of multi-heterogeneous attributes, and leverage features, attributes, and semantic knowledge to learn their relations. Third, multi-scale relational-regularized convolutional sparse learning is proposed to further improve the accuracy and speed of the whole system. The extensive experiments are performed on two real-world ISAR datasets, showing our proposed method outperforms the state-of-the-art methods.

INDEX TERMS Deep relation graph learning, dilated deformable, multi-scale relational-regularized convolutional sparse learning, inverse-synthetic-aperture-radar, real-world object recognition.

I. INTRODUCTION

Complex object recognition has been recognized as a most critical and challenging problem in computer vision tasks, and has attracted a great deal of interest [1], [2]. Particularly, real-world inverse synthetic aperture radar (ISAR) object (such as the moving helicopters, airplanes, naval vessels, cars) recognition [3], is more difficult than object recognition within natural and infrared image.

Because there are some serious obstacles to achieve an excellent ISAR object recognition system: (1) the challenging multi-modal problem, such as numberous different viewpoints, scales, poses, deformations, occlusions, blurs, low resolutions, polarizations and extreme off angle within ISAR images; (2) because of the special coherence tomography style and imaging environment, ISAR images are randomly covered with kinds of complex noises, the structure and scattering characteristics of ISAR objects are seriously weakened; (3) ISAR image objects are generally smaller than natural image objects. Fig. 1 shows the illustrations of the same real-world ISAR object, showing objects within ISAR images are more complex than the ones within natural and infrared images.

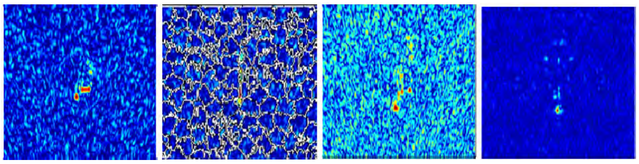

FIGURE 1. The illustrations of the same real-world ISAR object.

In general, there are some methods aimed to solve multi-modal problem, such as affine transformation [4] and scale invariant feature transform (SIFT) [5], but they only can handle a few fixed and known deformation styles with much expensive cost and insufficient deformable samples, which cannot solve the multi-modal problem well.

Recently, the successful application of deep convolutional neural networks (DCNNs) encourages us that DCNN may be one of the most promising means to handle the problem above [6]. Though DCNN has the powerful ability to extract features, there are still some problems. CNNs are inherently limited to model large, unknown transformations because of

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

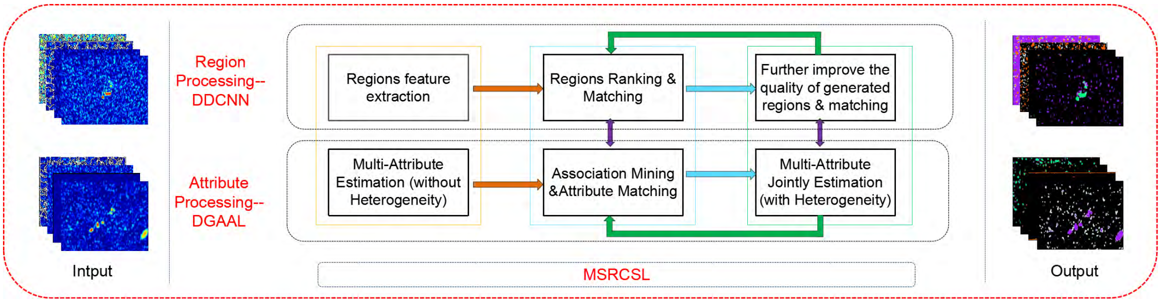
Region Processing--DDCNN
Attribute Processing--DGAAL
Regions feature extraction
Regions Ranking & Matching
Further improve the quality of generated regions & matching
Multi-Attribute Estimation (without Heterogeneity)
Association Mining &Attribute Matching
Multi-Attribute Jointly Estimation (with Heterogeneity)
Intput
MSRCSL
Output

FIGURE 2. Framework of the proposed method.

CNN modules' fixed geometric structures: the input feature map at fixed positions; a region-of-interest (RoI) pooling layer separates a RoI into fixed bins and so on. Moreover, most of current DCNNs owe very small sizes of receptive field for convolution sampling, limiting the sampling range severely. Simultaneously, the resolution of the output feature maps will be progressively reduced by convolutional networks for localization and recognition, until the image is represented by tiny feature maps that retain little spatial information, and the spatial structure is no longer discernible, such as some ISAR images' feature maps. There lacks internal mechanisms to handle the multi-modal recognition problem.

Moreover, although extensive efforts have been devoted for feature extraction in recent years [7], [8], most existing works neglected attributes mining and the correlations of features, attributes and classes during learning, while most of the rest, simply combined multiple familiar features with simple fusion strategies, and no one consider such complex multi-modal objects.

In this paper, an end-to-end DRGL is proposed for multi-modal ISAR object recognition, which not only handles the complex multi-modal problem fast and efficiently, but also exploits the relationships among feature, attribute and classes. (1) Dilated Deformable Convolution (DDC) and Dilated Deformable Location-Aware (DDLA) RoI Pooling are introduced to greatly improve CNNs' sampling and transformation ability, and increase the resolution of output feature maps. And an associated multi-modal regions ranking strategy is proposed. (2) Deep Graph Attribute-Association Learning (DGAAL) is proposed to jointly estimate large numbers of heterogeneous attributes, and exploit the features, attributes and the semantic knowledge to learn the inter-modal associations. (3) Multi-Scale Relational-Regularized Convolutional Sparse Learning (MSRCSL) is proposed to further improve the accuracy and speed of the whole system. The proposed method is evaluated on two multi-modal ISAR datasets with extensive experiments, showing that our proposed method outperforms several state-of-the-art recognition methods.

