A Confounder-Free Fusion Network for Aerial Image Scene Feature Representation

Wei Xiong, Zhenyu Xiong, and Yaqi Cui

***Abstract*—The increasing number and complex content of aerial images have made some recent methods based on deep learning not fit well with different aerial image processing tasks. The coarse-grained feature representation proposed by these methods is not discriminative enough. Besides, the confounding factors in the datasets and long-tailed distribution of the training data will lead to biased and spurious associations among the objects of aerial images. This study proposes a confounder-free fusion network (CFF-NET) to address the challenges. Global and local feature extraction branches are designed to capture comprehensive and fine-grained deep features from the whole image. Specifically, to extract the discriminative local feature and explore the contextual information across different regions, the models based on gated recurrent units are constructed to extract features of the image region and output the important weight of each region. Furthermore, the confounder-free object feature extraction branch is proposed to generate reasonable visual attention and provide more multigrained image information. It also eliminates the spurious and biased visual relationships of the image on the object level. Finally, the output of the three branches is combined to obtain the fusion feature representation. Extensive experiments are conducted on the three popular aerial image processing tasks: 1) image classification, 2) image retrieval, and 3) image captioning. It is found that the proposed CFF-NET achieves reasonable and state-of-the-art results, including high-level tasks such as aerial image captioning.**

***Index Terms*—Aerial image processing, causal inference, feature representation, visual attention.**

I. Introduction

With the advancement of satellite imaging techniques, both the volume and quality of aerial images have grown to an unprecedented level. Conventional image processing frameworks, which are based on handcrafted features, have exposed their deficiencies in image processing speed and accuracy [1], [2]. Due to the increasing quantity and complexity of aerial image data, an efficient method is urgently required.

Recently, several research studies, which are based on deep learning, have made huge leaps in the field of computer vision [3]–[5]. Many methods have successfully been applied to the aerial image processing fields. However, the variant scale of objects and "view of God" are still making the aerial images different from the natural scene images. Besides, some aerial images with similar labels still share different semantic information, whereas other images belonging to different categories are similar in visual content. Such intraclass dispersion and interclass similarity make the features that are directly extracted with convolutional neural networks (CNNs) from the whole aerial images not discriminative enough to handle different tasks. To tackle this challenge, some research based on attention mechanism [4], region feature extraction algorithm [5], and multiscale CNNs framework [6] have been proposed to improve the discriminative ability of the model by focusing on silent regions of the whole images.

Although these methods can improve the performance in different aerial image processing tasks, they only extract the region feature and discard the global semantic information. Besides, many dense objects and contents contained in the aerial images are very ambiguous in the real-world application. Moreover, the high resolution and wide range of aerial images are often related to multiple semantic object labels, it is hard to describe the region of interest with single image label information.

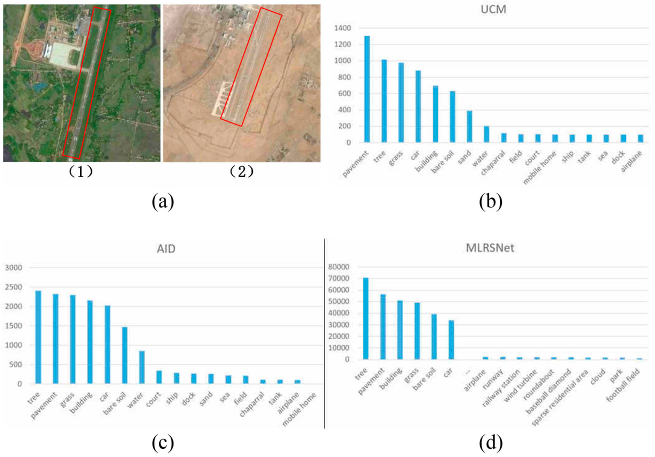
To this end, several multilabel aerial image datasets and their corresponding model [7]–[11] are proposed for a better understanding of the image. For instance, a graph-theoretic method is proposed in [7] to first retrieve multilabel aerial images in a semisupervised way. Hua *et al.* [9] newly construct a multilabel aerial image database (AID) dataset and propose an attention-aware label relational reasoning network to alleviate the label dependencies. Furthermore, a graph relation network with scalable neighbor discriminative loss is designed in [12] to learn a discriminative metric space for both the aerial image classification and retrieval tasks. Some aerial image caption frameworks [13]–[18] are also recently investigated to provide an in-depth description of the image content with flexible and precise sentences. For instance, Huang *et al.* [16] fuse the multiscale features to improve the feature representation of the model. A new attention mechanism is designed in [17] to effectively extract the scene information. To alleviate the overfitting problem during training of the image caption model, Li *et al.* [18] replace the conventional cross-entropy loss with the proposed truncation cross-entropy loss.

However, most of the existing methods are proposed by extracting the coarse-grained or single-grained features. Some intricate but discriminative objects with different scales can easily be ignored. More importantly, despite the discrimination power of the model, explainability is also a crucial ability. These deep learning methods just train the models to match and fit

Manuscript received 7 May 2022; revised 14 June 2022; accepted 3 July 2022. Date of publication 7 July 2022; date of current version 18 July 2022. This work was supported by the National Natural Science Foundation of China under Grants 61790550, 61790554, and 62001499. *(Corresponding author: Zhenyu Xiong.)*

The authors are with the Research Institute of Information Fusion, Naval Aviation University, Yantai 264001, China (e-mail: xiongwei@csif.org.cn; x_zhen_yu@163.com; cui_yaqi@126.com).

Digital Object Identifier 10.1109/JSTARS.2022.3189052



(a) (b)

(c) (d)

Fig. 1. (a) Two images with runway object labels from the MLRSNet dataset. (b)–(d) Number of images for each object in the UCM multilabel dataset, AID multilabel dataset, and MLRSNet dataset.

the training datasets without a reasonable explanation for the final decision results. The confounding factors in the datasets and long-tailed distribution of the training data will lead to biased and spurious associations among the objects of aerial images. Specifically, due to the wide imaging range of aerial images, foreground objects often co-occur with different rich background semantic objects under certain contexts. As shown in Fig. 1(a), the background information of two images with runway label vary a lot, such as farmland in Fig. 1(a) (1) and bareland in Fig. 1(a) (2). This context of aerial images is a confounder that will cause the model to make a wrong association between the input image and the predicted label. Moreover, the number of images for each object is listed in the UC Merced (UCM) multilabel dataset in Fig. 1(b); it is evident that the number of some objects such as pavement is significantly higher than that of any other objects. Furthermore, the long-tailed distribution also exists in other aerial image datasets [see Fig. 1(c)–(d)]. If the unbalanced data are directly utilized to train the model, the decision made by the learned model is biased and unreasonable and, hence, does not reflect the real causality.

In this study, a confounder-free fusion network (CFF-NET) is proposed to tackle the described problems. It mainly consists of three feature extraction branches: 1) global feature, 2) local feature, and 3) confounder-free object feature. For the global feature branch, based on the high-level features from the last convolutional layer of the CNNs, GAP is applied to compute the average value of each channel and, thus, extract the general features from images. For the local feature branch, the focus is put on extracting discriminative fine-grained image features. The high-level feature maps are divided into several patches from spatial dimensions and fed into a gated recurrent unit (GRU)-based network to generate the important weights of image regions. Attribute to the memory function of the GRU, the context information of different key regions can be captured by multiplying feature maps by importance weights. Furthermore, the confounder-free object feature extraction branch is proposed to understand aerial images from different levels and perspectives. Due to the spurious and biased associations between different objects in the aerial images, this study introduces a causal intervention model that is based on conventional likelihood to eliminate the confounding bias to explore the intrinsic relationships of different objects. Furthermore, it also pursues true causality from the input image and prediction outcome. After extraction, the outputs of each branch are concatenated to obtain the fusion features, which can help the network to capture the multi-grained information of the images. Overall, the main contributions of this study can be summarized as follows.

1) To improve the discriminative of the feature representation, a local feature extraction model is proposed based on GRU, which can help the network to extract fine-grained image features and capture context information of image regions.
2) To provide multigrained information in the process of feature representation, object-level image features are extracted in the proposed branch. Moreover, the true causality is explored in this branch to help the network in extracting semantically meaningful image features and find the intrinsic relationship of different objects in image scenes. To the best of our knowledge, this is the first work that constructs a model based on the causal inference for the aerial image processing tasks and successfully addresses the problems of furious and biased correlations between different objects.
3) The CFF-NET is applied to the three popular aerial image processing tasks: 1) image classification, 2) image retrieval, and 3) image captioning. Results of the experiments show the effectiveness of CFF-NET and its significant improvements over the existing state-of-the-art methods in each task, including high-level image processing tasks, such as aerial image captioning.

The rest of this article is organized as follows: Section II reviews existing literature related to aerial image processing tasks, feature representation for aerial images, and causal inference. Section III presents the details of the proposed CFF-NET. Section IV describes the experimental results and analyses of the three different aerial image processing tasks. Finally, Section V presents the conclusions drawn from this work.

