An Attention-Enhanced End-to-End Discriminative Network With Multiscale Feature Learning for Remote Sensing Image Retrieval

Dongyang Hou, Siyuan Wang, Xueqing Tian, and Huaqiao Xing

Abstract—The discriminative ability of image features plays a decisive role in content-based remote sensing image retrieval (CBRSIR). However, the widely-used convolutional neural networks cannot focus on the discriminative features of important scenes, resulting in unsatisfactory retrieval performance in complex contexts. In this article, an attention-enhanced end-to-end discriminative network with multiscale learning for CBRSIR is proposed to solve this issue. First, a multiscale dilated convolution module is embedded into some of ResNet50’s residual blocks to increase the perceptual field and capture the multiscale features of remote sensing image scenes. Then, a lightweight and efficient triplet attention module is added behind each residual block to capture the salient features of remote sensing images and establish the interdimensional dependencies using residual transform. In addition, the end-to-end training approach is performed using an online label smoothing loss to reduce the intraclass variance of features and enhance interclass differentiability. Experimental results on four publicly available remote sensing image datasets show that our network achieves state-of-the-art or competitive performance, especially on complex scene dataset UCMD with an average retrieval precision improvement of 3.23% to 29.35% compared to other new methods.

Index Terms—Attention mechanism, convolutional neural networks (CNNs), multiscale features, remote sensing image (RSI) retrieval.

I. INTRODUCTION

WITH the development of earth observation technology, the number of remote sensing images (RSIs) has increased exponentially [1]. Therefore, rapid and accurate search for scenes of interest from large RSI archives has been one of the keys to constrain the sharing and effective utilization of RSIs [2], [3].

At present, content-based remote sensing image retrieval (CBRSIR) can mine visually or semantically similar scenes from RSI archives by comparing the similarity of image features [4]. Their retrieval performance depends largely on the effectiveness and accuracy of feature extraction, which fully characterize the information of RSI content [5]. Previous feature extraction methods mainly focus on handcrafted low and mid-level RSI features (i.e., color and texture) [6], [7]. These methods are low-complexity and computationally simple, but they are highly subjective and have a direct semantic gap with the real image features [8].

In recent years, convolutional neural networks (CNNs) designed to simulate human cognitive processes have been used to extract high-level semantic information from RSIs, and have achieved encouraging retrieval results in CBRSIR [9]. However, CNNs perform poorly in some complex scenes with varying scales [4]. This is mainly because that CNNs focus on global features and have difficulty in accurately capturing features of important objects that occupy a relatively small portion in complex scenes [10], [11]. Therefore, attention mechanisms (i.e., DANet and residual attention) that focus on the important objects are often used to improve this problem, and have achieved better retrieval performance [11], [12], [13]. However, these attention modules introduce many external parameters, which increase the computational complexity. Besides, current attention-based methods take less account of the scale factor, which is also an important factor in complex scenes and affects the retrieval accuracy [14].

To address the abovementioned issues, a new attention-enhanced end-to-end discriminative network with multiscale learning is proposed for CBRSIR in this article. We introduce a lightweight and efficient triplet attention [15] to capture distinguishing features of important objects and use the residual transform to establish interdimensional dependencies. Before the attention module, a multiscale dilated convolution module is used to increase the reception field and capture the multiscale features. In addition, the network replaces the regular cross-entropy loss function with an online label smoothing loss function [16] to reduce intra-class differences and enhance interclass differentiability. In general, our network contains the following three important parts: triplet attention, dilated convolutions with multiscale, and online label smoothing loss function, so our network is simply called TDO-Net.

The main contributions and innovation of our work is to redesign the baseline ResNet50 network using three new modules of triplet attention, hybrid dilated convolutions, and online label...
smoothing loss function, making the novel network better suited for image retrieval of complex scenes. The experimental results on four public datasets indicate that our TDO-Net achieves better retrieval performance in complex scenes with fewer number of parameters and floating point of operations (FLOPs), compared to other state-of-the-art networks.

The rest of this article is organized as follows. Section II reviews related work about low and mid-level features for CBRSIR, CNN-based high-level features for CBRSIR, and attention mechanism. Section III outlines our TDO-Net, including framework of TDO-Net, feature extraction module with multiscale dilated convolution and triplet attention, and online label smoothing loss function. Experimental results are presented in Section IV. Finally, Section V concludes this article.

II. RELATED WORK

A. Low and Mid-Level Features for CBRSIR

Previously CBRSIR methods mainly relied on some handcrafted global image features such as color, edge, texture, and so on [17], [18], [19]. For example, Martins et al. [20] constructed a Gabor texture-based neural classifier for CBRSIR. Yao et al. [21] applied Gabor transformations with different scales and orientations for CBRSIR. Because a single low-level feature has limited ability to describe RSIs, some scholars have tried to combine multiple low-level features to enhance the retrieval performance [22]. For example, Maheshwary and Srivastava [22] designed a prototype CBRSIR system based on color moment and grayscale co-occurrence matrix. These basic low-level features facilitated the development of early RSI retrieval, but the retrieval accuracy of RSIs has been low because these features are easily affected by lighting conditions, occlusion, truncation, and other factors. Therefore, the focus of later research gradually shifted to the mid-level features represented by local feature aggregation, such as bag of visual words (BoVW) and vector of local aggregated descriptors (VLAD) [23], [24]. Compared with low-level features, the retrieval performance based on mid-level features has been greatly improved. However, because it does not consider the spatial association of local features and still relies on manual feature selection, there is still a high room for improvement in its retrieval performance and applicability [25], [26].