II. PROPOSED METHOD

A. NETWORK ARCHITECTURE

The overall framework of CIOR is shown in Fig. 2. DMCNN is used to handle the multimodal recognition problem, DGAAL is used to jointly estimate large numbers of heterogeneous attributes and exploit the inter-modal relationships among features, attributes, labels, classes, and semantic knowledge, while MSRCSL is used to improve the accuracy and speed of the entire system.

The proposed DRGL contains a deep network (the former) to learn shared attributes-features, followed by some shallow-layer networks (the latter) to learn class-individual attributes-features. The shared attributes-features are fine-tuned by the latter networks to get an optimal estimation of individual attributes, features and relationships. Particularly, a modified VGG-14 model is used as the default baseline deep model (the shared learning section includes 4 convolutional (Conv.), 4 batch normalization (BN), 4 pooling, 2 fully connected (FC) layers. Particularly, each BN layer is between each Conv. and pooling layer).

B. DILATED DEFORMABLE CONVOLUTIONAL NEURAL NETWORK

DDC and DDLA RoI Pooling are introduced to provide the receptive field with exponential level expansion without resolution loss or coverage, enable free form deformation of the sampling grid, which greatly improves CNNs' sampling and transformation ability, and increases the output feature maps' resolution. Both the two modules are light weight. They add small amount of computation for the offset learning. They can readily replace their plain counterparts in DCNNs and can be easily trained end-to-end with standard backpropagation (BP).

1) DILATED DEFORMABLE CONVOLUTION

In the dilated modules [9], [10], $F_i(i = 0, 1, ..., n - 1) : Z^2 \rightarrow R$ and $k_i(i = 0, 1, ..., n - 2) : \Omega_r \rightarrow R$ are defined as discrete functions and discrete filters of size $(2r + 1)^2$, respectively, $\Omega_r = [-r, r]^2 \cap Z^2$. Applying the filters with exponentially increasing dilation:

$$F_{i+1} = F_i{}^*[2i]k_i \tag{1}$$

$$(F_i{}^*[e]k)(l) = \sum_{s+et=l} F_i(s)k(t) \tag{2}$$

where e is a dilation factor, and $^*[e]$ is defined as the discrete l-dilated convolution operator. The receptive field of an element l in F_{i+1} is defined as the group of elements in F_0 that

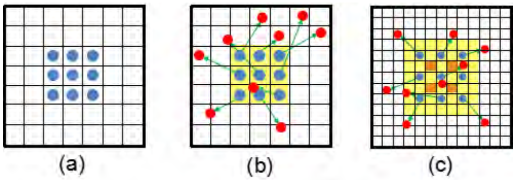

FIGURE 3. Illustration of the sampling positions in 3 × 3 SC and DDC.

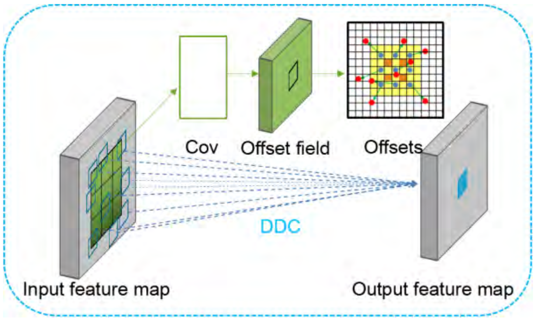

FIGURE 4. Illustration of 3 × 3 DDC.

revises $F_{i+1}(l)$. Let the size of the receptive field of l in F_{i+1} be the number of these elements. This is illustrated in Fig. 3 (yellow grid).

At the same time, with the deformable modules [11], [12], it adds 2D offests to the grid sampling positions in the standard convolution (SC). Fig. 3 is the illustration of the sampling positions in 3 × 3 SC and DDC. Fig. 3 (a) shows regular sampling grid $R$of SC. Fig. 3 (b) (c)are 1-dilated and 2-dilated deformed sampling with Fig. 3 (a)'s different cases of deformed sampling positions (red points) with augmented offsets (green arrows) in DDC. Fig. 4 shows the illustration of DDC. In DDC, for each location l_1 on the output feature map y, we have

$$y(l_1) = \sum_{l_n \in R} w(l_n) \cdot x(l_1 + l_n + \Delta l_n) \tag{3}$$

Regular grid R is augmented with offsets $\{\Delta l_n | n = 1, \ldots, N\}$, where $N = \{R\}$. Sampling is on the irregular and offset locations $l_n + \Delta l_n$.

2) DILATED DEFORMABLE LOCATION-AWARE RoI POOLING

It adds an offset to each bin position in the regular bin partition of the previous RoI pooling. Similarly, the offsets are learned from the preceding feature maps and the RoIs, enabling adaptive part localization with different shapes. In this paper, DDLA RoI Pooling [13] is in the last pooling layer in VGG-14. Fig.5 is the illustration of DDLA RoI Pooling.

Given the input feature map x and a RoI of size $w \times h$ and top-left corner l_1, RoI pooling divides the RoI into $k \times k$ bins and outputs a feature map y. In DDLA RoI pooling, for (i, j)-th bin $(0 \leq i, j < k)$, we have

$$y(i, j) = \sum_{l \in bin(i,j)} x(l_1 + l + \Delta l_{ij})/n_{ij} \tag{4}$$

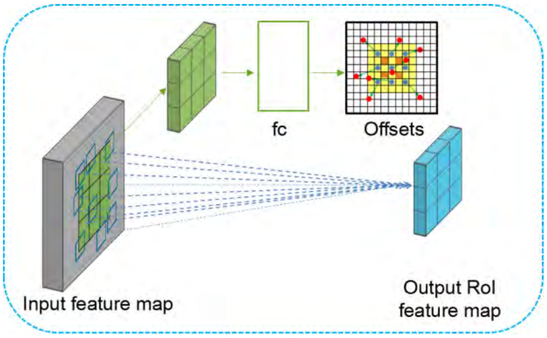

FIGURE 5. Illustration of 3 × 3 DDLA RoI pooling.

where n_{ij} is the number of pixels in the bin, the (i, j)-th bin spans $\lfloor i(w/k) \rfloor \leq l_x < \lceil (i+1)(w/k) \rceil$ and $\lfloor j(h/k) \rfloor \leq l_y < \lceil (j+1)(h/k) \rceil$, offsets $\{\Delta l_{ij} | 0 \leq i, j < k\}$ are added to the binning locations.

