

RESEARCH ARTICLE

Ultra-High-Definition Aerial Photo Categorization by an Enhanced Matrix Factorization Algorithm

JUNWU ZHOU¹, GUIFENG WANG², AND FUJI REN³, (Senior Member, IEEE)¹School of Higher Vocational and Technical College, Shanghai Dianji University, Shanghai 201306, China²Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321016, China³College of Computer Sciences, Hefei University of Technology, Hefei 242099, China

Corresponding author: Fuji Ren (fujiren.tokuma@gmail.com)

ABSTRACT In this work, we designed an effective ultra-high-definition (UHD) aerial photo categorization pipeline by designing an enhanced deep multi-clue matrix factorization (DMCMF). In detail, given a UHD aerial photo, those visually salient ground objects are extracted in the first place. In order to explicitly encode their spatial layout, multiple graphlets are constructed in each UHD aerial photo. Each is built by connecting those spatially neighboring object patches. Afterward, we propose a new matrix factorization (MF) model that intelligently uncover the underlying semantic features from graphlets. And multiple informative clues are encoded into the MF model. Notably, our DMCMF is optimized progressively. And we can represent each graphlet by a vector of binary hash codes. Lastly, each UHD aerial photograph can be effectively quantized into a feature vector by a kernel machine for multi-label categorization. Experiments have shown that our method is highly competitive in learning categorization model from imperfect labels at image-level.

INDEX TERMS Media analysis, aerial photo, multi-clue, matrix factorization.


I. INTRODUCTION

Thanks to the technology of delivering multiple satellites in a single rocket launch, thousands of earth observation satellites were sent into space recently. The satellites capture UHD aerial photos (typical resolutions over $5K \times 5K$) containing ground objects with sophisticated spatial interactions, such as reticular, star, and triangle. Semantically understanding these ground objects and their spatial topologies is a useful technique in many state-of-the-art intelligent systems. For example, it is significant to fast recognize the complicated street networks, *e.g.*, star and tree geometries, to optimize vehicle path planning (*i.e.*, calculating the shortest path between pairwise locations). Actually, it is feasible to represent the above topologies using a small graph. Each edge connects two adjacent streets.

In computer vision, dozens of shallow/deep visual categorization/annotation models were designed to describe aerial photos with regular resolutions (typically $800 \times 800 \sim 2K \times 2K$). Well-known models involves: 1) CNN (convolutional neural network)-based object localization using weak

labels [1], [2]; 2) graph models for semantically annotating aerial photos [3], [4]; and 3) elaborately-developed deep models for semantically understanding aerial images [5], [6], [7]. Nevertheless, to our best knowledge, the current deep models fail to satisfactorily characterize UHD aerial photos because of the following reasons:

- Actually, a high-resolution aerial photo typically has lots of objects distributed with various spatial configurations. Accurately uncovering the underlying semantic features is nontrivial. Possible challenges include: i) computationally and spatially modeling the complicated layout of the ground objects, and ii) formulating a deep model converting the spatially modeled features to fixed-length visual features. Moreover, spatially transforming different layouts inside each UHD aerial image into some traditional classification model [8] is a true difficulty;
- The large number of objects within each UHD aerial photo makes it labor-intensive to accurately annotate all the ground objects. Because of the progresses in weakly-labeled feature engineering, we solely require image-level labels are required for calculating semantics at region-level. Therefore, to exploit the regional-level