II. Related Work

A. Aerial Image Processing Tasks

Aerial image classification: The aerial image classification is a general image processing task in the aerial community. It is simply defined as labeling an aerial image with a semantic category according to the image content. With the fast development of deep learning, most existing studies [19]–[23] have shown great success. These methods are all the single-label image classification. Considering the rich semantic information of aerial images in the real world, multilabel aerial image classification is essential for deep image understanding. In [24], the radial basis function neural network is introduced to assign unmanned aerial vehicle (UAV) images with specific multilabels. Hua *et al.* [25] utilize an attention model and bidirectional network to eliminate the class dependence and capture discriminative multiclass

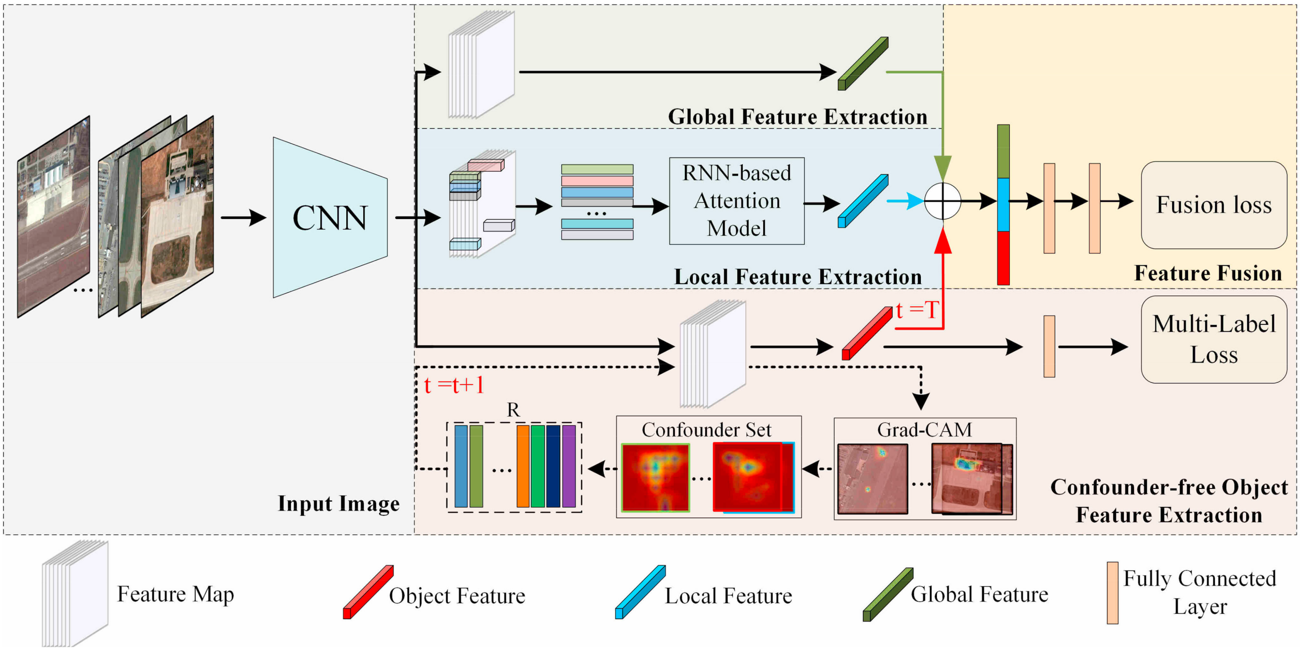


Fig. 2. Framework of our proposed CFF-NET.

features of the aerial images. In [9], the label relationships are exploited to effectively classify the multilabel aerial images.

Aerial image retrieval: The goal of aerial image retrieval is to search for aerial images of relevant information content from a large archive with a certain query image. Similar to aerial image classification, multilabel image retrieval [26]–[28] can provide more complete and valuable information about image scenes than single-label image retrieval [29]–[33]. However, the aerial image retrieval task is somehow different from the aerial image classification in that the learned feature in the metric space tends to be separable enough rather than just discriminative. In [7], Chaudhuri *et al.* construct the first multilabel aerial image dataset based on the UC-Merced dataset and propose a graph-theoretic framework to retrieve images in a semisupervised way. Sumbul *et al.* [26] incorporate both local information and multilabel relationships to retrieve aerial images' effectiveness. Kang *et al.* [12] propose a graph relation network to preserve semantic relations of complex aerial image scenes. It shows promising results both on the aerial image retrieval and classification tasks.

Aerial image captioning: Aerial image captioning aims to accurately describe the image by generating concise and flexible sentences. Recently, many aerial image captioning methods [34]–[38] have been proposed for a better understanding of image content. They are aided by three public datasets, UCM-captions dataset [39], Sydney-captions dataset [39], and RSICD [40]. Most of these methods are based on the classical encoder–decoder framework. A sound active attention framework is proposed in [34] to generate image descriptions with the guidance of sound. Furthermore, based on the existing five sentences in datasets, a topic word strategy is proposed in [35] to describe image with a memory network. In addition, a visual aligning attention model is proposed in [17] to provide accurate image captioning by focusing on regions of interest. An explainable word-sentence architecture is designed in [37] to open the black box of the encoder–decoder framework and describe the image in a human understanding way. Structured characteristics of the images have been explored in [38] to solve the problems in the field of aerial image captioning and describe the images in a fine-grained manner by generating pixelwise segmentation masks.

B. *Feature Representation for Aerial Image*

Feature representation is a key step in the process of aerial image processing. The conventional feature representation methods are based on low-level visual features [41], [42], which are not discriminative enough to represent complex image content. Recently, many learning-based methods [43]–[46] have achieved great success by representing high-level visual features of images. For instance, Xiong *et al.* [4] utilize the attention mechanism and multitask learning strategy to improve the discriminant capabilities of the model. In [43], Lin *et al.* combine both label correlation information and semantic visual information to represent meaningful information about aerial images. A region-based network [44] is proposed to represent image features with segmentation maps. This performs well by focusing on fine-grained information. To preserve the scene-level similarity relationships of images, attention mechanism and skip-layer connection strategy are combined to produce discriminative features for object-level aerial images annotation [45]. Huang *et al.* [46] fuse the multiscale features from different layers and, hence, introduce an attention model to increase the discriminative ability of feature representation. However, it is evident that these methods fail to extract the features with different grained levels. Besides, the spurious vision relationships between different classes are also not eliminated. These problems will largely hinder the effectiveness of feature representation for different image processing tasks under the complex aerial image scenes.

C. Causal Inference

Interpretability has always been a major problem in the development of deep learning. There is an urgent need to open the "black box" of neural networks and wish that the model makes a decision in a human logical manner. Causal inference is a powerful tool to design robust and explainable models. It has been widely used in the fields of medical science [47], economics [48], and social science [49]. However, these methods are proposed for the analysis of statistical models with few variables, which cannot be successfully applied in the field of computer vision. In the book of why [50], Pearl and Mackenzie mention that the existing learning-based models make a prediction by calculating the correlation and statistical distribution of the training datasets. It is reported that the model cannot perform well when the distribution of the testing dataset differs from the training dataset. However, most of the human knowledge exists in the form of cause and effect, which depend less on the observed data. To make the model imitate in this way, some recent works have successfully applied causal inference to the field of computer vision, including zero-shot learning [51], scene graph generation [52], visual dialog [53], incremental learning [54], image captioning [55], semantic segmentation [56], and visual question answering [57]. Inspired by these works, we use the causal intervention: $P(Y|do(X))$ to replace the conventional likelihood and $P(Y|X)$ to eliminate the spurious associations as well as explore the intrinsic relationships between objects in aerial images.

III. Proposed Method

A. Overview of the Method

As shown in Fig. 2, in the step of the input image, the high-level image features are extracted by the deep CNNs and the classical convolutional backbone VGG16 [58] is taken as an example. The input images are resized into 224×224 before feeding into the network. After a series of convolution and pooling operations, the high-level feature maps F_H with a size of 14×14×512 are obtained from the last convolutional layer.

Based on the high-level feature maps, three feature extraction branches are proposed in the next steps, to extract the features of the images in various grained, including global feature extraction, local feature extraction, and confounder-free object feature extraction. The three branches are detailed in Sections III-B–III-D, respectively.

To generate the final fusion feature, the three features (global, local, and confounder-free object features) are combined by the concatenation operation in the step of feature fusion. Two FC layers are then used to project the fusion feature into a one-dimensional vector. Finally, the binary cross-entropy loss is adopted as the loss function to train the network.

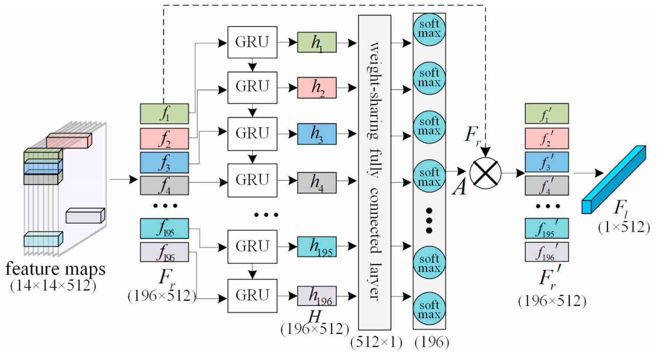


Fig. 3. Framework of local feature extraction branch.

B. Global Feature Extraction

The global feature extraction branch is proposed to produce the feature representation from the whole image to capture the comprehensive and coarse-grained information of the images. Specifically, the output of the last layer of the convolutional backbone is used as the input of the global branch. Each channel of feature maps can be considered as a type of high-level feature. The GAP is utilized to obtain the average value of each channel and the output of the pooling layer is viewed as a global feature $F_g \in \mathbb{R}^{512}$ of the images.