3) ASSOCIATED MULTI-MODAL REGIONS RANKING

It may produce many redundant or unlikely multi-modal object regions with the introducing of DDC and DDLA RoI Pooling. So a ranker is proposed, to offer an ordering of a group of associated multi-modal regions, which may belong to the same objects or the same scenes with different viewpoints or deformation styles and so on, and ensure that each object owes a set of top-ranked associated multi-modal regions, which simultaneously suppresses both the undesirable redundant and unlikely sample regions.

Our ranker incrementally adds regions, from best to worst, based on the combination of an object appearance score and a penalty for overlapping with previously added associated multi-modal regions. By taking into account the overlap with higher ranked associated multi-modal regions, our ranker ensures that redundant regions are suppressed, forcing the top-rank regions to be diverse.

By writing a scoring function $S(x, r, w)$ over the set of associated multi-modal regions x and their ranking r, we cast the ranking problem as a joint inference problem [14]. The goal is to find the parameters w such that $S(x, r, w)$ gives higher scores to rankings that place associated multi-modal regions for all objects in high ranks.

$$S(x, r, w) = \sum_i \alpha(r_i) \cdot (w_\alpha^T \Psi(x_i) - w_p^T \Phi(r_i)) \tag{5}$$

The score is a combination of appearance features $\Psi(x)$ and overlap penalty terms $\Phi(r)$, where r denotes the rank of a set of multi-modal regions, ranging from 1 to the number of regions M. This allows us to jointly learn the appearance model and the trade-off for overlapping regions. $\Phi(r)$ is the concatenation of two vectors $\Phi_1(r)$; $\Phi_2(r)$: $\Phi_1(r)$ penalizes regions with high overlap with previously ranked associated multi-modal regions, and $\Phi_2(r)$ further suppresses associated multi-modal regions that overlap with multiple higher ranked regions. The second penalty is necessary to continue to enforce diversity after many multi-modal regions have at least one overlapping associated multi-modal region. Since the strength of the penalty should depend on the amount of overlap (regions with 90% overlap should be suppressed more than regions with 50%), we want to learn overlap specific weights. To do this, we quantize the overlaps into bins of 10% and map the values to a 10 dimensional vector $q(ov)$ with 1 for the bin it falls into and 0 for all other bins.

$$\Phi_1(r_i) = q(\max_{\{j|r_j<r_i\}} ov(i,j)) \tag{6}$$

$$\Phi_2(r_i) = \sum_{\{j|r_j<r_i\}} q(ov(i,j)) \tag{7}$$

The overlap score between two regions is computed as the area of their intersection divided by their union, with A_i indicating the set of pixels belonging to region i:

$$ov(i,j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|} \tag{8}$$

Each region's score is weighted by $\alpha(r)$, a decreasing function. Because higher ranked associated multi-modal regions are given more weight, they are encouraged to have higher scores. We found the specific choice of $\alpha(r)$ is not particularly important, as long as it falls to zero for a moderate rank value. We use $\alpha(r) = \exp(\frac{(r-1)^2}{\sigma^2})$ with $\sigma = 100$.

Computing $\max_r \ S(x, r, w)$ cannot be solved exactly, so a greedy approximation that incrementally adds the associated multi-modal regions with the maximum marginal gain is used. We found that this works well for a test problem where full enumeration is feasible, especially when $ov(\cdot, \cdot)$ is sparse, which is true for this ranking problem.

DRGL takes a set of diverse hierarchies and computes the n-tuples up to a certain height of the ranked region in the tree. The n-tuples from each hierarchy can be interpreted as a ranked list of N_i regions that are put together to produce the final pool of N_p regions.

4) FURTHER IMPROVE THE QUALITY OF FEATURE MAPS AND THE SPEED OF REGIONS PROCESSING

To further reduce the quantity of the ranked multi-modal regions, a regressor is trained from low- and mid-level features. Since the top-ranked associated multi-modal regions are all formed by a set of regions from a reduced set of hierarchies, the features which can be computed fast and efficiently in a bottom-up style are focused on, including the boundary, size and position (area and perimeter of the candidate, scale and aspect ratio of the bounding box).

The object overlap with the ground-truth is regressed using a Random Forest [15], trained with these features above, and the ranking based on Maximum Marginal Relevance measures [16] is diversified, which the measure only considers the best overlap with each ground-truth object, and finally only the top ranked regions are retained.

To further reduce the dimensionality of the search space, we start by selecting two ranked lists L_1, L_2 and the list at S levels of number of regions are sampled. Then the full different parameters are scanned to combine the associated multi-modal regions from both. The final sets of regions are generated by combining the top N_1 from L_1 and the top N_2 from L_2. The process is iterated until all the ranked lists are combined. The number of sampled configurations using the proposed algorithm is $(R-1)S^2$, that is, we have reduced an exponential problem (S^R) to a quadratic one.

C. DEEP GRAPH ATTRIBUTE-ASSOCIATION LEARNING

A DGAAL [17]–[19] is presented to jointly estimate multiple heterogeneous attributes, and discovery the inter-modal of image dataset, features, and attributes with higher-level semantic knowledge, which takes into account both attribute correlation and heterogeneity into a single convolutional neural network. DGAAL contains of a deep network to lean shared attribute-features for all the attributes, followed by some shallow-layer networks to learn class-individual attribute-features. Fig. 6 is the illustration of DGAAL.