B. CNN-Based High-Level Features for CBRSIR

As the most representative deep learning algorithm, CNN has achieved an excellent performance in the field of computer vision [27]. It uses a multilayer network structure combining convolutional layers and pooling layers to extract features step by step [28]. These extracted high-level semantic features have been widely applied to CBRSIR. For example, Zhou et al. [3] retrained the mainstream CNN models on remote sensing benchmark dataset, and showed that their retrieval results were significantly better than the traditional low/mid-level features. Since there are significant differences between remotely sensed images and ordinary natural images in terms of camera angle and scale dependence, some scholars have tried to improve the basic CNN model by combining the characteristics of RSIs. For example, Fan et al. [29] proposed a new distribution consistency loss function for CBRSIR to solve the problem of nonuniformity in the distribution of RSIs sample data. Liu et al. [30] proposed a slice feature depth hashing algorithm for CBRSIR based on the smaller interclass distance feature of RSIs. The abovementioned methods have improved retrieval performance to some extent, but the retrieval performance in complex scenes is still poor because complex scenes often contain a large amount of redundant and complex background information [31], while CNNs mainly focus on global features and ignore local features [10].

C. CBRSIR Based on Attention Mechanisms

The visual attention mechanism is a resource allocation mechanism inspired by human perception, which focuses on the important region in the global image and then devotes more attention resources to this region [32]. In other words, the attention mechanism can suppress useless background information and improve visual information processing [33]. Therefore, the attention mechanism is often used to improve the retrieval performance of CNNs in complex scenes. For example, Imbriaco et al. [34] explored different attention mechanisms to select the most relevant features for CBRSIR. Xiong et al. [13] used different attention modules to focus on important features in spatial and channel dimensions for CBRSIR, respectively. Wang et al. [35] designed an integrated channel and spatial attention mechanism to extract globally consistent features of foreground objects for CBRSIR. Although these methods bring performance gains in complex scenes, they also introduce many external parameters that increase the complexity of the network. Besides, these methods combine spatial attention and channel attention in parallel or in series, respectively. However, the two types of attention in the human brain tend to work in tandem [36]. In addition, the embedding of multiple attention modules tends to reduce the stability of the model. Therefore, we choose a triplet attention to focus on important regions, and encode both channel and spatial information with negligible computational overhead.

III. METHODOLOGY

A. TDO-Net Framework

To solve the abovementioned problems and further improve the retrieval accuracy of complex scenes, we design an attention-enhanced end-to-end discriminative network TDO-Net. In our TDO-Net, the ResNet50 [37] based on a stack of residual structures with powerful feature learning capability and easy optimization, was chosen as the backbone network. The ResNet50 network structure consists of a single convolutional layer, four residual structures, an average pooling layer, and a fully connected layer. In the baseline ResNet50, the fully connected layer tends to produce global features of the image, while the convolutional layer describes local features [8]. Therefore, in our TDO-Net, the third and fourth residual structures are designed as multiscale feature extraction modules, and a triplet attention module is embedded after the last convolutional layer of each residual block. Finally, end-to-end training is performed using online label smoothing losses. In the testing phase, the output of the last average pooling layer of the feature extraction network
is selected as the image features for CBRSIR. The framework of our TDO-Net is shown in Fig. 1.

As can be seen from Fig. 1, unlike previous networks, our TDO-Net not only focuses on the feature extraction of important region, but also considers the scale factors affecting the feature representation. The multiscale feature learning module is composed of hybrid dilated convolution. That is, the dilated convolution with different expansion rates is assigned in the continuous residual structure to make it adaptively capture scene information at different scales. Besides, online label smoothing loss is used to replace the regular cross-entropy loss function. This is because online label smoothing loss can constrain the boundaries of different classes, enabling the network to discriminate visually similar but semantically irrelevant images more effectively.

B. Feature Extraction Network

1) Multiscale Feature Learning Module: In the convolutional operation, the reception field is used to represent the size of the perceptual range of different neurons inside the CNN for the original image. The size of the reception field of the standard convolution is consistent with its convolution kernel size, and the dilated convolution injects “zeros” into the standard convolutional map, making it possible to increase the reception field of the convolution with the same parameters and computational effort [38]. Therefore, output of the dilated convolution contains a larger range of feature information than the original convolution, which is beneficial to obtain global information of RSIs [39], [40].

The dilated rate of the convolution defines the spacing of the convolution kernel sampling, which determines the size of the reception field. Fig. 2 shows the reception field at different dilated rates, and the reception field increases as the dilated rate increases. However, dilated convolution increases the information reception field while the convolution space becomes discontinuous, which can cause the loss of local information and the lack of dependency among neighbouring pixels, resulting in inconsistent local information. Therefore, for RSIs with diverse scene scales, a network using dilated convolutions with only a fixed dilated rate is highly prone to information loss of small objects.

To address the shortcomings of dilated convolution, a multiscale feature learning module based on hybrid dilated convolution is designed in our TDO-Net, which aims to enable the network to flexibly capture multiscale contextual information of RSIs. Specifically, to ensure the continuity of image information, it is stipulated that the dilated rate of the superimposed convolution layer is specified to be only divisible by 1 and itself. After the pooling operation, the allocation of the dilated rate follows a sawtooth-like heuristic structure. As shown in Fig. 3, the dilated rate is designed as a rising group of [1], [2], [5], [9] in a residual
structure of our TDO-Net, which can adaptively extract feature information of different sizes. In this hybrid dilated convolution structure, the convolution with smaller dilated rate is used to capture the near feature information, and the convolution with larger dilated rate is used to capture the long-distance information, and the topmost convolution layer acquires information from a larger spatial range without causing sampling discontinuity in the case of constant perceptual field.

2) Attention-Enhanced Module: RSIs contain a large amount of background information, which affects the high-level feature discriminative ability of CNNs, while the attention mechanism focusing on important regions of images is often used to compensate for this deficiency of CNNs. However, currently widely used attention obtains better performance at the cost of higher model complexity. While these attention-based networks bring performance gains, they introduce many external parameters that increase the complexity of the network. In addition, the stacked use of multiple attention modules reduces the stability of the network. As an efficient attention module with almost no parameters, triplet attention uses residual transformations to establish interdimensional dependencies, encodes interchannel and spatial information with negligible computational overhead, and models both channel attention and spatial attention weights. Therefore, triplet attention is chosen as the basic component of the attention-enhanced module in our TDO-Net to highlight features of the important region and suppress cluttered background. Specifically, the triplet attention is embedded into each residual block of ResNet50 in our TDO-Net.