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

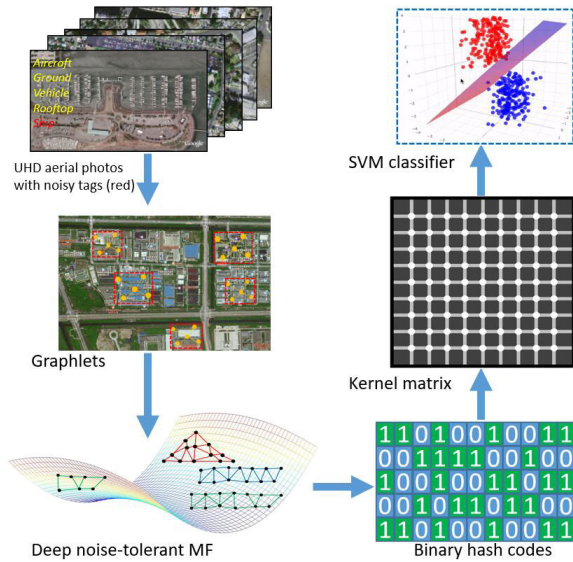


FIGURE 1. The pipeline of our proposed noise-robust binary matrix factorization (MF) framework for UHD aerial photo categorization.

visual semantics within a UHD aerial photo, it is necessary to uncover the corresponding weak and user-defined labels. Noticeably, the above user-defined labels are subjectively defined in practice. And sometimes they are noisy. In practice, establishing such a noise-tolerant refinement mechanism is difficult;

- To build a powerful UHD aerial photo categorization framework, we have to model the inherent sample distributions on manifold precisely. Practically, however, because of the contaminated user-defined labels, our previously calculated sample distribution is usually imperfect. Practically, what we need is a mathematical model that adaptively learns the optimal sample interaction when refining the labels. Actually, formulating a solvable and learnable framework requires domain experiences.

In our work, we propose a deep multi-clue matrix factorization (DMCMF) framework for multi-label UHD aerial photo categorization. The core technique is an enhanced MF which hierarchically converts the graphlets within a UHD aerial photo to corresponding binary features. Herein, the potentially noisy labels can be theoretically abandoned and the data graph can be progressively optimized. The entire pipeline of the proposed method is displayed in Fig. 1. In detail, for massive-scale UHD aerial photos, each contains user-defined labels that are potentially noisy, a succinct set of object-guided image regions are extracted in the first place. Thereafter, a rich set of object patches that spatially neighboring are collected to build different graphlets. They can accurately encode the various topologies inside multiple UHD aerial images. Based on this, we build a matrix factorization (MF) algorithm that is robust to label noises. The MF can effectively convert graphlets to the binary feature accordingly. Thereby, we can compare two graphlets rapidly and mathematically. Our proposed MF can optimally

encode four descriptive visual clues. By leveraging the binary vectors calculated from different graphlets, we convert graphlets within a UHD aerial image to the kernel-induced feature vector. In this way, an effective classifier is learned to classify different aerial images into the corresponding classes. Plenty of quantitative comparisons to the well-known deep classification models indicated the competitiveness of the learned classification model.

The main contributions can be summarized in the following: 1) a million-scale partially mislabeled UHD aerial photos collected from 100 metropolises for validating the superiority of our method, 2) a DMCMF that collaboratively and seamlessly incorporates four clues to compute the hash codes of each graphlet; and 3) a novel UHD aerial photo categorization model that avoids noisy image-level labels and adaptively updates the data distribution.

Organization of the remaining parts of article is given as follows. In Sec II we review the published work closely related to ours. Sec III delineates the proposed pipeline, including graphlet construction, our enhanced MF, and the kernel-induced feature learning. Experimental validation in Sec IV tests the effectiveness of our method. The last section concludes.

II. RELATED WORK

Dozens of computational visual models were developed to analyze aerial photos.¹ To semantically model the entire image, Zhang et al. [9] constructed a novel topological feature to model the inter-region connection inside each aerial photo. And a kernel-induced vector is calculated as the image representation for categorization. Xia et al. [10] formulated a weak learning model that semantically labels HR aerial photos at image-level. Akar [11] carefully combine the so-called random forest and object-level feature extractor to classify remote sensing images. Sameen et al. [12] developed a hierarchical deep architecture to calculate the multiple labels of HR aerial photos describing many downtown areas. In [13], Cheng et al. utilized a pre-trained deep CNN to classify high-resolution remote sensing images. A domain-specific scenic picture set is leveraged to fine tune the deep architecture. In [14], a cross-modality learning framework is proposed to collaboratively learn five deep models for categorizing aerial images, wherein pixel-level and spatial-level features are exploited complementarily. The authors [15] designed a novel inter-attentional algorithm to learn the weights of aerial image features both horizontally and vertically. In [16], Bazi et al. formulated a vision transformer for aerial image classification, wherein the long-term contextual dependencies among regions can be intrinsically encoded.