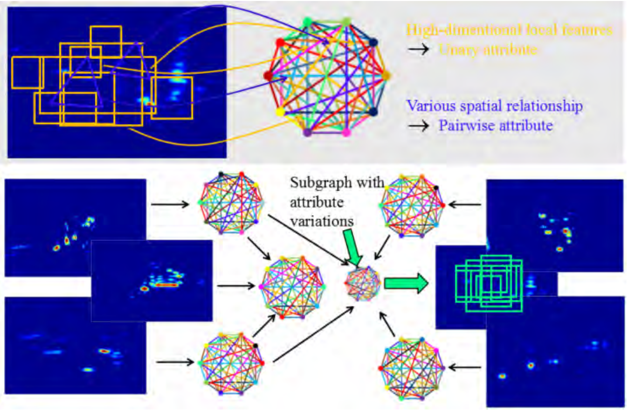
High-dimentional local features
→ Unary attribute
Various spatial relationship
→ Pairwise attribute
Subgraph with attribute variations

FIGURE 6. ARG, SAP, variable attribute relations discovery modeling.

1) THE DEFINITION OF ARG AND SAP

A Soft Attributed Pattern (SAP) is defined to represent the common conjunct subgraph pattern among a group of attributed relational graphs (ARGs) to estimate node correspondences, which takes into account both the attributes and architectures. And a novel mining method which efficiently draws the ARG is introduced.

a: ARG DEFINITION

An ARG G is defined as $G = (V, F_V, F_{V\times V})$, where V is the node set. Undirected edges connect each pair of nodes to form a complete graph. There are N_p types of local attributes for each node and N_Q types of pairwise attributes for each edge in G. $F_V = \{F_i^s | s \in V, i = 1, 2, \cdots, N_P\}$ and $F_{V\times V} = \{F_j^{st} | s, t \in V, s \neq t, j = 1, 2, \cdots, N_Q\}$ are the local and pairwise attribute sets, respectively.

b: SAP DEFINITION

Given a set of ARGs $GS = \{G_k^{'} | k = 1, 2, \cdots, N\}$ and a threshold τ, if graph template $G = (V, F_V, F_{V\times V})$

satisfies among the ARGs the following three conditions, G is an SAP:

(1) $\hat{x}^k = \arg\min_{x^k} \varepsilon(x^k|G, G_k^{'})$; we set

$\hat{x}^k = \{x_s^k|s \in V, k = 1, 2, \cdots, N\}$

(2) $(F_V, F_{V\times V}) \leftarrow \arg\min_{F_V, F_{V\times V}} \sum_{k=1}^{N} \varepsilon(\hat{x}^k|G, G_k^{'})$

(3) $\forall s \in V, E_s(\{\hat{x}^k\}|G, G_k^{'}) \leq \tau$

where $E_s(\{\hat{x}^k\}|G, G_k^{'})$ is the average matching penalty of node s in G among all the ARGs in GS.

$$E_s(\{\hat{x}^k\}|G, G_k^{'}) = \frac{1}{N}\sum_{k=1}^{N}\Bigg[P_s(\hat{x}^k|G, G_k^{'}) + \sum_{t\in V, t\neq s} Q_{st}(\hat{x}_s^k, \hat{x}_t^k|G, G_k^{'})\Bigg] \tag{9}$$

2) MULTI-ATTRIBUTE JOINTLY ESTIMATION

To jointly estimate multiple various attributes instead of individual attribute, and ingeniously perform a majority of attributes correlation and shared feature learning, DGAAL is formulated:

$$\arg\min_{W_c, \{W^j\}_{i=1}^{M}} \sum_{j=1}^{M}\sum_{i=1}^{N} L(y_i^j, F(X_i, W^j \circ W_c)) + \gamma_1\Phi(W_c) + \gamma_2\Phi(W^j) \tag{10}$$

where X_i is input, W_c and W^j are weight vectors, $F(\cdot)$ denotes the attribute forecasting function of X_i and W_c and W^j; y_i^j is the ground-truth values $F(X_i, W^j \circ W_c)$, $L(\cdot)$ means the loss function between $F(X_i, W^q \circ W_c)$ and y_i^j; $\Phi(\cdot)$ and γ denotes the regularization term and parameter respectively.

3) ATTRIBUTE GRAPH MATCHING

Given a set of ARGs $GS = \{G_k^{'}|k = 1, 2, \cdots, N\}$, $G_k^{'} = (V_k, F_{V_k}, F_{V_k\times V_k})$, the graph template G represents an attribute pattern among the ARGs in GS. The matching between G and $G_k^{'}$ aims to compute a set of matching assignments between G and $G_k^{'}$, denoted by $x^k = \{x_s^k|s \in V\}$. Each matching assignment $x_s^k \in V_k \cup \{\varepsilon\}$ maps node n in G to either a node in $G_k^{'}$ or a dummy choice ε is used when some nodes in G do not exist in $G_k^{'}$. The graph matching is formulated as a typical QAP with the following energy function:

$$\varepsilon(x^k|G, G_k^{'}) = \sum_{s\in V} P_s(x_s^k|G, G_k^{'}) + \sum_{(s,t)\in V, s\neq t} Q_{st}(x_s^k, x_t^k|G, G_k^{'}) \tag{11}$$

where $\varepsilon(x^k|G, G_k^{'})$ indicates the total matching energy. $P_s(\cdot)$ and $Q_{st}(\cdot,\cdot)$ denote matching penalties for local and pairwise attributes. Various graph matching optimization techniques can solve the energy minimization of $\varepsilon(x^k|G, G_k^{'})$, and we choose TRW-S (tree-reweighted message passing). In this study, matching penalties are defined using squared differences.

$$P_s(x^k|G, G_k^{'}) = \begin{cases} \sum_{i=1}^{N_p} \omega_i^P \left\| F_i^s - F_j^{x_s^k} \right\|^2, & x_s^k \in V_k \\ P_\varepsilon, & x_s^k = \varepsilon \end{cases} \tag{12}$$

$$Q_{st}(x_s^k, x_t^k|G, G_k^{'}) = \begin{cases} \dfrac{\sum_{j=1}^{N_Q} \omega_j^Q \left\| F_i^s - F_j^{x_s^k x_t^k} \right\|^2}{\|V\| - 1}, & x_s^k \neq x_t^k \in V_k \\ +\infty, & x_s^k = x_t^k \in V \\ \dfrac{Q_\varepsilon}{\|V\| - 1}, & x_s^k \ or\ x_t^k = \varepsilon \end{cases} \tag{13}$$

where P_ε and Q_ε are relatively large constant penalties for matching to the dummy node ε in the case of occlusions. $\|\cdot\|$ is the Euclidean norm. We use infinite penalties to avoid many-to-one matching assignments. ω_i^P and ω_j^Q denote the weights for local and pairwise attribute differences respectively.