The structure of triplet attention is shown in Fig. 4. Specifically, given an input feature map $X \in \mathbb{R}^{H \times W \times C}$, whose size is $H \times W \times C$ (H: Height, W: Width, C: Number of channels). The first branch is the spatial attention computation branch, where the input features are first subjected to channel pooling, i.e., Z-Pool, which reduces the zeroth dimension of $X$ to two. The pooling process can be expressed as

$$Z - \text{pool}(X) = [\text{MaxPool}_{0d}(X), \text{AvgPool}_{0d}(X)]$$  \hspace{1cm} (1)$$

where $0d$ denotes the 0th dimension where the maximum pooling and average pooling operations are performed. In this way, the input feature map $X$ changes its shape to $2 \times H \times W$ after Z-Pool, and the feature depth and computational effort are effectively reduced. After that, the input features are passed through the convolution and batch normalization layers in turn, and then the spatial attention weights are generated by the sigmoid activation function. The second branch is the channel $C$ and spatial $W$ dimensional interaction capture branch, where the input features $X$ are first transposed into $H \times C \times W$ dimensional features, followed by pooling in the $H$ dimension, $7 \times 7$ convolution and sigmoid activation function, and finally transposed into a $C \times H \times W$ feature. The third branch is the channel $C$ and space $H$ dimensions interaction capture branch, and the specific operation is basically the same as the second branch. Finally, the final attention-based features are obtained by feature aggregation of the features extracted from the three branches using average pooling.

C. Online Label Smoothing Loss

Compared with natural images, the complex and diverse backgrounds of RSIs tend to cause greater intraclass variation and higher interclass similarity. This leads to the problem of large intraclass gaps and unclear interclass boundaries among the extracted high-level features. In order to solve the abovementioned problems, it is necessary to increase the interclass separability and intraclass tightness during the training process, so that similar images are divided into more compact clusters. However, the most widely used cross-entropy loss obtains interclass distances that are tiny and susceptible to noisy labels [41]. To address this problem, Zhang et al. [16] proposed an online label smoothing loss and it has been shown to be effective in the classification of natural pictures. Therefore, we try to introduce it into our CBRSIR network TDO-Net for further differentiating interclass differences and narrowing intraclass differences in RSIs.

In the online label smoothing loss, different smoothing thresholds are given for different categories, and the smoothing weight matrix is dynamically generated during the training process to constrain the variance distance for images of different categories. The specific formula is shown as follows [16]:

$$L_{\text{hard}} = - \sum_{k=1}^{K} q(k | x_i) \cdot \log p(k | x_i)$$ \hspace{1cm} (2)

$$L_{\text{soft}} = - \sum_{k=1}^{K} S_{y_{i,k}}^{t-1} \cdot \log p(k | x_i)$$ \hspace{1cm} (3)

$$S_{y_{i,k}}^{t-1} = S_{y_{i,k}}^{t-1} + p(k | x_i)$$ \hspace{1cm} (4)

$$S_{y_{i,k}}^{t} = \frac{SS_{y_{i,k}}^{t-1}}{\sum_{l=1}^{K} S_{y_{i,l}}^{t-1}}$$ \hspace{1cm} (5)

where $L_{\text{hard}}$ is the cross-entropy loss, $x_i$ denotes the input image, $y_i$ denotes the true category of the input image, $k$ is the predicted category of the input image, $K$ is the total number of image categories, $p(k | x_i)$ denotes the probability that the input image $x_i$ is predicted to be category $k$, and $q$ denotes the
distribution of \( y_i \), \( q(k = y_i \mid x_i) = 1 \) and \( q(k \neq y_i \mid x_i) = 0 \). \( L_{\text{soft}} \) is the online label smoothing loss, \( t \) is the number of training iterations, \( S_{y_i,k}^{t-1} \) is the weight of label smoothing. At the end of one iteration process, the weights are updated according to (4) and (5) to obtain the smoothing threshold \( S_{y_i,k}^t \) for the next iteration.

Initially, the parameters of the online label smoothing loss are initialized according to the following equation:

\[
S_{y_i,k}^0 = \begin{cases} 
\mu & \text{if } k = c \\
\frac{1-\mu}{K-1} & \text{if } k \neq c 
\end{cases}
\] (6)

where \( c \) is the correct category of the input image, \( \mu \) is the initial smoothing parameter, generally set between 0.9 and 1.0. Due to the lack of hard labels, the model is difficult to converge. Therefore, cross-entropy loss and online label smoothing are used to jointly constrain the training of the model. The total training loss can be expressed as (7), where \( \alpha \) is used to balance the cross-entropy loss and the online label smoothing loss

\[
L = \alpha L_{\text{hard}} + (1 - \alpha) L_{\text{soft}}. \tag{7}
\]

### IV. Experiments and Results

To verify the effectiveness of our TDO-Net, three types of experiments are carried out on four RSI benchmark datasets. All experiments are conducted on Ubuntu 20 OS with an Intel 3.7 GHz i9-10900K processor and an NVIDIA GeForce RTX3090 graphics card. Our TDO-Net and other comparative networks are implemented based on the PyTorch library in Python language. To ensure the fairness of the performance evaluation, all experiments are conducted in the same environment.

#### A. Experimental Setup

1) Datasets: Four publicly available RSI datasets of UCMD [42], NWPU [43], PatternNet [3], and VArcGIS [44], are selected as validation datasets. Table I shows the details of the four datasets.