C. Local Feature Extraction

In the execution of this step of CFF-NET, the local feature extraction branch is designed to capture the context and fine-grained information of the images. The feature maps of the last layer of the convolutional backbone are again adopted for the local branch. As shown in Fig. 3, the feature maps are arranged from left to right and from top to bottom (scanning patterns of the human eye) in a sequential manner to obtain the high-level regional feature set $F_r = \{f_1, f_2, \ldots, f_{196}\}$, $f \in \mathbb{R}^{512}$. Then, F_r is fed into the GRUs [59] to learn the context and important region of the image. The weights of regional features are modified after training the GRUs. Given the inputs f_t and h_{t-1}, the updates of GRU are as follows:

$$z_t = \sigma\left(W_{fz} f_t + W_{hz} h_{t-1} + b_z\right) \tag{1}$$

$$r_t = \sigma\left(W_{fr} f_t + W_{hr} h_{t-1} + b_r\right) \tag{2}$$

$$n_t = \tanh\left(W_{fg} f_t + r_t * W_{hg} h_{t-1} + b_g\right) \tag{3}$$

$$h_t = (1 - z_t) * n_t + z_t * h_{t-1} \tag{4}$$

where z_t denotes the update gate, r_t denotes the reset gate, σ denotes the sigmoid function, $*$ denotes the elementwise multiplication. To further adjust the importance weight of different image regions, the output sequence, which can be denoted as, $H = \{h_1, h_2, \ldots, h_{196}\}$, $h \in \mathbb{R}^{512}$, is fed into the two fully connected (FC) layers to generate the attention weights $Q = \{q_1, q_2, \ldots, q_{196}\}$ of the regional feature. Of note, the first FC layer is the weight-sharing layer with the ReLU function, whereas the second FC layer has 196 units with the softmax function. After outputting the attention weights, the revised

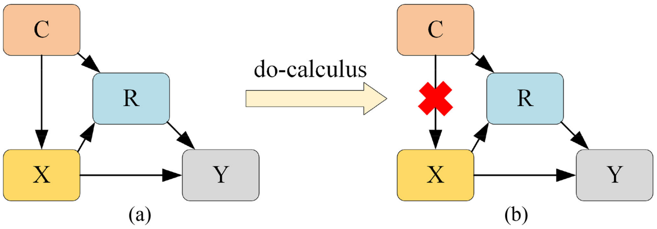


Fig. 4. (a) Constructed causal graph. (b) Causal graph after causal intervention with do-calculus.

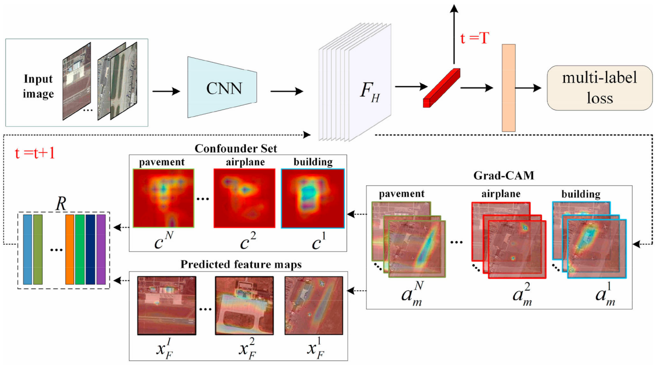


Fig. 5. Framework of confounder-free object feature extraction branch.

high-level regional feature set $F'_r = \{f'_1, f'_2, \ldots, f'_{196}\}$, $f' \in \mathbb{R}^{512}$ is calculated by the elementwise multiplication and the local feature F_l can be obtained by summing the elements in the feature set F'_r

$$F_l = \sum_{i=1}^{196} f'_i = \sum_{i=1}^{196} (q_i \times f_i). \tag{5}$$

D. Confounder-Free Object Feature Extraction

Causal intervention model: As shown in Fig. 4, a causal graph [50] is constructed to present the cause and effect between two nodes. Each node and arrow represents the variable and direct causal effects, respectively. By training a large training dataset, the conventional learning-based model generates the specific feature representation R of the input images X and output predicted labels Y [see Fig. 4(a)] in the field of computer vision. But semantic relationships among different objects [60] in the aerial image scene lead to the confounders $c \in C$, which largely distort the common relationship between X and Y resulting in erroneous conclusions by learning the $P(Y|X)$. The traditional Bayes rule can be implemented in

$$P(Y|X) = \sum_c P(Y|X,c)\, P(c|X). \tag{6}$$

In this branch, to remove the effects of confounders C on the prediction labels Y and explore the true causality from X to Y, the causal intervention model is formulated to cut off the causal link from C to X [see Fig. 4(b)]. We apply the causal intervention with do-calculus based on the Bayes rule

$$P(Y|do(X)) = \sum_c P(Y|do(X), c)\, P(c|do(X)). \tag{7}$$

Considering that c and X are independent of each other after the do-calculus operation: $P(c|do(X)) = P(c)$. So, the equation (7) can be implemented as

$$P(Y|do(X)) = \sum_c P(Y|X,c)\, P(c). \tag{8}$$

Taking the specific feature representation R and causal graph into consideration, (8) can be further represented as

$$\begin{aligned} P(Y|do(X)) &= \sum_c P(Y|X,c)\, P(c) \\ &= \sum_c P(Y|X,c,R)\, P(R|X,c)\, P(c) \\ &= \sum_c P(Y|X,c,R = g(X,c))\, P(c) \\ &= \sum_c P(Y|X,R = g(X,c))\, P(c). \end{aligned} \tag{9}$$

This can be viewed as an objective function of this branch for optimization in the following step. The function $g(\cdot)$ will be defined in later (13) and (14). To reduce computation complexity and storage cost for all the classes in calculating equation (6), normalized weighted geometric mean (NWGM) [61] is applied to move the outer expectation into the feature level. The approximate algorithm for expectation is presented as follows:

$$P(Y|do(X)) \overset{\text{NWGM}}{\approx} P\left(Y \,\middle|\, X, R = \sum_c g(X,c)\, P(c)\right). \tag{10}$$

Based on the causal intervention model, the expectation–maximum (EM) algorithm [62] is utilized to optimize the parameters in an iteration manner to generate the confounder-free object-level features. E-step and M-step are for calculating the expectation and optimizing the parameters in (10), respectively. The process of iterative optimization is detailed in the following steps.

Attention map generation: As shown in Fig. 5, the images are only put into the network and set $t = 0$ in the first stage. To generate the attention maps for each object, the multilabel classification model is trained by minimizing the multilabel loss [63]. The loss function is presented in

$$\begin{aligned} P(Y|do(X)) = \sum_{i=1}^{I} & y_i * \log \frac{1}{1+\exp(-s_i)} \\ & + (1-y_i) * \log\left(\frac{\exp(-s_i)}{1+\exp(-s_i)}\right) \end{aligned} \tag{11}$$

where $y_i \in \{0,1\}$ is the image label, $s_i = f(X,R)$presents the label prediction of the image x_i. I denotes the number of images in the dataset. R is the specific feature representation, which is an empty set when $t = 0$ and can be updated in the following iterative steps.

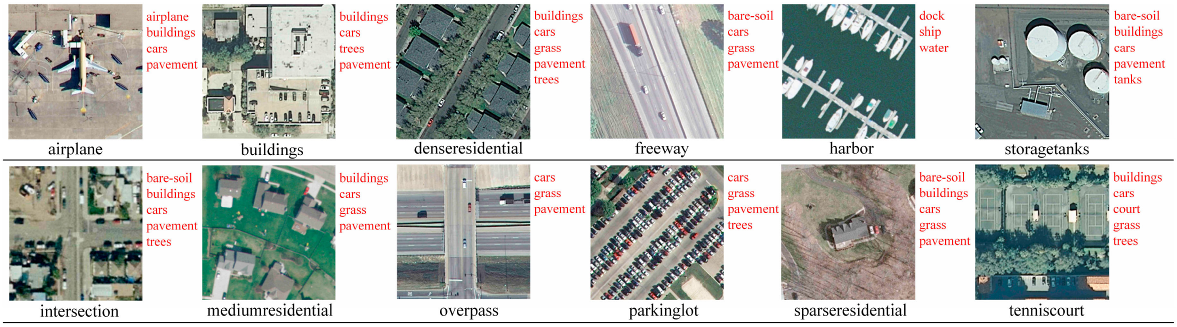


Fig. 6. Examples from the UCM multilabel dataset, the object labels in each image are signed with red font.

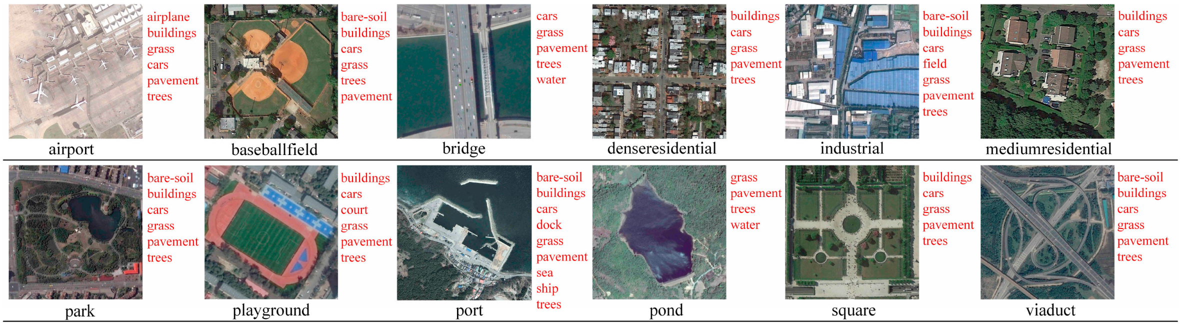


Fig. 7. Examples from the AID multilabel dataset, the object labels in each image are signed with red font.