For region-level modeling, Wang et al. [17] designed a hierarchical deep architecture for discovering attractive objects with different scales. In [18], a focal loss deep

¹A more comprehensive survey of deep-learning-based aerial photo understanding is illustrated in [23].

architecture is proposed that optimally discovers vehicles from aerial images. In [19], researchers developed a geographic object detection model toward remote sensing images by intelligently extracting intersections as well as streets. In [20], Yu et al. integrated feature enhancement and soft label assignment into an anchor-independent object detector toward aerial images. In [21], Wang et al. proposed a deep rotation-invariant detector that effectively estimates the angles of multi-scale objects inside aerial images. In [22], Chalavadi et al. proposed a parallel deep model called mSODANet that hierarchically learns contextual features from multi-scale and multi-FoV (field-of-views) ground objects. Notably, different from the above methods, our approach is bionic-inspired and accurately mimics human gaze behavior.

III. OUR PROPOSED METHOD

A. TOPOLOGICAL FEATURE ENGINEERING

Nowadays, in each UHD aerial image, we can observe tens of multi-scale ground objects. Based on the psychological progresses [24] in the past decade, human observers practically attend to the foreground objects that are visually or semantically prominent when they perceiving the world. Specifically, if humans tend to understand a UHD aerial photo, their visual and cognition subsystem will perceive visually or semantically salient ground objects firstly, *e.g.*, an aircraft and its components. Meanwhile, it is observable that those background regions are nearly neglected. Obviously, it is necessary to encode the visually/semantically perceptual experience when building a UHD aerial image classification pipeline. Herein, a rapid object patches generation as well as a manifold-guided active feature selection are adopted to obtain the ground salient image patches describing different objects.

In our implementation, the well-known BING [25] descriptor is deployed to capture different ground objects because of its following superiorities: 1) achieving a sufficiently high object discovery precision and speed; 2) obtaining multiple highly descriptive and representative object-level patches that effectively simulate how human perceiving different UHD aerial images; and 3) is capable to be generalized to new UHD aerial photo classes. Therefore, the trained categorization algorithm can be transferred onto different data sets. Noticeably, by adopting the aforementioned BING, we still observe too many object-level patches within the UHD aerial images. Actually, we observe that during human visual perception, typically < 10 objects are perceived in each UHD aerial image. To mimic this, a powerful active learning [26] algorithm is utilized to select K descriptive object-level patches from a UHD aerial photo. The algorithm integrates two features: 1) the spatial configurations to each image and 2) visual semantics from image-level labels.

Based on these K patches, we can build a graphlet by the random walking algorithm [27]. The random walk is conducted on multiple neighboring patches. As shown in Fig. 2, we first construct a multi-layer spatial pyramid, based

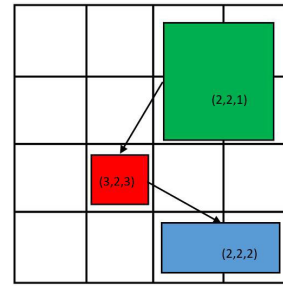


FIGURE 2. An example elaborating pairwise object patches that are adjacent spatially. Herein, the red object patch is spatially adjacent with the green one and blue one respectively. The coordinate denotes the position of each object patch in the multi-layer spatial pyramid.

on which two patches are treated as spatially adjacent if the corresponding cells are neighboring. Afterward, we randomly choose an initial patch. Next, we jump to a spatially adjacent object patch. Such jumping operation step if the graphlet size reaches our pre-defined upper bound. Afterward, we collect these adjacent patches (along the path of random walk) to form a graphlet. Herein, we present an example of how to build graphlet with three vertices is Fig. 2.