Among these four datasets, the UCMD and NWPU datasets are originally designed for RSI scene classification and some images contain a large amount of background information [45]. This redundant background information poses a challenge for accurate and reliable similar image search [3]. The remaining datasets with different sources and the same classification systems are designed for CBRSIR and contain less background information. That is, the abovementioned four datasets can be divided into two categories of complex scenes (UCMD and NWPU) and simple scenes (PatternNet and VArcGIS). This diversity of datasets can improve the credibility of the assessment results. Figs. 5–8 give some example images of each category in the four datasets.

2) Implementation Details: In the experiments, the ResNet50 network without pretrained weights was chosen as the baseline model. In the training phase, the training batch is 40 epochs, the batch size is 32, the optimizer is Adam, the initial learning rate is 3e–4, and the weight decay is 3e–4. In all experiments, the size of the input image has been resized to 224 x 224 pixels. To make the model converge more efficiently, we set the weight value of balanced cross-entropy loss and online label smoothing loss to 0.4.

For the benchmark datasets, we randomly divide the images of each category into training set and test set in the ratio of 8:2. Besides, the training set is redivided into two parts, with 80% of the images used for training and the remaining 20% for validation.

In the testing phase, the Euclidean distance is used to measure the similarity of features. The closer the distance between the
Fig. 7. Example images of the PatternNet dataset. (1) Airplane. (2) Baseball field. (3) Basketball court. (4) Beach. (5) Bridge. (6) Cemetery. (7) Chaparral. (8) Christmas tree farm. (9) Closed road. (10) Coastal mansion. (11) Crosswalk. (12) Dense residential. (13) Ferry terminal. (14) Football field. (15) Forest. (16) Freeway. (17) Golf course. (18) Harbor. (19) Intersection. (20) Mobile home park. (21) Nursing home. (22) Oil gas field. (23) Oil well. (24) Overpass. (25) Parking lot. (26) Parking space. (27) Railway. (28) River. (29) Runway. (30) Runway marking. (31) Shipping yard. (32) Solar panel. (33) Sparse residential. (34) Storage tank. (35) Swimming pool. (36) Tennis court. (37) Transformer station. (38) Wastewater treatment plant.

Fig. 8. Example images of the V ArcGIS dataset. (1) Airplane. (2) Baseball field. (3) Basketball court. (4) beach. (5) Bridge. (6) Cemetery. (7) Chaparral. (8) Christmas tree farm. (9) Closed road. (10) Coastal mansion. (11) Crosswalk. (12) Dense residential. (13) Ferry terminal. (14) Football field. (15) Forest. (16) Freeway. (17) Golf course. (18) Harbor. (19) Intersection. (20) Mobile home park. (21) Nursing home. (22) Oil gas field. (23) Oil well. (24) Overpass. (25) Parking lot. (26) Parking space. (27) railway. (28) River. (29) Runway. (30) Runway marking. (31) Shipping yard. (32) Solar panel. (33) Sparse residential. (34) Storage tank. (35) Swimming pool. (36) Tennis court. (37) Transformer station. (38) Wastewater treatment plant.

Fig. 9. Retrieval results of the top five images on the NWPU dataset. The incorrect and correct results are tagged in red and green, respectively. It can be seen that the networks of ResNet50 with one of MFE, Triplet, or OLS modules outperform the baseline network ResNet50 on the two datasets. However, on the complex scene dataset UCMD, the mAP of the networks with a single module improves only 3.16% to 13.13% compared to the baseline network ResNet50, while our TDO-Net improves the mAP by 24.46%. That is, our TDO-Net still improves the mAP by 11.33% to 21.3% compared to these networks with a single-module. On the simple scene dataset V ArcGIS, our TDO-Net still achieves an improvement of 2.07% to 9.67% compared to the abovementioned networks. These experimental results show that our TDO-Net effectively combines the three modules of triplet attention, hybrid dilated convolutions, and online label smoothing loss function to significantly improve the image retrieval performance under scene complexity.

B. Experimental Results

1) Ablation Experiments: To verify the effectiveness of our proposed method, we conducted ablation experiments using the baseline ResNet50 network combining different modules. It should be noted that the abbreviation MFE stands for multiscale feature extraction module, Triplet stands for attention-enhanced module and OLS stands for online label smoothing loss. These ablation experiments are conducted on one complex scene dataset UCMD and one simple scene dataset V ArcGIS. Table II shows the retrieval performances of these methods on the two datasets.

It can be seen that the networks of ResNet50 with one of MFE, Triplet, or OLS modules outperform the baseline network ResNet50 on the two datasets. However, on the complex scene dataset UCMD, the mAP of the networks with a single module improves only 3.16% to 13.13% compared to the baseline network ResNet50, while our TDO-Net improves the mAP by 24.46%. That is, our TDO-Net still improves the mAP by 11.33% to 21.3% compared to these networks with a single-module. On the simple scene dataset V ArcGIS, our TDO-Net still achieves an improvement of 2.07% to 9.67% compared to the abovementioned networks. These experimental results show that our TDO-Net effectively combines the three modules of triplet attention, hybrid dilated convolutions, and online label smoothing loss function to significantly improve the image retrieval performance under scene complexity.

2) Comparisons With the Baseline ResNet50: To further validate the effectiveness of our TDO-Net, we comprehensively evaluated the performance difference between the TDO-Net and the baseline ResNet50 on four benchmark datasets. Table III shows their retrieval performances on the four datasets.

It can be seen from Table III that our TDO-Net improves the mAP by 24.46% and 33.84% on the complex scene datasets UCMD and NWPU, and by 6.57% and 9.67% on the simple scene datasets PatternNet and V ArcGIS. The trends of other evaluation metrics are consistent with the abovementioned changes. The experimental results show that our TDO-Net can improve the performance of CBRSIR, especially for complex scenes. In addition to these quantitative comparisons, we also perform some visual comparisons. Fig. 9 shows the top five images of the retrieval results on the NWPU dataset. The incorrect and correct results are tagged in red and green, respectively. It can be seen that our TDO-Net can return more correct results. For example, for the retrieval of stadium and sparse building images, the baseline method incorrectly retrieves semantically irrelevant images, such as golf courses and dense houses, while all images retrieved by our network are correct. This is mainly because our network has more and stronger feature discrimination ability.
To clearly demonstrate the impact on different categories, the retrieval performances of each category on the four datasets are shown in Figs. 10–13. It can be seen that the performance improvement for different categories is irregular, but there is an overall trend of improvement. The retrieval improvement of our TDO-Net is significant in the categories with complex or redundant background (e.g., airplane, intersection, cemetery, sparse residential, storage tank, etc.).