We require the pairwise penalty to be symmetric, i.e., $Q_{st}(x_s, x_t|G, G_k^{'}) = Q_{ts}(x_t, x_s|G, G_k^{'})$, and to be normalized by $\|V\| - 1$. Penalties P_ε and Q_ε and attribute weights $\{\omega_i^P\}$ and $\{\omega_j^Q\}$ can be manually set or automatically mined.

4) ASSOCIATED MULTI-MODAL REGIONS MATCHING

A probabilistic Bayesian model is proposed for related multi-modal regions matching, and an efficient matching strategy based on local regularization is described.

a: ① PROBABILISTIC BAYESIAN MODEL FOR ASSOCIATED MULTI-MODAL REGIONS MATCHING

R and R' denote two groups of region samples extracted from images I and I', respectively. $r = (f, s)$ is a region sample with appearance feature f and spatial support s, $r \in R$. f and s represent the region's visual descriptor and the set of all pixel locations respectively. Given the data $D = (R, R')$, to estimate the posterior matching probability score of region sample $r_{\text{in}}R$ matches region sample r' in R', $r \mapsto r'$, we have:

$$p(r \mapsto r'|D) = p(f \mapsto f')p(s \mapsto s'|D) \tag{14}$$

where $p(f \mapsto f')$ denotes a similarity between feature descriptors f and f', and $p(s \mapsto s'|D)$ is computed by the spatial supports s and s', $s, s' \in D$. Assign the best match $\phi(r)$ for every region sample in R:

$$\phi(r) = \arg\max_{r'\in R'} p(r \mapsto r'|D) \tag{15}$$

If $(f', s') = \phi(f, s), f' = \phi(f)$ and $s' = \phi(s)$.

b: ② LOCAL OFFSET GEOMETRIC MATCHING STRATEGY (LOGMS)

To achieve reliable correspondences considering both the appearance, geometric relationship, and the noised clutter, which distracts outlier regions, a novel local offset geometric matching strategy is proposed.

For each region sample, to estimate a reliable offset for each region sample r in a robust manner without any information about objects and their positions, a reliable offset is considered for translation and scale which exploits only the neighboring region samples. For each region r, its neighborhood $N(r)$ is defined as the set of regions with overlapping spatial support:

$$N(r) = \{\hat{r} | s \cap s' \neq \emptyset, \hat{r} \in R\} \tag{16}$$

Using an initial correspondence $\psi(r)$, determined by the best match, each neighboring region $\hat{r}$ is assigned its own offset, and all of them form a set of neighbor offsets:

$$X(r) = \{r(\hat{s}) - r(\psi(\hat{s})) | \hat{r} \in N(r)\} \tag{17}$$

From the set of neighbor offsets, a local offset x_r^* for the region r by the geometric median [21] is estimated:

$$x_r^* = \arg\min_{x \in R^3} \sum_{y \in X(r)} \|x - y\|_2 \tag{18}$$

The local offset x_r^* for the region r is estimated by regression using its local neighboring offsets $X(r)$. Based on x_r^* optimized for each region, we define the geometric consistency function:

$$g(s \mapsto s'|D) = p(s \mapsto s'|x_r^*) \sum_{\hat{r} \in N(r)} p(\hat{f} \mapsto \psi(\hat{f})) \tag{19}$$

which can be interpreted as the fact that the region $r_{\text{in}}R$ is likely to match r' in R' where its offset $\gamma(s) - \gamma(s')$ is close to the local offset x_r^*, and the region r has many neighboring matches with a high appearance fidelity. By using $g(s \mapsto s'|D)$ as a proxy for $p(s \mapsto s'|D)$, LOGMS imposes local smoothness on offsets between neighboring regions.

5) ATTRIBUTE AND FEATURE HETEROGENEITY

Although attribute correlation and region matching are considered in feature learning above, the attribute heterogeneity still needs to be considered. We treat each of the heterogeneous attribute categories separately, but attributes within each category are expected to share feature learning and classification model to a larger extent. To accomplish this, the objective function is written as:

$$\begin{aligned} &\arg\min_{W_c, \{W^j\}_{i=1}^M} \sum_{g=1}^{G} \sum_{j=1}^{M} \sum_{i=1}^{N} \lambda^g L^g(y_i^j, F(X_i, W^g \circ W)) \\ &\quad + \arg\min_{W_c, \{W^j\}_{i=1}^M} \sum_{g=1}^{G} \sum_{j=1}^{M} \sum_{i=1}^{N} \gamma_1 \Phi(W_c) + \gamma_2 \Phi(W^g) \end{aligned} \tag{20}$$

where G is the number of heterogeneous attribute categories, and M is the number of attributes within each attribute category; λ^g balances the importance of each attribute category; W^g refines the shared features w.r.t. each of the heterogeneous attribute class. $L^g(\cdot)$ is a loss function for each of the heterogeneous attribute categories, given the estimated values by F and the corresponding ground-truth y_i^j.

TABLE 1. Definitions of the symbols in (21).