Based on graph theory, the inherent patches associated with the spatial distribution collaboratively determine the visual appearance. In our implementation, for a graphlet, it can be naturally represented by a matrix $A = [A_1, A_2]$, where A_1 denotes a matrix whose size is $K \times 137$. Herein, the 137 dimensions are obtained by concatenating a 9-dimensional color moment [28] and a 128-dimensional HOG [29]. A_2 denotes a matrix whose size is $K \times K$. In detail, when the i -th and j -th patches are spatially adjacent, then we set $A_2(i, j)$ to one, and otherwise we set $A_2(i, j)$ to zero. In order to receive a conventional feature, we row-wise condense matrix A into a long vector \mathbf{x} .

B. OUR DMCMF

In our work, the have to efficiently and effectively calculate the distance between graphlets from two UHD aerial image, whose image labels might be contaminated. Herein, we formulate a multi-component MF technique optimally handling noises from image labels. Noticeably, our MF can keep the highly informative integer feature encoded in the binary matrix. In theory, the above operation can be represented as:

$$\min_{\mathbf{Q}, \mathbf{R}} \mathcal{H}(\mathbf{U}, \mathbf{Q}\mathbf{R}^T) + \Delta(\mathbf{R}, \mathbf{Q}), \quad s.t., \quad \mathbf{Q} \in \{-1, 1\}, \quad (1)$$

where $\mathbf{R} \in \mathbb{R}^{c \times t}$ and $\mathbf{Q} \in \mathbb{R}^{n \times t}$ are respectively the image labels as well as aerial images distributed in the hidden space. \mathcal{J} calculate the error during the MF and $\Theta(\cdot)$ denotes some pre-defined regularizer to avoid overfitting. In practice, we notice that image labels \mathbf{T} are usually noisy. Obviously, label noises will cause unsatisfactory MF operation. In order to optimally tackle this problem, we proposed to derive a noise-free image label matrix \mathbf{M} from the potentially noisy label matrix. Theoretically, by inspecting the label matrix construction, element \mathbf{M}_{ij} denotes a binary indicator

reflecting the correlation between the labels associated with pairwise aerial images. Thereby, an upgraded optimized task obtained:

$$\begin{aligned} & \min_{\mathbf{M}, \mathbf{Q}, \mathbf{R}} \mathcal{H}(\mathbf{M}, \mathbf{Q}\mathbf{R}^T) + \mathcal{H}_l(\mathbf{M}, \mathbf{U}) + \Theta(\mathbf{R}, \mathbf{Q}), \\ & \text{s.t. } \mathbf{M} \in \{-1, 1\}, \mathbf{Q} \in \{-1, 1\}, \end{aligned} \quad (2)$$

where \mathcal{H}_l denotes the error of constructing these noise-free label matrix using the potentially noisy one.

It the graphlet hashing stage, we generally believe the significance of maintaining the local sample distribution [26], *e.g.*, the spatial relationships among spatially neighboring graphlets in the feature space. Meanwhile, we can derive the the hash function accordingly. Such hash function allows pairwise graphlets comparison highly scalable. Such hash function is employed to calculate different binary hash codes, that is, $\mathbf{i} = \text{sgn}(g(x)\mathbf{B})$. In summary, the below optimization can be obtained:

$$\begin{aligned} & \min_{\mathbf{I}, \mathbf{B}, \mathbf{g}} \beta \sum_{i=1}^n \mathcal{H}(\mathbf{i}^i, f(\mathbf{y}_i)\mathbf{B}) + \frac{\delta}{2} \sum_{i=1}^n \sum_{j=1}^n \mathbf{N}_{ij} \|\mathbf{c}^i - \mathbf{c}^j\|, \\ & \text{s.t. } \mathbf{I} \in \{-1, 1\}^{n \times M}, \end{aligned} \quad (3)$$