### Comparisons With Other Methods
To further validate the performance of our TDO-Net, six newly proposed methods of discriminative feature learning (DFL) [13], Attention boosted bilinear pooling (ABP) [35], tree-triplet-classification network (TT-C) [48], context attended graph convolutional network (GCN), and context attended Siamese graph convolutional network (SGCN) [10], and deep feature learning with latent relationship embedding [49] are selected as the comparison methods. All the

---

**TABLE II**

| Dataset  | Method               | ANMRR | mAP  | P5   | P10  | P20  | P50  | P100 | P1000 |
|----------|----------------------|-------|------|------|------|------|------|------|-------|
| UCMD     | ResNet50             | 0.3456| 0.6091| 0.7438| 0.7262| 0.6981| 0.6483| 0.5036| 0.0783|
|          | MFE                  | 0.3148| 0.6407| 0.7971| 0.7764| 0.7506| 0.6848| 0.5260| 0.0780|
|          | Triplet              | 0.2835| 0.6776| 0.7462| 0.7400| 0.7292| 0.7010| 0.5607| 0.0780|
|          | OLS                  | 0.2992| 0.6586| 0.7538| 0.7448| 0.7273| 0.6907| 0.5422| 0.0781|
|          | MFE+Triplet          | 0.2438| 0.7192| 0.8376| 0.8245| 0.8050| 0.7542| 0.5854| 0.0793|
|          | MFE+OLS              | 0.2240| 0.7404| 0.8557| 0.8388| 0.8221| 0.7782| 0.6021| 0.0791|
|          | Triplet+OLS          | 0.3117| 0.6450| 0.7338| 0.7281| 0.7114| 0.6711| 0.5356| 0.0790|
|          | Our TDO-Net (MFE+Triplet+OLS) | **0.1254** | **0.8537** | **0.8948** | **0.8936** | **0.8683** | **0.8671** | **0.6903** | **0.0794** |

The larger the mAP and Pk, and the smaller the ANMRR indicate the performance is better. The best results are marked in bold.

**TABLE III**

| Dataset  | Method               | ANMRR | mAP  | P5   | P10  | P20  | P50  | P100 | P1000 |
|----------|----------------------|-------|------|------|------|------|------|------|-------|
| UCMD     | ResNet50             | 0.3456| 0.6091| 0.7438| 0.7262| 0.6981| 0.6483| 0.5036| 0.0783|
|          | MFE                  | 0.3148| 0.6407| 0.7971| 0.7764| 0.7506| 0.6848| 0.5260| 0.0780|
|          | Triplet              | 0.2835| 0.6776| 0.7462| 0.7400| 0.7292| 0.7010| 0.5607| 0.0780|
|          | OLS                  | 0.2992| 0.6586| 0.7538| 0.7448| 0.7273| 0.6907| 0.5422| 0.0781|
|          | MFE+Triplet          | 0.2438| 0.7192| 0.8376| 0.8245| 0.8050| 0.7542| 0.5854| 0.0793|
|          | MFE+OLS              | 0.2240| 0.7404| 0.8557| 0.8388| 0.8221| 0.7782| 0.6021| 0.0791|
|          | Triplet+OLS          | 0.3117| 0.6450| 0.7338| 0.7281| 0.7114| 0.6711| 0.5356| 0.0790|
|          | Our TDO-Net (MFE+Triplet+OLS) | **0.177** | **0.9773** | **0.9797** | **0.9791** | **0.9785** | **0.9779** | **0.9776** | **0.9754** |

The larger the mAP and Pk, and the smaller the ANMRR indicate the performance is better. The best results are marked in bold.

---

**Fig. 10.** mAP for each category in the UCMD dataset.
abovementioned methods use a pre-trained network (PT_Net) as the backbone network, so we also use PT_ResNet50 to train our TDO-Net on the UCMD dataset (PT_TDO-Net). In addition, seven low/mid-level feature-based method of color moment, color histogram, Gabor texture, GLCM texture, GIST, BOVW, and VLAD are also selected as comparisons. The retrieval performances of the 15 different methods are shown in Table IV.

It can be seen from Table IV that the high-level feature-based method significantly outperform the low/mid-level feature-based methods in CBRSIR. It also can be seen that our TDO-Net without pre-trained ResNet50 achieves better retrieval performance than the four PT_Net DFL, T-T-C, CA-GCN, and CA-SGCN, with an mAP improvement of 1.37%, 13.2%, 18.6%, and 12.15%, respectively. After adding pre-trained ResNet50, our network outperforms all other pre-trained methods, with an mAP improvement of 3.23% to 29.35%. The abovementioned experimental results demonstrate that our TDO-Net achieves state-of-the-art or competitive results.

To further validate the efficiency of our TDO-Net, the number of parameters and FLOPs are selected as evaluation metrics. Their smaller values indicate better efficiency. Some baselines models (i.e., PT_AlexNet [50], PT_VGG16 [51], and PT_ResNet50 [37]) and similar new multiple attention fusion-based models (i.e., PT_CEL [4], PT_ABP [35], and PT_W-CAN [11]) are chosen as comparison models. Experimental results are shown in Table V.