Sym.	Definition
$\hat{Z}$	Includes the sparse representations for $B^{d\times K}$ in X
$\|\cdot\|_F^2$ $\|\cdot\|_{2,1}^2$	The matrix's F- and $l_{2,1}$-norm, respectively
D, γ	A non-negative matrix and a regularization parameter, respectively.
	A Hadamard product operator of matrices
R_1	A X-X (X denotes feature, attribute, or knowledge) association-based regularization term
R_2	A subject-subject association-based regularization term
φ_1 φ_2	R_1 and R_2's regularization parameters, respectively.

D. MULTI-SCALE RELATIONAL-REGULARIZED CONVOLUTIONAL SPARSE LEARNING

To get good translation invariance, accelerate, and improve the performance of joint learning multi-modal relational features, and provide an unsupervised transfer learning method based feature attribute and relationship knowledge, we combine multi-scale convolutional sparse coding with relational regularization on the grid structures. Define a local descriptor set $X = [x_1, x_2, \cdots, x_N] \in R^{d\times N}$ as an image, $B^{d\times K}$is a dictionary which represents the local descriptors, where x_i is the *ith* local descriptor column of X, K is the dictionary's size. A descriptor set's sparse representations can be expressed as (TABLE 1 shows the define of the symbols in (21)):

$$\begin{aligned} \hat{Z} = &\arg\min_Z \|X + D \odot C - BZ\|_F^2 \\ &+ \gamma \|Z\|_{2,1}^2 + \varphi_1 R_1(Z) + \varphi_2 R_2(Z) \end{aligned} \tag{21}$$

$$R_1(Z) = \sum_{u,v=1}^{FT} \exp(-\|x^u - x^v\|_2^2) \|z^u - z^v\|_2^2 \tag{22}$$

$$R_2(Z) = \sum_{j,k=1}^{S} \exp(-\|x_j - x_k\|_2^2) \|x_j Z - x_k Z\|_2^2 \tag{23}$$

where x^u is the uth column of the input data X, z^u is the uth row of Z, X_j is the jth row of X, FT and S denotes feature, attribute and subject, respectively. Particularly, feature-subject associated context are integrated in a regularized discriminative least squares regression module with $l_{2,1}$-norm to exploit the underlying internal associations.

Define the image-level features with a concatenation of max pooling features described on m^2 spatial girds:

$$\beta = \bigcup_{c=1}^{m^2} [\xi_{\max}(\hat{Z}_{I_c})] \tag{24}$$

where $\beta \in R^{m^2 K}$, $\xi_{\max}$ is defined on each row of $\hat{Z}$, $\bigcup[\,]$ is vector concatenation operator, and I_c is index set for the descriptors falling into the receptive field of *cth* grid. The max pooling feature is invariant to translations of the local descriptors within each grid.

E. TRAINING

There are quite massive parameters to learn, so DCNN requires a large amount of training data. However, there is just a small amount of available annotated training data, which becomes one of the bottlenecks in feasible DCNN training. To relief the problem, the several DRGL sections are weakly-supervised pre-trained, which started from supervised pre-training, ended to unsupervised pre-training and then the integrated detection model is fine-tuned to detect. It demonstrates that adapt the pre-trained DCNNs processed using some state-of-the-art methods in an approximate way with a spot of annotated and sufficent unlabeled training data also can achieve a good performance. The DCNN was pre-trained on ISAR-1 ISAR-2 supervisedly and unsupervisedly jointly dataset, which is different with orthers' strategies and is significantly benificial to the following training, and the pre-training is performed with Caffe [22].

Forward computation (FC), BP and stochastic gradient descent with momentum (MSGD) is used to train DRGL end-to-end. It is efficient to utilize FC and BP strategies when the receptive fields of the final layer overlap seriously. We train the dictionary for local descriptors through BP, by minimizing the training error of the image level features, which are extracted by max pooling over the sparse codes. The achieved dictionary is remarkably more effective than the unsupervised one in terms of classification. And the pooling procedure over different spatial scales equips the proposed model with local translation-invariance similar to the convolutional network. All new layers are randomly initialized by drawing weights from a zero-mean Gaussian distribution with standard deviation 0.01.

III. EXPERIMENTS AND RESULTS

A. DATASETS

Two multi-modal ISAR datasets are constructed for multi-modal ISAR object recognition: ISAR-1 and ISAR-2, which consider intra-class variations and multi-modal conditions. Assuming the ISAR objects present in the images and their parts may undergo shape deformation, occlusion, thin plate spline (TPSs) is used to interpolate sparse keypoints to generate ground-truth considering the multi-modal conditions. TPS Warping is used directly to estimate a warping function using ground truth keypoint annotations from sparse correspondences. ISAR-1 consists of 5 ISAR object classes with 10 keypoint annotations for each image.

Note that these images contain more multi-modal issues than existing datasets, which include mainly images with tightly cropped regions. According to viewpoint or background clutter, the images are split into 8 sub-classes. Given the images and regions, ground-truth data between all possible image pairs with each sub-class are produced. The two multi-modal ISAR datasets are more challenging than other existing datasets for multi-modal object recognition.

B. ATTRIBUTE ANNOTATION

The training sets of the two multi-modal ISAR datasets have labels indicating the identities of the ISAR objects. In addition, a total of 86 binary user-defined attributes have been annotated. We remove the user-defined attributes which do not appear in each dataset, and the numbers of the remaining attributes are 62 and 36 for ISAR-1 and ISAR-2, respectively.

C. EXPERIMENT IMPLEMENTATIONS

The same training strategies are utilized in the whole process, the proposed method was evaluated in two multi-modal ISAR datasets with Caffe. Experiment with 6000, 3000 and 2000 images on the training, validation, and testing set respectively. Using fixed learning rates of 10^{-4}, IoU 0.7, weight decay of $5*10^{-4}$. DRGL is trained end-to-end using forward computation, BP and stochastic gradient descent. The mean Average Precision (mAP) is selected as the evaluation metric. Fig. 7 shows the results with the proposed method.