(3) can be updated into the corresponding matrix form, *i.e.*,

$$\begin{aligned} & \min_{\mathbf{I}, \mathbf{B}, \mathbf{g}} \beta \mathcal{H}(\mathbf{I}, g(\mathbf{Y}\mathbf{B})) + \delta \text{tr}(\mathbf{I}^T \mathbf{L}\mathbf{I}), \\ & \text{s.t. } \mathbf{I} \in \{-1, 1\}^{n \times M}, \end{aligned} \quad (4)$$

Herein, γ and δ denote two positive parameters reflecting the significance of the two terms accordingly. M is the Hamming space's dimension. $\mathbf{L} \in \mathbb{R}^{n \times n}$ represents the Laplacian matrix derived based on $\mathbf{L} = \mathbf{C} - \mathbf{L}$. Herein, \mathbf{C} denotes a diagonal matrix wherein each diagonal entity is calculated as $\mathbf{C}_{ii} = \sum_{j=1}^n \mathbf{N}_{ij}$. As the formulation in (4), the hash codes corresponding to aerial images as well as the hash function can be jointly calculated.

To obtain sufficiently compatible MF and the corresponding hash codes, it is naturally to assume that the inherent geometry revealed by our designed MF and the hashing have a shared feature space. In this way, the constructed latent space by our designed MF is the same to the aforementioned Hamming space. In this way, the potential semantics uncovered by the formulated MF with noise-free image labels is utilized to enhance the hashing model. Theoretically, we can set $\mathbf{I} = \mathbf{Q}$ and $M = t$. Therefore, we can obtain the below formulation as:

$$\begin{aligned} & \min_{\mathbf{M}, \mathbf{R}, \mathbf{I}, \mathbf{B}, \mathbf{g}} \mathcal{H}(\mathbf{M}, \mathbf{I}\mathbf{R}^T) + \mathcal{H}_l(\mathbf{M}, \mathbf{U}) + \beta \mathcal{H}(\mathbf{I}, f(\mathbf{Y}\mathbf{B})) \\ & \quad + \frac{\delta}{2} \text{tr}(\mathbf{I}^T \mathbf{L}\mathbf{I}), \\ & \text{s.t. } \mathbf{M} \in \{-1, 1\}^{n \times S}, \mathbf{H} \in \{-1, 1\}^{n \times L}, \end{aligned} \quad (5)$$

Herein, S denotes the category number of the UHD aerial images.

Notably, the optimizing task formulated above aims to drive the hash function as well as the binary codes by

leveraging the pre-constructed sample graph. Such graph is built upon the potentially contaminated image labels. This pre-constructed sample graph maintains intact in the learning stage. This is practically sub-optimal. Actually, we have to progressively adjust the sample graph in the hash codes learning. To this end, the sample graph updating module is also integrated into the learning model. Mathematically, to refine the possibly contaminated image labels, we expect the sample graph \mathbf{N} can be progressively updated during model learning. In the model, the similarities between each graphlet and the entire different ones sum to one, and $\mathbf{N}_{ii} = 0$. Thus, the mathematical formulation in (5) is reorganized as:

$$\begin{aligned} & \min_{\mathbf{M}, \mathbf{R}, \mathbf{I}, \mathbf{B}, \mathbf{N}, \mathbf{g}} \mathcal{H}(\mathbf{M}, \mathbf{I}\mathbf{R}^T) + \mathcal{H}_l(\mathbf{M}, \mathbf{U}) + \beta \mathcal{H}(\mathbf{N}, \mathbf{N}_0) \\ & \quad + \gamma \mathcal{H}(\mathbf{I}, g(\mathbf{Y}\mathbf{B})) + \frac{\delta}{2} \text{tr}(\mathbf{I}^T \mathbf{L}\mathbf{I}) + \Theta(\mathbf{R}, \mathbf{B}), \\ & \text{s.t. } \mathbf{M} \in \{-1, 1\}^{n \times S}, \mathbf{I} \in \{-1, 1\}^{n \times M}, \sum_{j=1}^n \mathbf{N}_{ij} = 1, \end{aligned} \quad (6)$$