It is apparent from this table that compared to the baseline network PT_ResNet50, our PT_TDO-Net increases the mAP by 17.6% with only 0.02% increase in parameters and 1.12% increase in FLOPs. Besides, it also can be seen from the table that compared to multiple attention fusion-based methods PT_W-CAN and PT_CEL, our network has improved retrieval performance with 78.3% and 58.6% reduction in the number of parameters, and 51.9% and 43.3% decrease in FLOPs, respectively. Although the FLOPs of the PT_ABP method is slightly better than our model, our mAP is 5.56% higher than
its mAP. This is because hybrid dilated convolution does not change the number of parameters, while the rotation operation and residual transformation of triplet attention can reduce part of the convolution computation process. The abovementioned experimental results show that our method improves the retrieval performance but not at the cost of reducing the efficiency.

Table IV

| Method       | ANMRR  | mAP   | P5   | P10  | P20  | P50  | P100 | P1000 |
|--------------|--------|-------|------|------|------|------|------|-------|
| Color Moment | 0.7616 | 0.1859| 0.4933| 0.4029| 0.3264| 0.2386| 0.1733| 0.0630|
| Color Histogram | 0.7877 | 0.1610| 0.4686| 0.3869| 0.3114| 0.2119| 0.1538| 0.0575|
| Gabor Texture | 0.8085 | 0.1381| 0.4152| 0.3398| 0.2668| 0.1840| 0.1390| 0.0569|
| GLCM Texture | 0.7947 | 0.1481| 0.4143| 0.3407| 0.2724| 0.1962| 0.1500| 0.0611|
| GIST         | 0.7509 | 0.2091| 0.4257| 0.3690| 0.3074| 0.2398| 0.1825| 0.0646|
| BoVW         | 0.6690 | 0.2581| 0.4971| 0.4617| 0.4089| 0.3260| 0.2440| 0.0647|
| VLAD         | 0.6733 | 0.2542| 0.4871| 0.4526| 0.4101| 0.3204| 0.2408| 0.0647|
| **PT_DFL**   | 0.089  | 0.840 | 0.919 | 0.914 | 0.921 | 0.845 | -    | -    |
| **PT_ABP**   | -      | 0.9056| 0.8638| 0.8367 | 0.7227| 0.5939| -    | -    |
| **PT_T-T-C** | 0.2135 | 0.7217| 0.9435| 0.9081 | 0.7852| 0.6771| -    | -    |
| **PT_CA-GCN**| 0.31   | 0.6677| -     | 0.9008 | -    | -    | -    | -    |
| **PT_CA-SGCN**| 0.29  | 0.7322| -     | 0.9462 | -    | -    | -    | -    |
| **PT_DFLR**  | 0.0483 | 0.9289| 0.9547| 0.9238 | 0.8911| 0.4723| 0.0500| -    |
| **Our TDO-Net** | 0.1254 | 0.8537| 0.8948| 0.8936 | 0.8863| 0.8671| 0.6903| 0.0794|
| **Our PT_TDO-Net** | **0.0328** | **0.9612** | **0.9652** | **0.9652** | **0.9638** | **0.9607** | **0.7734** | **0.0799** |

The larger the mAP and Pk, and the smaller the ANMRR indicate the performance is better. The best results are marked in bold.

Table V

| Method         | mAP    | Parameters(M) | FLOPs(G) |
|----------------|--------|---------------|----------|
| PT_AlexNet     | 0.7235 | 61.101        | 7.147    |
| PT_VGG16       | 0.7301 | 138.358       | 15.486   |
| PT_ResNet50    | 0.7852 | 25.557        | 4.112    |
| PT_CEL         | 0.9424 | 61.678        | 7.355    |
| PT_ABP         | 0.9056 | 28.074        | 4.123    |
| PT_W-CAN       | 0.9510 | 117.85        | 8.638    |
| **Our PT_TDO-Net** | **0.9612** | **25.562**    | **4.158** |

4) Visual Interpretation of the Network: In order to better understand the superiority of our TDO-Net, the output feature heatmap of our TDO-Net is visualized. Fig. 14 shows the feature heatmap of some example images in the UCMD dataset using the Grad-CAM++ [52] tool. It is worth noting that the more the color leans towards red, the more sensitive the model is to the pixel value at that location, i.e., the higher the level of attention.

As shown in Fig. 14, the focus of the heatmap generated by ResNet50 is generally inaccurate or even mislocated. For example, on the image in the fourth column, the heatmap is incorrectly positioned to cover the background area, while the heatmap of our TDO-Net completely covers the building. In the second row and second column, the heat map focus of the ResNet50 method is broad, while the heatmap of our TDO-Net better covers the important objects with higher level of detail. The results indicate that our TDO-Net can better focus on important scenes with higher discriminative power, which is especially important for improving the performance of image retrieval in complex scenes.

5) Parameter Sensitivity Analysis: The setting of hyperparameters has a significant impact on retrieval performance. The loss function weight in (7) and the dilated rate of the hybrid dilated convolution are the important hyperparameters of our proposed TDO-Net. Therefore, to verify the effect of hyperparameters, we do some additional experiments on the UCMD dataset. In these experiments, to minimize the randomness of the trained model, we repeat each set of experiments 10 times and randomly divide the training and validation sets for each group.

Table VI represents the experimental results on the sensitivity of the loss function weight. It can be seen from the table that the retrieval performance increases and then decreases as the weight increases, and the best performance is achieved at a weight of 0.4. The lowest retrieval performance is achieved when only the
cross-entropy loss function ($\alpha = 1$) is available due to the small distance between feature classes. Therefore, we choose 0.4 as the loss fusion weight in the abovementioned experiments.

To demonstrate the power of hybrid dilated convolution in TDO-Net, we also experimentally analyze the effect of convolution kernels with different dilated rates. Specifically, we compare the dilated convolution setups studied by CEL [4], HDCFE-Net [53], and LDFN [54]. The experimental results for different dilated rates are shown in Table VII. It can be found that the dilated convolution is better than the standard convolution. Furthermore, the hybrid dilated convolution of [1], [2], [5], [9] used in our TDO-Net has superiority compared to others.