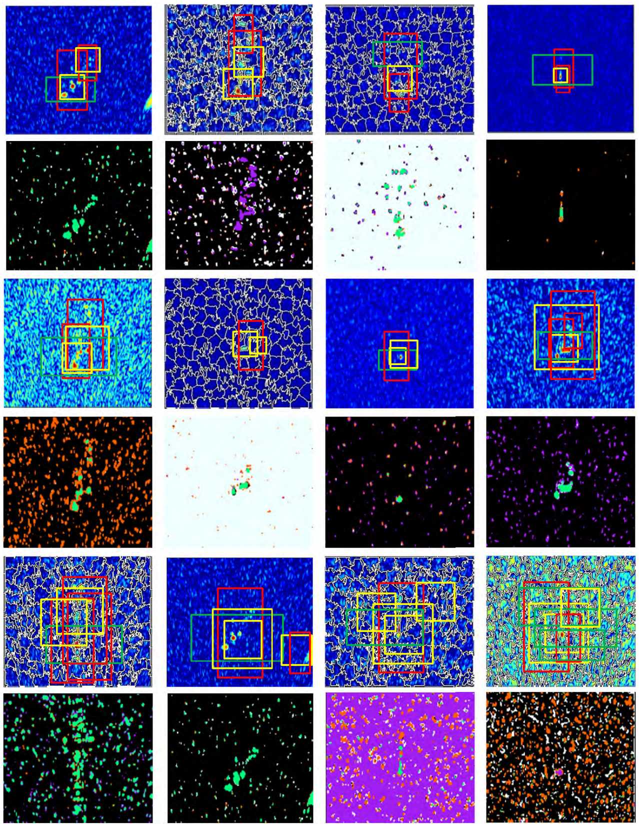

FIGURE 7. Complex ISAR object recognition results and the last layer feature maps of the ISAR objects. The aspect ratio of the red, green and yellow box denote 2:1, 1:2 and 1:1, respectively. And there are 4 scales of the box, 128 × 128, 256 × 256, 512 × 512, and 1024 × 1024.

D. EVALUATIONS

Extensive ablation studies are performed to validated the efficiency of our method. Table 2 evaluates the effect of DDC with our proposed method and several state-of-the-art models, such as RCNN [23], Faster R-CNN [24], R-FCN [25]

TABLE 2. Evaluation of the effects of the proposed modules with several methods.

Usage of DDC	None(0,baseline)	VGG c(1)	VGG b,c(2)	VGG a,b,c(3,default)	Runtime/s
RCNN	38.6	43.5	46.8	45.2	12.635
Faster R-CNN	45.9	51.2	54.1	53.6	7.914
R-FCN	48.1	52.6	54.6	56.9	6.590
Mask R-CNN	49.5	53.3	56.1	58.2	5.534
He [32]	43.5	-	-	-	15.915
Tang [33]	39.8	-	-	-	28.392
Jiao [34]	43.1	-	-	-	46.148
*1	49.6	52.5	53.9	54.2	7.568
*2	53.8	56.1	58.3	61.1	5.893
*3	51.3	54.5	58.4	59.6	6.392
*4	53.2	57.8	61.5	63.8	4.268
*5	54.5	56.7	58.8	60.1	5.489
*6	58.3	61.5	63.1	65.8	3.136

TABLE 3. Evaluation of the effects of all the proposed modules with several state-of-the-art methods.

Method	None	DDC	DDLA RoI Pooling	DDC&DDLA RoI Pooling	Runtime/s
RCNN	38.6	45.2	N/A	45.2	12.635
Faster R-CNN	45.9	53.6	48.8	56.1	6.821
R-FCN	48.1	56.9	53.4	59.3	4.628
Mask R-CNN	49.5	58.2	56.5	59.5	4.269
He [32]	43.5	-	-	-	15.915
Tang [33]	39.8	-	-	-	28.392
Jiao [34]	43.1	-	-	-	46.148
*1	49.6	54.2	53.3	57.4	5.315
*2	53.8	61.1	56.2	63.4	3.164
*5	54.5	60.1	56.7	61.3	4.257
*6	58.3	65.8	61.9	67.2	2.621

and Mask R-CNN [26]. The performance of the proposed method is evaluated using the VGG-16 [27] and the modified VGG-14 (default) with or without DGAAL and MSRCSL, respectively. The effect of DCC in the several state-of-the-art models also are invested in the last 3 convolutional layers.

Particularly, DGAAL and MSRCSL are not used in RCNN, Faster R-CNN, R-FCN and Mask R-CNN in this paper, because the associated multi-modal region sample based methods with DDC cost much more time with DGAAL and MSRCSL. Table 2 shows that accuracy of all these models steadily improves, when more DDC layers are used while the running time also increased. But we find that it cannot always achieve the best performance when all the Conv. layers are replaced with DDC, so the last 3 Conv. layers of these models are selected. In Table 2 and 3, *1, *2 *3, *4, *5, *6 denote our proposed method with VGG-16 (without DGAAL and MSRCSL), with the modified VGG-14 (without DGAAL and MSRCSL), with VGG-16 (with DGAAL without MSRCSL), with the modified VGG-14 (with DGAAL without MSRCSL), with VGG-16 (with DGAAL and MSRCSL), and with the modified VGG-14 (with DGAAL and MSRCSL), respectively.

Table 3 shows the evaluation of the effect of all the DGAAL and MSRCSL. As shown in TABLE 3, using the proposed modules, including DDC, DDLA RoI Pooling, DDC and DDLA RoI Pooling generates noticeable performance gains. For Faster R-CNN, R-FCN in Table 2, the improvement are increased slowly with the increasing of the number of DDC layers. When both DDC and DDLA RoI Pooling are used, significant accuracy improvements are obtained to Faster R-CNN, R-FCN and Mask R-CNN. As shown in Table 3, Our proposed method achieves the best accuracy with the modified VGG-14 net.