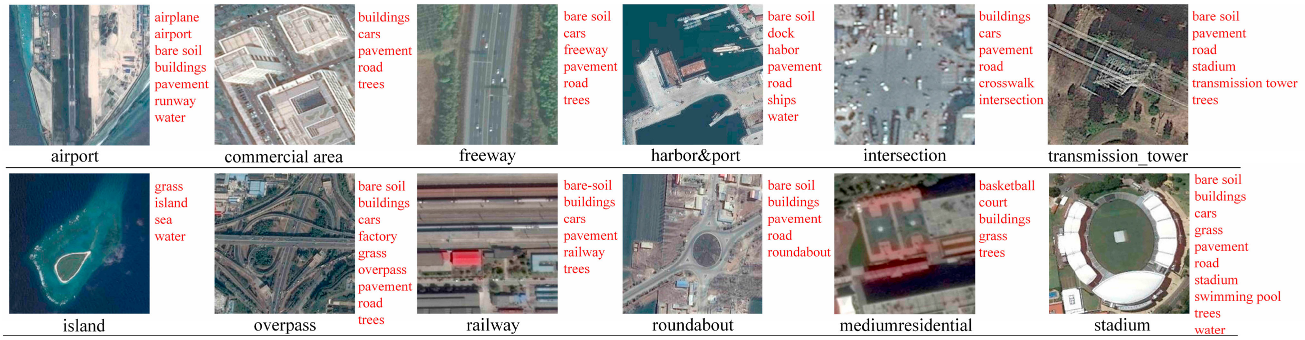


Fig. 8. Examples from the MLRSNet dataset, the object labels in each image are signed with red font.

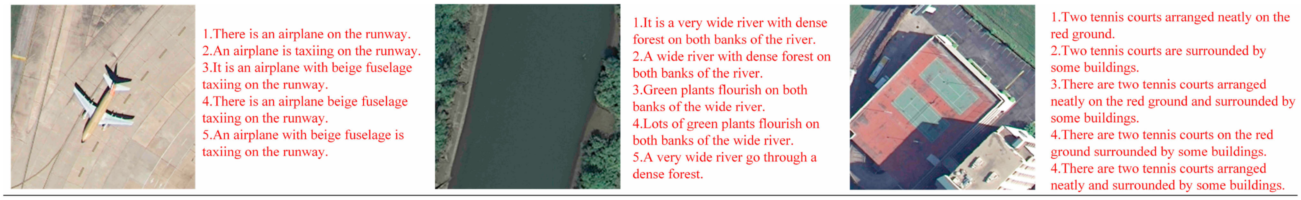
Inspired by the attention maps generation methods [64]–[66], Gradient-weighted Class Activation Mapping (Grad-CAM) [65] is chosen to produce attention maps $A^n{=}(a_1^n, a_2^n, \ldots, a_M^n)$ for every class n during the training step. M denotes the number of images in class n. The implementation process is as follows:

$$a_m^n{=}\,\mathrm{ReLU}\left(\sum_k w_k^n f_k\right) \tag{12}$$

where a_m^n presents the attention maps of *m*th images for class *n*. w_k^n represents the importance weights of the unit *k* for class *n* and f_k is the feature map activation in the last layer of the convolutional backbone.

Specific feature representation generation: Since it is hard to find all the confounders in the image scenes, it is partially revealed by constructing a confounder set $C{=}(c^1, c^2, \ldots, c^N)$ based on the feature representation of each object. N denotes the category size, whereas c^n in C is the average value of attention maps A^n within all the corresponding *n*th class in the dataset. Additionally, we collect predicted feature maps $X_F{=}(x_F^1, x_F^2, \ldots, x_F^I)$ of every training image to help in generating a specific feature representation R. Based on (10), R can be calculated using the following equations:

$$R = \sum_{n=1}^{N} \alpha_n c^n P\left(c^n\right) \tag{13}$$



1.There is an airplane on the runway.
2.An airplane is taxiing on the runway.
3.It is an airplane with beige fuselage taxiing on the runway.
4.There is an airplane beige fuselage taxiing on the runway.
5.An airplane with beige fuselage is taxiing on the runway.

1.It is a very wide river with dense forest on both banks of the river.
2.A wide river with dense forest on both banks of the river.
3.Green plants flourish on both banks of the wide river.
4.Lots of green plants flourish on both banks of the wide river.
5.A very wide river go through a dense forest.

1.Two tennis courts arranged neatly on the red ground.
2.Two tennis courts are surrounded by some buildings.
3.There are two tennis courts arranged neatly on the red ground and surrounded by some buildings.
4.There are two tennis courts on the red ground surrounded by some buildings.
4.There are two tennis courts arranged neatly and surrounded by some buildings.

Fig. 9. Examples from the UCM-captions dataset, each image is described with five sentences.

TABLE I
Classification Results Obtained by Different Convolutional Backbone With Three Proposed Branches

Conv	Models	UCM				AID				MLRSNET			
		F_s^1	F_s^2	P_s	R_s	F_s^1	F_s^2	P_s	R_s	F_s^1	F_s^2	P_s	R_s
VGG16	G	78.14	79.02	78.23	80.03	80.25	80.88	79.02	80.18	71.16	72.14	72.25	73.58
	G+L	87.21	86.14	88.01	88.26	87.24	86.02	88.25	86.14	84.23	85.26	85.14	86.21
	G+L+O	92.23	92.04	92.18	92.01	90.23	90.01	90.68	91.19	89.85	90.11	90.12	89.86
Resnet38	G	79.82	80.01	79.01	80.12	80.11	82.01	83.23	83.25	78.14	77.36	76.55	75.26
	G+L	88.15	87.88	88.57	89.14	90.21	90.85	90.11	90.96	86.33	87.25	88.24	88.21
	G+L+O	92.89	93.08	92.96	92.86	92.95	91.08	91.88	**94.86**	92.25	91.02	**94.85**	90.76
Resnet50	G	80.14	80.26	81.61	81.82	80.41	83.52	79.55	81.68	85.85	85.23	84.65	84.36
	G+L	89.42	88.85	88.98	89.10	91.85	91.24	91.14	90.11	89.24	90.22	89.85	89.88
	G+L+O	**93.28**	**93.63**	**93.91**	**94.51**	**95.32**	**94.02**	**94.97**	94.01	**95.11**	**94.01**	94.52	**93.66**

The performances are evaluated by four conventional evaluation metrics.

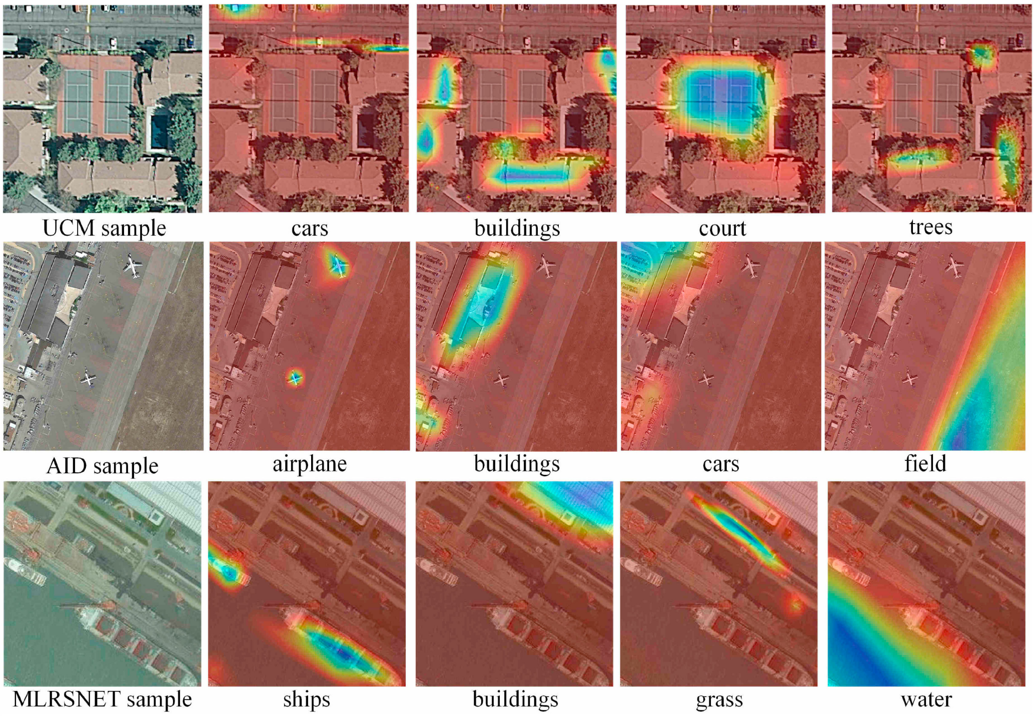


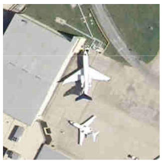
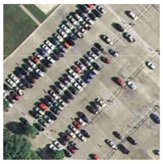
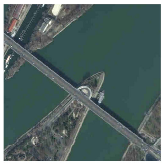
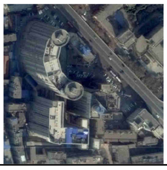
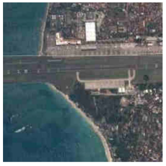
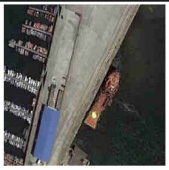
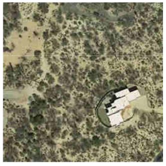
Fig. 10. Examples of class attention maps on three multilabel aerial image dataset.

$$\alpha_n = \text{softmax}\left(\frac{(W_1 X_F)(W_2 c^n)}{\sqrt{N}}\right) \tag{14}$$

where α_n is the similarity coefficient between X_F and c^n. $P(c^n)$ is set to $\frac{1}{n}$. W_1 and W_2 are the two trainable parameters in the model, which can map X_F and c^n to the common feature space.

With the specific feature representation R, the counter t is bumped up by 1 and channelwise concatenation is applied to combine R with high-level feature map to obtain $[R, F_H]$, which is used for the next training iteration ($t = 1$). The final confounder-free object feature F_o is output when $t = T$.