## V. CONCLUSION

In this article, a novel attention-enhanced end-to-end discriminative network with multiscale learning is proposed for CBRSIR. In this network, hybrid dilated convolutions with different dilated rates are employed to replace the regular convolution in the last two residual blocks to obtain the multiscale features of RSI scenes. Besides, a triplet attention module is embedded to each residual block to learn more discriminative spatial and channel information adaptively through cross-dimensional interaction of spatial and channel. Finally, online label smoothing loss is used for end-to-end training to reduce intraclass variance and enhance interclass differentiability of features.

Extensive experiments are conducted on four benchmark datasets to evaluate the effectiveness of our network. The experimental results indicate that our network achieves the best retrieval performance, especially in complex scenes, compared to the baseline ResNet50 and other recently proposed networks.

Our future work will be concentrated on the effect of noisy samples on attention mechanisms and multisource RSI retrieval.

Acknowledgment

The authors would like to thank the anonymous reviewers for their comments to improve this article. The authors also would like to thank the UCMD, NWPU, and PatternNet datasets for their open access.

## REFERENCES

[1] Z. Shao, W. Zhou, X. Deng, M. Zhang, and Q. Cheng, “Multilabel remote sensing image retrieval based on fully convolutional network,” IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens., vol. 13, pp. 318–328, 2020.

[2] W. Xiong, Y. Lv, X. Zhang, and Y. Cui, “Learning to translate for cross-source remote sensing image retrieval,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 7, pp. 4860–4874, Jul. 2020.

[3] W. Zhou et al., “PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval,” ISPRS J. Photogramm. Remote Sens., vol. 145, pp. 197–209, Nov. 2018.

[4] S. Wang, D. Hou, and X. Xing, “A novel multi-attention fusion network with dilated convolution and label smoothing for remote sensing image retrieval,” Int. J. Remote Sens., vol. 43, no. 4, pp. 1306–1322, Mar. 2022.

[5] X. Tang et al., “Unsupervised deep feature learning for remote sensing image retrieval,” Remote Sens., vol. 10, no. 8, Aug. 2018, Art. no. 1243.

[6] A. Ma, N. Yu, Z. Zheng, Y. Zhong, and L. Zhang, “A supervised progressive growing generative adversarial network for remote sensing image scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5618818.

[7] C. Xu, G. Zhu, and J. Shu, “A combination of lie group machine learning and deep learning for remote sensing scene classification using multi-layer heterogeneous feature extraction and fusion,” Remote Sens., vol. 14, no. 6, Mar. 2022, Art. no. 1445.

[8] Y. Ge et al., “Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval,” Multimodal Tools Appl., vol. 77, no. 13, pp. 17489–17151, Jul. 2018.

[9] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, “Deep hash learning for remote sensing image retrieval,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 4, pp. 3420–3443, Apr. 2021.

[10] U. Chaudhuri, B. Banerjee, A. Bhattacharya, and M. Datcu, “Attention-Driven graph convolution network for remote sensing image retrieval,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 8019705.

[11] H. Wang, Z. Zhou, H. Zong, and L. Miao, “Wide-Context attention network for remote sensing image retrieval,” IEEE Geosci. Remote Sens. Lett., vol. 18, no. 12, pp. 2082–2086, Dec. 2021.

[12] Q. Cheng et al., “A novel ensemble architecture of residual attention-based deep metric learning for remote sensing image retrieval,” Remote Sens., vol. 13, no. 17, Aug. 2021, Art. no. 3445.

[13] W. Xiong et al., “A discriminative feature learning approach for remote sensing image retrieval,” Remote Sens., vol. 11, no. 3, Feb. 2019, Art. no. 281.

[14] P. Wang, X. Sun, W. Diao, and K. Fu, “FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 5, pp. 3377–3390, May 2020.

[15] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, “Rotate to attend: Convolutional triplet attention module,” in Proc. IEEE Winter Conf. Appl. Comput. Vis., 2021, pp. 3139–3148.

[16] C.-B. Zhang et al., “Delving deep into label smoothing,” IEEE Trans. Image Process., vol. 30, pp. 5984–5996, 2021.

[17] A. K. Jain and A. Vailaya, “Image retrieval using color and shape,” Pattern Recognit., vol. 29, no. 8, pp. 1233–1244, Aug. 1996.

[18] B. S. Manjunath and W.-Y. Ma, “Texture features for browsing and retrieval of image data,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 8, pp. 837–842, Aug. 1996.

[19] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” Int. J. Comput. Vis., vol. 42, no. 3, pp. 145–175, May 2001.
M. P. Martins, L. F. Guimarães, and L. M. G. Fonseca, “Texture feature neural classifier for remote sensing image retrieval systems,” in Proc. Braz. Symp. Comput. Graph. Image Process., Oct. 2002, pp. 90–96.

H. Yao, B. Li, and W. Cao, “Remote sensing imagery retrieval based on Gabor texture feature classification,” in Proc. 7th Int. Conf. Signal Process., 2004, pp. 733–736.

P. Maheshwary and N. Srivastava, “Prototype system for retrieval of remote sensing images based on color moment and gray level co-occurrence matrix,” Int. J. Comput. Sci. Issues, vol. 3, pp. 20–23, Aug. 2009.

D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” Int. J. Comput. Vis., vol. 60, no. 2, pp. 91–110, Nov. 2004.

Sivic and Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in Proc. IEEE Int. Conf. Comput. Vis., 2003, pp. 1470–1470.

Y. Li, J. Ma, and Y. Zhang, “Image retrieval from remote sensing big data: A survey,” Inf. Fusion, vol. 67, pp. 94–115, Mar. 2021.

S. K. Sudha and S. Aji, “A review on recent advances in remote sensing image retrieval techniques,” J. Indian Soc. Remote Sens., vol. 47, no. 12, pp. 2129–2139, Dec. 2019.

Manjula and S. Kumar, “A comprehensive study on deep learning approach for CBIR,” in Proc. 10th IEEE Int. Conf. Commun. Syst. Netw. Technol., 2021, pp. 560–564.