Considering the tradeoff between accuracy and speed, the proposed method achieves the best performance. The proposed method performs much better than the six models with the modified VGG-14 net. Ours outperforms Faster RCNN, R-FCN and Mask R-CNN by about 11.1, 7.9 and 7.7 percent in mAP respectively. This indicates that the significant performance improvement is from the capability of modeling geometric transformation and noise-occlusions, other than increasing model parameters.

IV. CONCLUSION

In this paper, we have proposed a novel efficient real-world ISAR object recognition and relation discovery method based on Deep Relation Graph Learning. It not only can greatly improve CNNs' sampling and transformation modeling ability, handle the complex multi-modal recognition problems, but also can leverage image datasets, region features, attributes and their high-level semantic descriptions to learn about their inter-modal associations between the features, attributes and classes with attribute-association learning. We evaluated the proposed method to recognize real-world ISAR objects with other methods on two real-world ISAR datasets, and the results show that our proposed method outperforms several state-of-the-art methods.

REFERENCES

[1] X. Zhang, Y. Zhuang, W. Wang, and W. Pedrycz, "Online feature transformation learning for cross-domain object category recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 7, pp. 2857–2871, Jul. 2018.

[2] B. Xue and N. Tong, "DIOD: Fast and efficient weakly semi-supervised deep complex ISAR object detection," *IEEE Trans. Cybern.*, to be published.

[3] W. Y. Yang, H. P. Ji, J. W. Bae, H. N. Myung, and C. H. Kim, "Automatic algorithm for estimating the jet engine blade number from the radar target signature of aircraft targets," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 30, no. 7, pp. 18–29, Jul. 2015.

[4] Q.-C. Pham and Y. Nakamura, "A new trajectory deformation algorithm based on affine transformations," *IEEE Trans. Robot.*, vol. 31, no. 4, pp. 1054–1063, Aug. 2015.

[5] S. L. Al-Khafaji, J. Zhou, and A. Zia, "Spectral-spatial scale invariant feature transform for hyperspectral images," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 837–850, Feb. 2018.

[6] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, and S. Mougiakakou, "Lung pattern classification for interstitial lung diseases using a deep convolutional neural network," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1207–1216, 2016.

[7] B. Xue, N. Tong, and X. Xu, "DIOD: Fast, semi-supervised deep ISAR object detection," *IEEE Sensors J.*, vol. 19, no. 3, pp. 1073–1081, Feb. 2019.

[8] E. Salahat and M. Qasaimeh, "Recent advances in features extraction and description algorithms: A comprehensive survey," in *Proc. IEEE Int. Conf. Ind. Technol.*, Mar. 2017, pp. 1059–1063.

[9] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 636–644.

[10] I. Berkes and M. Weber, "On series of dilated functions," *Quart. J. Math.*, vol. 65, no. 1, pp. 25–52, 2016.

[11] K. M. Digumarti, A. T. Conn, and J. Rossiter, "Euglenoid-inspired giant shape change for highly deformable soft robots," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2302–2307, Oct. 2017.

[12] J. Li, H.-C. Wong, S.-L. Lo, and Y. Xin, "Multiple object detection by a deformable part-based model and an R-CNN," *IEEE Signal Process. Lett.*, vol. 25, no. 2, pp. 288–292, Feb. 2018.

[13] Y. Qin, S. He, Y. Zhao, and Y. Gong, "RoI pooling based fast multi-domain convolutional neural networks for visual tracking," in *Proc. Int. Conf. Artif. Intell. Ind. Eng.*, 2016, pp. 1–5.

[14] R. Meymandpour and J. G. Davis, "A semantic similarity measure for linked data: An information content-based approach," *Knowl.-Based Syst.*, vol. 109, pp. 276–293, Oct. 2016.

[15] D. C. Alexander, D. Zikic, J. Zhang, H. Zhang, and A. Criminisi, "Image quality transfer via random forest regression: Applications in diffusion MRI," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2014, vol. 17, no. 3, pp. 225–232.

[16] B. Xue, N. Tong, X. Xu, and X. He, "Dynamical rain attenuation short-term prediction in the w band based on a nonstationary time series ARIMA model," *IEEE Antennas Propag. Mag.*, to be published.

[17] P. Yang, P. Zhao, and X. Gao, "Robust online multi-task learning with correlative and personalized structures," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 11, pp. 2510–2521, Nov. 2017.

[18] X. Huang, J. Li, and X. Hu, "Accelerated attributed network embedding," in *Proc. SIAM Int. Conf. Data Mining*, 2017, pp. 633–641.

[19] G. Zhu and C. A. Iglesias, "Sematch: Semantic similarity framework for knowledge graphs," *Knowl.-Based Syst.*, vol. 130, pp. 30–32, Aug. 2017.

[20] J. Choi and R. A. Rutenbar, "Video-rate stereo matching using Markov random field TRW-S inference on a hybrid CPU+FPGA computing platform," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 2, pp. 385–398, Feb. 2016.

[21] E. Eftelioglu, "Geometric median," in *Encyclopedia of GIS*. 2015.

[22] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 675–678.

[23] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2016.

[24] 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.

[25] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 379–387.

[26] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[27] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

BIN XUE was born in Shangluo, Shaanxi, in 1990. He received the B.S. degree in computer science and technology from the Zhongnan University of Economics and Law, in 2013, and the M.S. degree in information and communication engineering from Airforce Engineering University, Xi'an, China, in 2015, where he is currently pursuing the Ph.D. degree in electronic science and technology.

His research interests include ISAR object recognition, deep learning, time series prediction, modern statistical analysis, and machine learning.

NINGNING TONG received the B.S., M.S., and Ph.D. degrees from Air Force Engineering University, Xi'an, China, in 1984, 1988, and 2009, respectively, where she is currently a Professor. She has authored or co-authored over 60 research papers and four books. Her current research interests include wireless communication, radar signal processing, and electronic countermeasures.