TABLE II
EXAMPLE PREDICTIONS ON THREE MULTILABEL AERIAL IMAGE DATASET OBTAINED WITH THREE PROPOSED BRANCHES

Images from the UCM Multi label Dataset				
Ground truths	airplane, buildings, cars, grass, pavement	buildings, cars, pavement, trees	bare soil, grass, pavement, sand	cars, grass, pavement, trees
G	airplane, buildings, cars, grass, pavement	buildings, cars, pavement, trees, grass	bare soil, grass, pavement, sand	cars, grass, pavement, trees
G+L	airplane, buildings, cars, grass, pavement	buildings, cars, pavement, trees	bare soil, grass, pavement, sand	cars, grass, pavement, trees
G+L+O	airplane, buildings, cars, grass, pavement	airplane, buildings, cars, grass, pavement	bare soil, grass, pavement, sand	cars, grass, pavement, trees
Images from the AID multi label dataset				
Ground truths	buildings, cars, dock, grass, pavement, ship, trees, water	buildings, cars, grass, pavement, trees	bare soil, buildings, cars, grass, pavement, trees	bare soil, buildings, cars, court, dock, grass, pavement, ship, trees, water
G	buildings, cars, dock, grass, pavement, ship, trees, water, bare soil	buildings, cars, grass, pavement, trees	bare soil, buildings, cars, grass, pavement, trees	bare soil, buildings, cars, court, dock, grass, pavement, ship, trees, water, sand
G+L	buildings, cars, dock, grass, pavement, ship, trees, water	buildings, cars, grass, pavement, trees, water	bare soil, buildings, cars, grass, pavement, trees	bare soil, buildings, cars, court, dock, grass, pavement, ship, trees, water
G+L+O	buildings, cars, dock, grass, pavement, ship, trees, water	buildings, cars, grass, pavement, trees	bare soil, buildings, cars, grass, pavement, trees	bare soil, buildings, cars, court, dock, grass, pavement, ship, trees, water
Images from the MLRSNET multi label dataset				
Ground truths	airport, bare soil, buildings, grass, pavement, runway, trees, water	buildings, cars, dense residential area, pavement, road, trees	buildings, dock, harbor, pavement, road, ships, water	buildings, chaparral, pavement, road, sand, sparse residential area, trees, water
G	airport, bare soil, buildings, grass, pavement, runway, trees, water , road	buildings, cars, dense residential area, pavement, road, trees, grass	buildings, dock, harbor, pavement, road, ships, water, sea	buildings, chaparral, pavement, road, sand, sparse residential area, trees, water, bare soil, grass
G+L	airport, bare soil, buildings, grass, pavement, runway, trees, water	buildings, cars, dense residential area, pavement, road, trees	buildings, dock, harbor, pavement, road, ships, water, sea	buildings, chaparral, pavement, road, sand, sparse residential area, trees, water
G+L+O	airport, bare soil, buildings, grass, pavement, runway, trees, water	buildings, cars, dense residential area, pavement, road, trees	buildings, dock, harbor, pavement, road, ships, water, sea	buildings, chaparral, pavement, road, sand, sparse residential area, trees, water

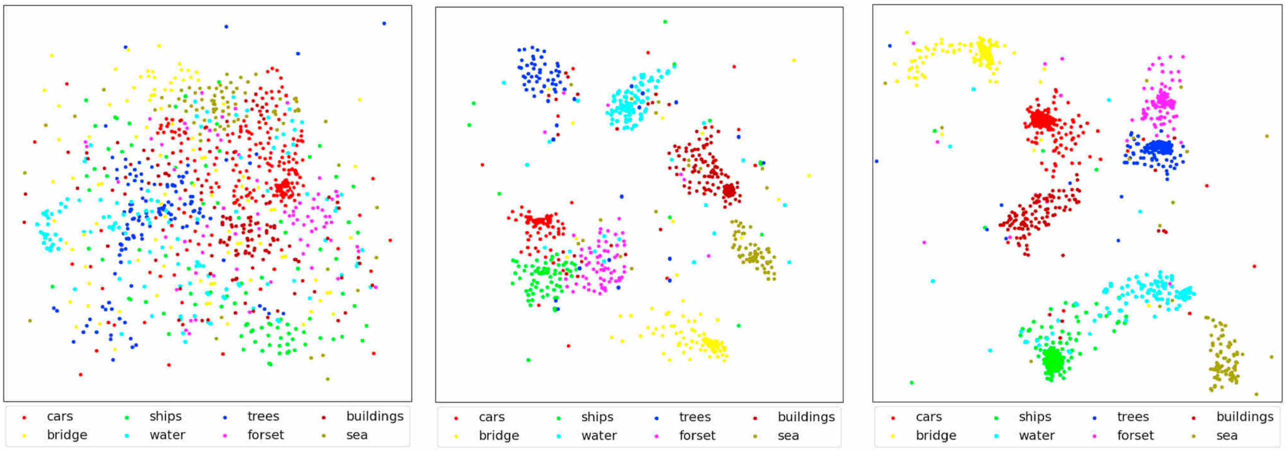
The false positives are marked in red, and the false negatives are marked in blue.

IV. EXPERIMENTS AND ANALYSIS

A. Dataset Description

To evaluate the performance of the CFF-NET, three challenging multilabel aerial datasets, UCM multilabel dataset [7], AID multilabel dataset [9], and multilabel high spatial resolution remote sensing dataset (MLRSNet) [11] are introduced for both multilabel aerial image classification and retrieval tasks. Besides, an aerial image caption dataset, the UCM-captions dataset [39], is also utilized for the aerial image caption task.

1) The UCM multilabel dataset is the first wildly used dataset for multilabel aerial image retrieval, which is reproduced based on UC Merced Land-Use dataset [41]. The original UC Merced Land-Use dataset specifically, consists of 2100 aerial images and 21 scene classes, including airplane, building, agricultural, tennis courts, dense residential, forest, freeway, golf course, mobile home park, parking lot, beach, harbor, intersection, chaparral, storage tank, medium residential, overpass, sparse residential, river, runway, and baseball diamond. Each of these land-use categories contains 100 images of 256×256 pixels and a 30-cm resolution. The UCM multilabel dataset is relabeled with 17 object classes, including airplane, water, sand, bare soil, building, ship, car, chaparral, court, tank,



(a) (b) (c)

Fig. 11. Feature visualization of the learned features under three combinations of branches. (a) G. (b) G+L. (c) G+L+O.

TABLE III
Comparison of F_s^1 With State-of-the-Art Methods

method	UCM	AID	MLRSNET
RNN-CNN [22]	67.18	60.08	59.26
RBFNN [8]	82.19	83.77	84.02
CA-BiLSTM [23]	85.11	87.63	87.46
AL-RN [24]	87.76	88.72	88.52
CFF-NET	**93.28**	**95.32**	**95.11**

grass, pavement, tree, mobile home, sea, dock, and field. The aerial image samples from this multilabel dataset are shown in Fig. 6.

2) The AID multilabel dataset is reconstructed from a large-scale aerial scene classification dataset AID [67]. The original AID dataset specifically consists of 10 000 aerial images and 30 scene categories including airport, square, baseball field, bridge, commercial, church, desert, forest, farmland, industrial, meadow, center, bare land, park, port, dense residential, playground, pond, viaduct, railway station, beach, river, resort, school, parking, stadium, sparse residential, storage tanks, medium residential, and mountain. Each of them is collected from Google Earth imagery with an image size of 600×600. However, it is noted that the number of images in each category varies from 220 to 420 and the spatial resolution ranges from 0.5 to 8 m/pixel. Furthermore, 3000 images from 30 scene categories are selected and relabeled with 17 object classes to build the AID multilabel dataset. It is observed that the multiple object labels are similar to those in the UCM multilabel dataset. Fig. 7 illustrates the aerial image samples from this multilabel dataset.

3) MLRSNet is the large multilabel aerial image dataset with the most sufficient semantic label information. The dataset consists of 109 161 aerial images with a size of 256×256. It is divided into 46 categories at the scene level including airplane, airport, dense residential area, basketball court, beach, bridge, commercial area, freeway, intersection, farmland, forest, industrial area, lake, meadow, mountain, island, harbor and port, railway, mobile home park, swimming pool, overpass, park, ground track field, parking lot, wind turbine, baseball diamond, railway station, storage tank, vegetable greenhouse, transmission tower, golf course, chaparral, cloud, desert, shipping yard, eroded farmland, river, bareland, parkway, terrace, snowberg, tennis court, roundabout, sparse residential area, wetland, and stadium. Moreover, each image from MLRSNet is categorized into one of the 60 object-level labels. The number of images in each category varies from 1500 to 3000 and the number of multiple labels corresponding to each image ranges from 1 to 13. The spatial resolution ranges from 10 to 0.1 m/pixel. Some aerial image samples from this multilabel dataset are illustrated in Fig. 8.

4) The UCM-captions dataset is the first aerial image caption dataset, which is constructed based on the UC Merced Land-Use dataset. Each image from this dataset is described with five different sentences and some image samples with the associated sentences are displayed in Fig. 9.

B. Experimental Setup

To validate the effectiveness of the proposed CFF-NET, three different aerial image processing tasks (multilabel aerial image classification, multilabel aerial image retrieval, and aerial image captioning) are conducted. In this section, some experimental setups are detailed for each task.

Multilabel aerial image classification: For the multilabel datasets UCM and AID, we randomly select 80% of images of each class for training while the remaining 20% are used for testing. For the multilabel dataset MLRSNet, 40% of images per class are selected for training our model, and the remaining 60% for testing our model. Inspired by conventional settings [68], different existing metrics are adopted to evaluate the performance of the CFF-NET on the multilabel aerial image classification

TABLE IV
RETRIEVAL RESULTS OBTAINED BY DIFFERENT CONVOLUTIONAL BACKBONE WITH THREE PROPOSED BRANCHES

method	UCM	AID	MLRSNET
IAH [81]	59.23	61.38	62.36
IDHN [82]	65.26	66.82	64.85
GTDRL [26]	84.30	88.02	86.75
GRN [12]	**99.92**	99.78	96.98
CFF-NET	**99.92**	**99.89**	**98.76**

The performances are evaluated by three conventional evaluation metrics.