E. Li, J. Xia, P. Du, C. Lin, and A. Samat, “Integrating multilayer features of convolutional neural networks for remote sensing scene classification,” IEEE Trans. Geosci. Remote Sens., vol. 55, no. 10, pp. 5653–5665, Oct. 2017.

L. Fan, H. Zhao, and H. Zhao, “Distribution consistency loss for large-scale remote sensing image retrieval,” Remote Sens., vol. 12, no. 1, Jan. 2020, Art. no. 175.

E. Liu et al., “Slice-feature based deep hashing algorithm for remote sensing image retrieval,” Infrared Phys. Technol., vol. 107, Jun. 2020, Art. no. 103299.

Z. Xiao et al., “High-Resolution remote sensing image retrieval based on CNNs from a dimensional perspective,” Remote Sens., vol. 9, no. 7, Jul. 2017, Art. no. 725.

W. Huang and F. Zhou, “DA-CapsNet: Dual attention mechanism capsule network,” Sci. Rep., vol. 10, no. 1, Jul. 2020, Art. no. 11383.

W. Li et al., “Attention mechanism and depthwise separable convolution aided 3DCNN for hyperspectral remote sensing image classification,” Remote Sens., vol. 14, no. 9, May 2022, Art. no. 2215.

R. Imbriaco et al., “Aggregated deep local features for remote sensing image retrieval,” Remote Sens., vol. 11, no. 5, Feb. 2019, Art. no. 493.

Y. Wang et al., “Attention boosted bilinear pooling for remote sensing image retrieval,” Int. J. Remote Sens., vol. 41, no. 7, pp. 2704–2724, Apr. 2020.

L. Yang et al., “SimAM: A simple, parameter-free attention module for convolutional neural networks,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 11863–11874.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.

P. Wang et al., “Understanding convolution for semantic segmentation,” in Proc. IEEE Winter Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1451–1460.

Z. Liu, R. Feng, L. Wang, Y. Zhong, and L. Cao, “D-Resunet: Resunet and dilated convolution for high resolution satellite imagery road extraction,” in Proc. IEEE Int. Geosci. Remote Sens. Symp., 2019, pp. 3927–3930.

J. Wang, Y. Zhong, Z. Zheng, A. Ma, and L. Zhong, “RSNet: The search for remote sensing deep neural networks in recognition tasks,” IEEE Trans. Geosci. Remote Sens., vol. 59, no. 3, pp. 2520–2534, Mar. 2021.

L. Peng et al., “Can cross entropy loss be robust to label noise?,” in Proc. 29th Int. Joint Conf. Artif. Intell., 2021, pp. 2206–2212.

Y. Yang and S. Newsam, “Bag-of-visual-words and spatial extensions for land-use classification,” in Proc. ACM SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst., 2010, pp. 270–279.

G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” Proc. IEEE, vol. 105, no. 10, pp. 1865–1883, Oct. 2017.

D. Hou et al., “Two novel benchmark datasets from ArcGIS and Bing world imagery for remote sensing image retrieval,” Int. J. Remote Sens., vol. 42, no. 1, pp. 240–258, Jan. 2021.

D. Hou, S. Wang, and H. Xing, “A novel benchmark dataset of color steel sheds for remote sensing image retrieval,” Earth Sci. Inf., vol. 14, no. 2, pp. 809–818, Jun. 2021.

D. Hou, Z. Miao, H. Xing, and H. Wu, “V-RRSR: An open access web-based image annotation tool for remote sensing image retrieval,” IEEE Access, vol. 7, pp. 83852–83862, 2019.

D. Hou et al., “Exploiting low dimensional features from the mobilenets for remote sensing image retrieval,” Earth Sci. Inf., vol. 13, no. 4, pp. 1437–1443, Dec. 2020.

Y. Liu et al., “Remote-sensing image retrieval with tree-triplet classification networks,” Neurocomputing, vol. 405, pp. 48–61, Sep. 2020.

L. Liu, Y. Wang, J. Peng, and A. Plaza, “DFLLR: Deep feature learning with latent relationship embedding for remote sensing image retrieval,” IEEE Trans. Geosci. Remote Sens., vol. 60, 2022, Art. no. 5608114.

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.

K. Simonovsky and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Represent., 2015, pp. 1–13.

A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks,” in Proc. IEEE Winter Conf. Comput. Vis. Pattern Recognit., 2018, pp. 839–847.

R. Liu, W. Cai, G. Li, X. Ning, and Y. Jiang, “Hybrid dilated convolution guided feature filtering and enhancement strategy for hyperspectral image classification,” IEEE Geosci. Remote Sens. Lett., vol. 19, 2022, Art. no. 5508105.

C. Li et al., “Hybrid dilated convolution with multi-scale residual fusion network for hyperspectral image classification,” Microachines, vol. 12, no. 5, May 2021, Art. no. 545.

Dongyang Hou received the Ph.D. degree in geographic information system from the China University of Mining and Technology, Xuzhou, China, in 2016.

He is currently an Associate Professor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. His main research interests include spatial-temporal data crawling, mining, and remote sensing image retrieval.

Siuyuan Wang received the B.S. degree in geographic information science from Shandong Jianzhu University, Jinan, China, in 2020. He is currently working toward the M.S. degree in cartography and geographical information engineering from Central South University, Changsha, China.

His research interests include remote sensing image retrieval and deep learning.

Xueqing Tian received the B.S. degree in geographic information science from Shandong Normal University, Jinan, China, in 2021. She is currently working toward the M.S. degree in cartography and geographical information engineering from Central South University, Changsha, China.

Her research interests include remote sensing image retrieval and deep learning.

Huaqiao Xing received the Ph.D. degree in geographic information system from the China University of Mining and Technology (Beijing), Beijing, China, in 2017.

He is currently an Associate Professor with the School of Surveying and Geo-Informatics, Shandong Jianzhu University, Jinan, China. His main research interests include land cover change detection and remote sensing image retrieval.