TABLE V
MAP RESULTS OF COMPARISON WITH OTHER METHODS

method	UCM		AID		MLRSNET	
	train	test	train	test	train	test
IAH [81]	0.1h	7.6s	0.3h	14s	1.2h	33s
IDHN [82]	0.08h	4.8s	0.2h	9.3s	1.2h	38s
GTDRL [26]	0.4h	19s	1.1h	31s	1.9h	53s
GRN [12]	0.25h	11s	0.9h	23s	1.8h	42s
CFF-NET	0.6h	23s	1.4h	38s	2.6h	79s

task, including the average sample-based F-scores (F_s^1 and F_s^2), precision (P_s), and recall (R_s).

Multilabel aerial image retrieval: For the image retrieval task, the splitting ratio of the multilabel datasets UCM, AID, and MLRSNet are the same as the image classification task. Three commonly used evaluation metrics [average normalized modified retrieval rank (ANMRR), the mean average precision (mAP), and the hamming loss (HL)] are employed to validate the effectiveness of the CFF-NET.

Aerial image captioning: For the aerial image captioning task, the popular encoder–decoder framework [69] is adopted to generate sentences from the aerial images. The whole training process is divided into two stages, namely image representation and sentence generation. Our proposed CFF-NET is applied to obtain the fusion feature F_f in the first stage. Of note, we only retain the feature fusion layer and discard the FC layer as well as the loss function in the feature fusion branch of CFF-NET. Long-short term memory networks [70] are then adapted to generate accurate sentences from F_f in the second stage. The negative log-likelihood function [71] is utilized to optimize the parameters at the training stage followed by the other encoder–decoder framework. For the UCM-captions dataset, 80% of the images in UCM-captions are taken as training samples, 10% for validation, and the rest 10% are used for testing our model. Several metrics are introduced for aerial image caption tasks including BiLingual Evaluation Understudy (BLEU) [72], Recall-Oriented Understudy for Gisting Evaluation (ROUGE_L) [73], Metric for Evaluation of Translation with Explicit ORdering (METEOR) [74], and Consensus-based Image Description Evaluation (CIDEr) [75].

In the following experiments, three convolutional backbones (pretrained on ImageNet [76]) are adopted to generate the high-level feature from the input images, including VGG16 [58], Resnet38 [77], and Restnet50 [78]. The experiments are conducted under the same experimental setups, which makes the comparison fair. All the input images are resized to 224×224 pixels. Furthermore, the number of iterations in the confounder-free object feature extraction branch is set to 3. The stochastic gradient descent optimizer with minibatch is employed for training the model with an initial learning rate set at 0.0005. The batch size is set at 64. All experiments are implemented based on PyTorch deep learning framework with Ubuntu16.04, 32GB of RAM, 8 Intel(R) Core(TM) i7-6770K CPU, and NVIDIA GTX 1080Ti.

C. Multilabel Aerial Image Classification

1) Effectiveness of the CFF-NET: To verify the effectiveness of the three different branches of the proposed CFF-NET on the classification task, an ablation study is conducted on the three benchmark datasets. Three branches in the CFF-NET are divided into three combinations to show their performance.

1) G indicates that only the global feature extraction branch is used in the network during the training stage.
2) G+L refers to the proposed CFF-NET without a confounder-free object feature extraction branch.
3) G+L+O denotes that all the three branches are combined to train the model.

The experimental results under four conventional metrics are shown in Table I. It is easily observed that G+L+O achieved the best performance with the same convolutional backbone on the three datasets. The combination of global and local branches (G+L) significantly improves the classification accuracies by exploiting the discriminative fine-grained and context features of the aerial images. This is specifically in comparison with only the adoption of a global branch (G). G+L+O obtains higher accuracies by extracting confounder-free multigrain features. Furthermore, the average F_s^1 and F_s^2 scores can achieve above 90% on three datasets. Another important observation is that the classification accuracies improve with the widening and deepening of the convolutional backbone. However, the improvement is not apparent in the UCM multilabel dataset compared with the results on the AID multilabel dataset and the MLRSNet.

Some predicted examples from the three datasets are displayed in Table II. It is evident that G+L outperforms the G. Furthermore, some small and intricate objects can be classified accurately by fusing the discriminative global and local features. Notably, G+L+O can further improve the classification accuracy by combining three well-designed branches. In addition, it can perform well even in some images with low resolutions and complex label information, attributed to the extracted confounder-free multigrained features.

To better show the explainability and reasonableness of the CFF-NET, Grad-CAM [65] is also applied to three datasets to visualize how the network is affecting the learning of the class-specific features. The experimental results of class attention maps are displayed in Fig. 10. Evidently, the decision making of the proposed network can focus on the corresponding target category accurately and integrally, even in the case of complex background and low resolution. Meanwhile, the image regions that are unrelated to label information are also less activated.

TABLE VI
Time Comparison With Different Methods

Conv	Models	UCM			AID			MLRSNET		
		ANMRR	mAP	HL	ANMRR	mAP	HL	ANMRR	mAP	HL
VGG16	G	0.2882	79.23	0.29	0.2952	83.21	0.27	0.5713	61.12	0.36
	G+L	0.2203	92.82	0.19	0.1936	95.02	0.18	0.2652	94.23	0.23
	G+L+O	0.1878	96.01	0.14	0.1801	97.25	0.16	0.1982	96.85	0.18
Resnet38	G	0.2809	88.12	0.21	0.2652	89.23	0.19	0.2877	80.03	0.18
	G+L	0.1503	97.01	0.12	0.1301	96.11	0.09	0.1509	96.23	0.14
	G+L+O	0.1102	**99.84**	0.09	0.1026	99.10	**0.07**	0.1198	98.17	**0.10**
Resnet50	G	0.2796	88.20	0.16	0.2603	89.01	0.19	0.2633	88.25	0.18
	G+L	0.1431	97.25	0.09	0.1399	97.52	0.14	0.1452	96.36	0.12
	G+L+O	**0.1062**	99.21	**0.07**	**0.1012**	**99.89**	**0.07**	**0.1278**	**98.76**	0.11

TABLE VII
Image Captioning Results Obtained From Different Convolutional Backbone With Three Proposed Branches

Conv	Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
VGG16	G	0.6217	0.5054	0.3189	0.2012	0.1768	0.5108	1.5284
	G+L	0.7892	0.7385	0.6812	0.6301	0.4017	0.7310	3.0152
	G+L+O	0.8156	0.7869	0.7128	0.6701	0.4392	0.7718	3.1175
ResNet38	G	0.6401	0.5358	0.3346	0.2514	0.1792	0.5225	1.7145
	G+L	0.8301	0.7705	0.7014	0.6711	0.4598	0.7610	3.1158
	G+L+O	0.8710	**0.8492**	**0.7873**	0.7245	0.4812	0.8305	**3.4796**
ResNet50	G	0.6422	0.5285	0.3298	0.2610	0.1789	0.5231	1.7028
	G+L	0.8313	0.7739	0.7107	0.6613	0.4585	0.7685	3.1585
	G+L+O	**0.8718**	0.8466	0.7831	**0.7291**	**0.4893**	**0.8312**	3.4712

The performances are evaluated by three conventional evaluation metrics.

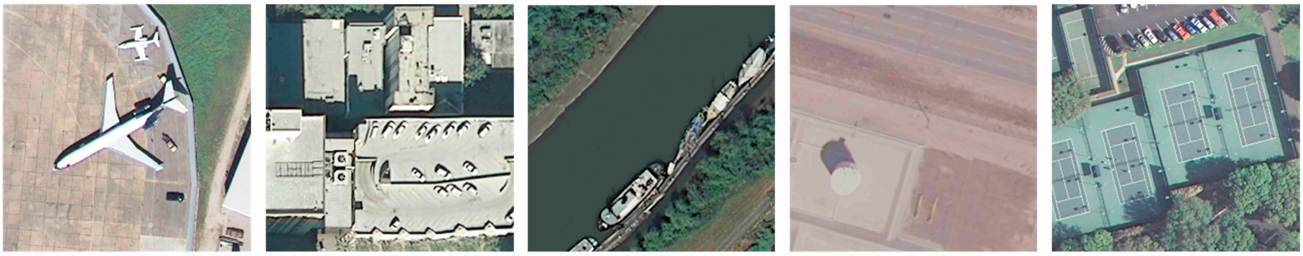
2) Comparisons With Other Methods: The performance of the proposed CFF-NET is determined by comparing F_s^1 with various state-of-the-art methods on the three datasets that are shown in Table III. Despite the fact that all the methods utilize the learning-based deep features, the performance of the RNN-CNN [79] is inferior compared to other methods. This is mainly because the RNN-CNN facilitates the multilabel image classification task of natural scene images. The complex content of aerial images limits the performance of this method. The classification accuracy of RBFNN [24] is significantly improved compared with the RNN-CNN. Although both the aerial images captured by satellite and the UAV images have a similar visual angle, the objects observed by satellite are more intricate. Discriminative fine-grained features need to be extracted to further improve the accuracy. The CA-BiLSTM [25] and AL-RN [9] protect the class-specific information by designing the attention. However, the spurious and biased relationships among the semantic objects still limit their performance. The proposed CFF-NET yields the best classification results among all the methods.

D. Multilabel Aerial Image Retrieval

1) Effectiveness of the CFF-NET: The ablation study results of the image retrieval obtained by different combinations of branches are shown in Table IV. It can be observed that the retrieval results of G+L are largely improved compared with the G, which is consistent with the classification results. This is mainly because the intraclass distance of aerial images is often larger than interclass distance in the feature space. Furthermore, only extraction of global features is not sufficient to accurately retrieve images with the same classes. The G+L can provide more discriminative information by fusing the global and local image features. G+L+O achieves the best results under all the convolutional backbones on three datasets by providing rich confounder-free multigrained information.

The two-dimensional feature representations are obtained with three combinations of branches by employing the *t*-distributed stochastic neighbor embedding (*t*-SNE) algorithm [80] on the eight classes of the MLRSNet dataset (see Fig. 11). This is to visualize the learned feature distribution in the metric space so as to show the reasonableness and explainability of the extracted feature through the proposed CFF-NET. The feature distribution observed in Fig. 11(b) is more compact than that in Fig. 11(a). The features with similar classes pull together and the features with different classes push away by fusing the discriminative global and local features. Moreover, the feature distribution in Fig. 11(c) illustrates a more reasonable outcome compared with the results of Fig. 11(a) and (b). Specifically, it is observed that without a confounder-free object feature extraction branch [see Fig. 11(b)], the feature distribution of semantically related classes, such as forests and trees, leads to a large separation in the metric space. In addition, the unrelated feature classes get closer, such as forests, ships, and cars. However, in Fig. 11(c), the feature distributions of trees and ships via the proposed CFF-NET are reasonably closer to the forest and water, respectively.

2) Comparisons With Other Methods: To show the superiority of the CFF-NET, the mAP results of comparison with



(a) An airplane in the airport.(with lack of cars)
(b) Two airplanes and some cars are in the airport.
(c) Two white airplanes and some cars are in the airport.

(a) Some buildings with grey roof .(with lack of cars)
(b) Many cars parked in the parking lot.(with lack of buildings)
(c) Many cars are parked in the roof of some buildings.

(a) Many boats docked at the harbor and the water is green.
(b) A river go through a dense forest and some boats docked at the harbor.
(c) Green plants flourish on both banks of the river and some boats moored to the bank.

(a) A airplane is on the airport.
(b) A white building is on the ground.
(c) A small white storage tank is on the ground with a road beside.

(a) Some plants are around the buildings.
(b) Some tennis courts are surrounded by a dense forest.
(c) Some tennis courts are surrounded by some plants.

Fig. 12. Some image captioning examples of the three combinations of branches on the UCM-captions dataset. The wrong words are marked in red, and the missing words are shown in blue.

TABLE VIII
IMAGE CAPTIONING RESULTS OF COMPARISON WITH OTHER METHODS

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L	CIDEr
Sound-a-a [34]	0.7484	0.6837	0.6310	0.5896	0.3623	0.6579	2.7281
Word-sentence [37]	0.7931	0.7237	0.6671	0.6202	0.4395	0.7132	2.7871
RTRMN [35]	0.8028	0.7322	0.6821	0.6393	0.4258	0.7726	3.1270
VAA [17]	0.8192	0.7511	0.6927	0.6387	0.4380	0.7824	3.3946
SAL [83]	0.8154	0.7575	0.6936	0.6458	0.4240	0.7632	3.1864
HA [40]	0.8157	0.7312	0.6702	0.6182	0.4263	0.7698	2.9947
Multi-level Att [84]	0.8330	0.7712	0.7154	0.6623	0.4371	0.7763	3.1684
TCE [18]	0.8210	0.7622	0.7140	0.6700	0.4775	0.7567	2.8547
Structured Att [38]	0.8538	0.8035	0.7572	0.7149	0.4632	0.8141	3.3489
CFF-NET	**0.8718**	**0.8492**	**0.7873**	**0.7291**	**0.4893**	**0.8312**	**3.4796**

other methods are shown in Table V. It is evident that the performance of the IAH [81] and IDHN [82] is inferior in the three datasets. This is because the methods are proposed for processing the image in natural scenes. The aerial images are quite different from the natural images, which exist in large intraclass variations and small interclass differences. It is hard to transfer their success to the aerial image retrieval task. The proposed CFF-NET achieves the best retrieval results on both AID and MLRSNET datasets and is equivalent to that of GRN [12] on the UCM dataset.

For the image retrieval tasks, computational efficiency is also an important measure. The computational complexity of the proposed method can be analyzed by comparing the training and testing time of our method with various baselines as shown in Table VI. Considering that both IAH and IDHN have high retrieval speeds, which use hash codes to represent image features. The increased performance brings the increased computational costs. The proposed methods obtain the highest retrieval accuracy and control the retrieval time within an acceptable range.

E. Aerial Image Captioning

1) Effectiveness of the CFF-NET: For the aerial image captioning task, an ablation study is also conducted with several evaluation metrics to analyze the importance of the three branches (see Table VII). It is observed that the accuracy of the ResNet is higher than that of the VGG. It is also evident that the results of G+L+O with the ResNet38 and ResNet50 are similar in all kinds of evaluation metrics. Furthermore, under the same convolutional backbone, the results show that G+L+O achieves the best results on all metrics and outperform the other two models with a significant advantage.

Some image captioning examples of the three combinations of branches on the UCM-captions dataset are shown in Fig. 12 to intuitively display the effectiveness of the proposed CFF-NET. The three sentences marked (a), (b), and (c) in each image represent the G, G+L, and G+L+O, respectively. The wrong words in the generated sentences are marked in red color, whereas the missing words are shown in blue color. Notably, the adoption of the global branch can only extract the coarse-grained feature, which will neglect the tiny object information such as cars. Besides, this branch does not utilize the semantic-related information of the various objects. This leads to the generated sentences failing to correctly describe the scene information of the target. Moreover, G+L generates a complete and full description by extracting the discriminative fine-grained feature and also by capturing the context information from the whole image. However, the long-tailed distribution caused by imbalanced frequency occurrences of some objects in the datasets results in wrong descriptions. Furthermore, the high-frequency co-occurrences of some objects such as buildings also lead to the bias and furious associations among objects in the process of image captioning. In comparison, G+L+O describes most of the images accurately with the generated sentences. The learned multigrained features and confounder-free feature extraction strategy eliminate the cognitive errors as well as describe the image reasonably and in detail.

2) Comparisons With Other Methods: To verify the superiority of the proposed CFF-NET on the aerial image captioning task, the comparison experiments are also conducted with some state-of-the-art methods. The experimental results are also reported in Table VIII. Compared with all the state-of-the-art methods, the proposed CFF-NET has the best performance in terms of all evaluation metrics. It can also be observed that CFF-NET has an obvious advantage in the most of entries. Therefore, these results completely validate the effectiveness of the proposed image processing strategy.

V. Conclusion

In this article, a novel CFF-NET is proposed for aerial image feature representation. Considering the rich and complex content of aerial images, a local feature extraction branch is proposed to extract the discriminative fine-grained and context feature of the image. The problems of long-tailed distribution of the dataset and spurious correlation between objects in image scenes are rarely explored in the field of aerial image processing. This is one of the main reasons why the existing models lack being explainable. To address this problem, the confounder-free object feature extraction branch is proposed to extract semantically meaningful features and find the reliable causality between the input image as well as the predicted outcome by utilizing a causal intervention strategy. To the best of our knowledge, this is the first work that constructs a model based on the causal inference for the aerial image processing tasks and successfully addresses the problems of furious and biased correlations between different objects. Extensive experiments are conducted under three popular aerial image processing tasks. The proposed CFF-NET has shown state-of-the-art results, including a high-level image processing task such as aerial image captioning.

References

[1] S. Ramirez-Gallego, A. Fernandez, S. Garcia, M. Chen, and F. Herrera, “Big data: Tutorial and guidelines on information and process fusion for analytics algorithms with mapreduce,” *Inf. Fusion*, vol. 42, pp. 51–61, 2018.

[2] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, “Image matching from handcrafted to deep features: A survey,” *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, 2021.

[3] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, “Image fusion meets deep learning: A survey and perspective,” *Inf. Fusion*, vol. 76, pp. 323–336, 2021.

[4] W. Xiong, Y. Lv, Y. Cui, X. Zhang, and X. Gu, “A discriminative feature learning approach for remote sensing image retrieval,” *Remote Sens.*, vol. 11, no. 4, 2019, Art. no. 281.

[5] P. Li, P. Ren, X. Zhang, Q. Wang, X. Zhu, and L. Wang, “Region-wise deep feature representation for remote sensing images,” *Remote Sens.*, vol. 10, 2018, Art. no. 871.

[6] X. Zheng, Y. Yuan, and X. Lu, “A deep scene representation for aerial scene classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 4799–4809, Jul. 2019.

[7] B. Chaudhuri, B. Demir, S. Chaudhuri, and L. Bruzzone, “Multilabel remote sensing image retrieval using a semisupervised graphtheoretic method,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1144–1158, Feb. 2018.

[8] Y. Hua, L. Mou, and X. X. Zhu, “Label relation inference for multi-label aerial image classification,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5244–5247.

[9] Y. Hua, L. Mou, and X. Zhu, “Relation network for multilabel aerial image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 7, pp. 4558–4572, Jul. 2020.

[10] F. Ömrüuzun, B. Demir, L. Bruzzone, and Y. Çetin, “Content based hyperspectral image retrieval using bag of endmembers image descriptors,” in *Proc. 8th Workshop Hyperspectral Image Signal Process.*, Aug. 2016, pp. 1–4.

[11] X. Qi, P. Zhu, and Y. Wang, “MLRSNet: A multi-label high spatial resolution remote sensing dataset for semantic scene understanding,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 16, no. 9, pp. 337–350. 2020.

[12] J. Kang, R. Fernandez-Beltran, D. Hong, J. Chanussot, and A. Plaza, “Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4355–4369, May 2021.

[13] G. Hoxha, F. Melgani, and B. Demir, “Toward remote sensing image retrieval under a deep image captioning perspective,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4462–4475, 2020.

[14] G. Sumbul, S. Nayak, and B. Demir, “SD-RSIC: Summarization-driven deep remote sensing image captioning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 8, pp. 6922–6934, Aug. 2021.

[15] B. Wang, X. Lu, X. Zheng, and X. Li, “Semantic descriptions of high-resolution remote sensing images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1274–1278, Aug. 2019.

[16] W. Huang, Q. Wang, and X. Li, “Denoising-based multiscale feature fusion for remote sensing image captioning,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 436–440, Mar. 2021.

[17] Z. Zhang, W. Zhang, W. Diao, M. Yan, X. Gao, and X. Sun, “VAA: Visual aligning attention model for remote sensing image captioning,” *IEEE Access*, vol. 7, pp. 137355–137364, 2019.

[18] X. Li, X. Zhang, W. Huang, and Q. Wang, “Truncation cross entropy loss for remote sensing image captioning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5246–5257, Jun. 2021.

[19] Y. Zhang, Y. Ma, X. Dai, H. Li, X. Mei, and J. Ma, “Locality-constrained sparse representation for hyperspectral image classification,” *Inf. Sci.*, vol. 546, pp. 858–870, 2021.

[20] J. Jiang, J. Ma, and X. Liu, “Multilayer spectral-spatial graphs for label noisy robust hyperspectral image classification,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 839–852, Feb. 2022.

[21] B. Rasti and P. Ghamisi, “Remote sensing image classification using subspace sensor fusion,” *Inf. Fusion*, vol. 64, pp. 121–130, 2020.

[22] Y. Li, Y. Zhang, and Z. Zhu, “Error-tolerant deep learning for remote sensing image scene classification,” *IEEE Trans. Cybern.*, vol. 51, no. 4, pp. 1756–1768, Apr. 2021.

[23] Y. Yao *et al.*, “Continuous multi-angle remote sensing and its application in urban land cover classification,” *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 413.

[24] A. Zeggada, F. Melgani, and Y. Bazi, “A deep learning approach to UAV image multilabeling,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 694–698, May 2017.

[25] Y. Hua, L. Mou, and X. Zhu, “Recurrently exploring class-wise attention in a hybrid convolutional and bidirectional LSTM network for multi-label aerial image classification,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 149, pp. 188–199, 2019.

[26] G. Sumbul and B. Demir, “A novel graph-theoretic deep representation learning method for multi-label remote sensing image retrieval,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 266–269.

[27] Z. Shao, K. Yang, and W. Zhou, “Performance evaluation of single-label and multi-label remote sensing image retrieval using a dense labeling dataset,” *Remote Sens.*, vol. 10, no. 6, 2018, Art. no. 964.

[28] R. Imbriaco, C. Sebastian, E. Bondarev, and P. H. N. de With, “Toward multilabel image retrieval for remote sensing,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4703214.

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

[30] Y. Li, Y. Zhang, X. Huang, and J. Ma, “Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 11, pp. 6521–6536, Nov. 2018.

[31] Y. Li, Y. Zhang, X. Huang, H. Zhu, and J. Ma, “Large-scale remote sensing image retrieval by deep hashing neural networks,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 950–965, Feb. 2018.

[32] W. Zhou, S. Newsam, C. Li, and Z. Shao, "PatternNet: A benchmark dataset for performance evaluation of remote sensing image retrieval," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 197–209, 2018.

[33] W. Zhou, N. Shawn, C. Li, and Z. Shao, "Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval," *Remote Sens.*, vol. 9, no. 5, 2017, Art. no. 489.

[34] X. Lu, B. Wang, and X. Zheng, "Sound active attention framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1985–2000, Mar. 2020.

[35] B. Wang, X. Zheng, B. Qu, and X. Lu, "Retrieval topic recurrent memory network for remote sensing image captioning," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 256–270, 2020.

[36] G. Hoxha and F. Melgani, "A novel SVM-based decoder for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5404514.

[37] Q. Wang, W. Huang, X. Zhang, and X. Li, "Word–sentence framework for remote sensing image captioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10532–10543, Dec. 2021.

[38] R. Zhao, Z. Shi, and Z. Zou, "High-resolution remote sensing image captioning based on structured attention," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603814.

[39] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *Proc. Int. Conf. Comput.*, Jul. 2016, pp. 1–5.

[40] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 4, pp. 2183–2195, Apr. 2018.

[41] J. Yang, J. Liu, and Q. Dai, "An improved bag-of-words framework for remote sensing image retrieval in large-scale image databases," *Int. J. Digit. Earth*, vol. 8, no. 8, pp. 273–292, 2015.

[42] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3304–3311.

[43] D. Lin, J. Lin, L. Zhao, Z. Wang, and Z. Chen, "Multilabel aerial image classification with a concept attention graph neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5602112.

[44] 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 Observ. Remote Sens.*, vol. 13, pp. 318–328, 2020.

[45] P. Zhu *et al.*, "Deep learning for multilabel remote sensing image annotation with dual-level semantic concepts," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 6, pp. 4047–4060, Jun. 2020.

[46] R. Huang, F. Zheng, and W. Huang, "Multilabel remote sensing image annotation with multiscale attention and label correlation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 6951–6961, 2021.

[47] Q. Zhao, E. Adeli, and M. Pohl, "Training confounder-free deep learning models for medical applications," *Nat. Commun.*, vol. 11, 2020, Art. no. 6010.

[48] A. Elizabeth, "Matching methods for causal inference: A review and a look forward," *Stat. Sci.: Rev. J. Inst. Math. Statist.*, vol. 25, no. 1, 2010, Art. no. 1.

[49] M. Lechner, "Earnings and employment effects of continuous off-the-job training in east Germany after unification," *J. Bus. Econ. Statist.*, vol. 17, no. 1, pp. 74–90, 1999.

[50] J. Pearl and D. Mackenzie, *The Book of Why: the New Science of Cause and Effect*. London, U.K.: Penguin Books, 2018.

[51] Z. Yue, T. Wang, Q. Sun, X. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15399–15409.

[52] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3713–3722.

[53] J. Qi, Y. Niu, J. Huang, and H. Zhang, "Two causal principles for improving visual dialog," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10857–10866.

[54] X. Hu, K. Tang, C. Miao, X. Hua, and H. Zhang, "Distilling causal effect of data in class incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3956–3965.

[55] X. Yang, H. Zhang, and J. Cai, "Deconfounded image captioning: A causal retrospect," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[56] D. Zhang, H. Zhang, J. Tang, X. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2020.

[57] Y. Niu, K. Tang, H. Zhang, Z. Lu, X. Hua, and J. Wen, "Counterfactual VQA: A cause effect look at language bias," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2021.

[58] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, May 2015.

[59] K. Cho *et al.*, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. IEEE Conf. Computation Lang.*, 2014, pp. 1724–1734.

[60] D. Marr, *Vision: A computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA, USA: MIT Press, 1982.

[61] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, vol. 5.

[62] C. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 1, no. 1, pp. 95–103, 1983.

[63] E. Rudd, M. Günther, and T. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *Proc. Eur. Conf. Comput. Vis.*, 2016, vol. 5.

[64] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–662.

[66] 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. Appl. Comput. Vis.*, 2018, pp. 839–847.

[67] G. Xia *et al.*, "AID: A benchmark data set for performance evaluation of aerial scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3965–3981, Jul. 2017.

[68] Z. Chen, X. Wei, P. Wang, and Y. Guo, "Learning graph convolutional networks for multi-label recognition and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[69] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[70] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[71] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.

[72] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, pp. 311–318.

[73] C. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association Computational Linguistics, 2004.

[74] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Statist. Mach. Transl.*, 2014, pp. 376–380.

[75] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[76] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[77] Z. Wu, C. Shen, and A. Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognit.*, vol. 90, no. 1, pp. 119–133, 2019.

[78] 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.

[79] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2285–2294.

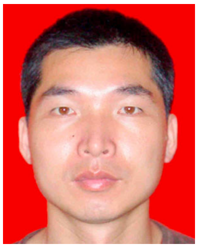
[80] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 8, pp. 2579–2605, 2008.

[81] H. Lai, P. Yan, X. Shu, Y. Wei, and S. Yan, "Instance-aware hashing for multi-label image retrieval," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2469–2479, Jun. 2016.

[82] Z. Zhang, Q. Zou, Y. Lin, L. Chen, and S. Wang, "Improved deep hashing with soft pairwise similarity for multi-label image retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 2, pp. 540–553, Feb. 2020.

[83] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, "Description generation for remote sensing images using attribute attention mechanism," *Remote Sens.*, vol. 11, no. 6, Mar. 2019, Art. no. 612.

[84] Z. Yuan, X. Li, and Q. Wang, "Exploring multi-level attention and semantic relationship for remote sensing image captioning," *IEEE Access*, vol. 8, pp. 2608–2620, 2020.



Wei Xiong received the B.S., M.S., and Ph.D. degrees from Naval Aviation University, Yantai, China, in 1998, 2001, and 2005, respectively.

From 2007 to 2009, he was a Postdoctoral Researcher with the Department of Electronic Information Engineering, Tsinghua University, Beijing, China. He is currently a Full Professor with Naval Aviation University. He is one of the founders and the directors of the Research Institute of Information Fusion, Naval Aviation University. He is the Member and the Director General of Information Fusion Branch of the Chinese Society of Aeronautics and Astronautics. His research interests include pattern recognition, remote sensing, and multisensor information fusion.

Zhenyu Xiong received the B.S. and M.S. degrees in 2018 from Naval Aviation University, Yantai, China, where he is currently working toward the Ph.D. degree in information and communication engineering.

His research interests include information fusion, and deep learning with their applications in remote sensing.

Yaqi Cui received the B.S., M.S., and Ph.D. degrees in information and communication engineering from Naval Aviation University, Yantai, China, in 2008, 2011, and 2014, respectively.

Since 2014, he has been a Lecturer with Naval Aviation University. His research interests include information fusion, machine learning, and deep learning with their applications in information fusion.