During hash codes learning, the updating of the Laplacian matrix is $\mathbf{L} = \mathbf{A}(\mathbf{N} + \mathbf{N}^T)/2$. Herein, in the model initialization, \mathbf{N}_0 is computed by leveraging \mathbf{U} . And this objective function optimally incorporates hash codes calculation, semantic feature learning, and sample graph adjustment into an effective model.

To tackle (6), it is necessary to explicitly define \mathcal{H} , \mathcal{H}_l and Θ . In our implementation, we employ the least square loss $\mathcal{H}(a, b) = \frac{1}{2}(a - b)^2$. In order to maximally eliminate the noisy image labels, we have $\langle_l(c, d) = \mu|c - d|$. To obtain the regularization terms, it is common to have $\Theta(\mathbf{U}, \mathbf{V}) = \frac{\lambda}{2} \|\mathbf{U}\|_F^2 + \frac{\eta}{2} \|\mathbf{V}\|_F^2$. Totally, we can upgrade (6) as follows:

$$\begin{aligned} & \min_{\mathbf{M}, \mathbf{R}, \mathbf{I}, \mathbf{B}, \mathbf{N}} \frac{1}{2} \|\mathbf{M} - \mathbf{I}\mathbf{R}^T\| + \nu \|\mathbf{M} - \mathbf{U}\|_1 + \frac{\beta}{2} \|\mathbf{N} - \mathbf{N}_0\|_F^2 \\ & \quad + \frac{\gamma}{2} \|\mathbf{I} - f(\mathbf{Y}\mathbf{B})\|_F^2 + \frac{\delta}{2} \text{tr}(\mathbf{I}^T \mathbf{L}\mathbf{I}) + \frac{\mu}{2} \|\mathbf{R}\|_F^2 \\ & \quad + \frac{\theta}{2} \|\mathbf{B}\|_F^2 \\ & \text{s.t. } \mathbf{M} \in \{-1, 1\}^{n \times S}, \mathbf{I} \in \{-1, 1\}^{n \times M}, \sum_{j=1}^n \mathbf{N}_{ij} = 1, \end{aligned} \quad (7)$$

Obviously, (7) is a non-convex optimization. In our implementation, we propose to iteratively solve this objective function. The specific solutions are presented in the following link.² Moreover, following [30], we extend (7) into a deep learning architecture containing F layers.

C. KERNEL MACHINE FOR MULTI-LABEL UHD AERIAL PHOTO CATEGORIZATION

As aforementioned, many graphlets are extracted from each UHD aerial photo and are subsequently converted into binary hash codes. We observe that: 1) the graphlet numbers from different UHD aerial photos are generally

²<https://drive.google.com/file/d/1r6-b-mPKAcIuU6SAaAXyOneqvQIhF9W6/view?usp=sharing>

inconsistent; 2) the dimensionalities of binary hash codes calculated from graphlets with different vertices are different. Therefore, we cannot directly send them to a standard support vector machine (SVM) for visual categorization. In our implementation, a kernel-based quantizing method is deployed to calculate the feature at image-level, *i.e.*, fixed-length feature vector corresponding to a UHD aerial photo. For each UHD aerial photo, the BING [25]-based object patches are extracted to build graphlets, which are subsequently converted into binary hash codes using our DMCMF. Finally, graphlets within each UHD aerial photo are accumulated into vector $\mathbf{u} = \{u_1, u_2, \dots, u_A\}$, where A counts the training UHD aerial photos. Mathematically, we can calculate feature vector \mathbf{u} 's entity is computed as:

$$\mathbf{u}_i \propto \exp\left(-\frac{1}{AA'} \sum_{a=1}^A \sum_{b=1}^{A'} \text{dist}(\mathbf{h}_a, \mathbf{h}_b)\right), \quad (8)$$

where A and A' count the equally-sized graphlets from two UHD aerial photos respectively; $\text{dist}(\cdot, \cdot)$ computes the Jaccard similarity between binary hash codes.

By leveraging the vector quantized using (8), we train a SVM classifier for multi-label classification. In theory, to train an SVM distinguishing UHD aerial photos belonging to two different classes, the SVM can be mathematically represented in the following:

$$\begin{aligned} \max_{c \in \mathbb{R}^{N_{ab}}} \kappa(c) &= \sum_{i=1}^{N_{ab}} c_i - \frac{1}{2} \sum_{i=1}^{N_{ab}} \sum_{j=1}^{N_{ab}} d_i d_j t_i t_j k(\mathbf{v}_i, \mathbf{v}_j) \\ \text{s.t. } 0 &\leq d_i \leq D, \quad \sum_{i=1}^{N_{ab}} d_i t_i = 0, \end{aligned} \quad (9)$$

where \mathbf{v}_i denotes the calculated vector from each UHD aerial photo during training; t_i labels the i -th UHD aerial image; κ denotes the hyperplane separating samples from different categories; $D > 0$ trades the machine complexity off those mislabeled samples; and N_{ab} denotes the number of samples from the all the categories.

By calculating a quantized vector \mathbf{u} corresponding to a testing UHD aerial image, we can obtain the image labels is obtained as:

$$\text{sgn}\left(\sum_{j=1}^{N_{ab}} d_j t_j k(\mathbf{u}_i, \mathbf{u}^*) + \eta\right), \quad (10)$$

Herein, $\eta = 1 - \sum_{i=1}^{N_{ab}} d_i t_i k(\mathbf{u}_i, \mathbf{u}_k)$. Besides, \mathbf{u}_k represents the support vector corresponding to category label '+1'. In testing, we assign \mathbf{u}^* to the label set receiving the maximum number of votes.

IV. EXPERIMENTAL EVALUATIONS

Herein, we evaluate the performance of our UHD aerial photo categorization using three experiments. We first introduce our self-compiled data set, which includes 2.3 million UHD aerial images crawled from 100 well-known metropolises from different countries. Based on this, our algorithm is compared with 17 carefully-designed visual categorization models from three perspectives: accuracy and stability. Meanwhile, we carefully explain the high performance advantage of our classification model. Then, we carefully evaluate each

key module during UHD aerial image categorization. Lastly, we report our categorization accuracy of our method under different parameters. Based this, the optimal parameter settings are suggested.

After collecting the million-scale UHD aerial photos, we have to annotate them to obtain the corresponding image-level labels. Herein, 82 volunteers³ first manually annotate 14.7% UHD aerial photos in each metropolitan city, wherein a total of 47 different image-level labels were utilized. Afterward, we train a multi-label SVM and employ it to annotate the image-level labels of the rest UHD aerial images. Then, the same 82 volunteers manually correct the labels calculated by SVM. It is noticeable that multiple image-level labels are associated with intolerably small number of UHD aerial photos. This makes it infeasible to train a generalizable categorization model corresponding to these image-level labels. In our implementation, if the the number of UHD aerial photos corresponding to an image-level label is smaller than 200,000, Then we abandon this label. In this way, we finally obtain 18 different image-level labels. Thereafter, we notice that 99.973% UHD aerial photos have fewer than four image-level labels, while the rest very few UHD aerial photos have larger numbers of image-level labels (from five to 13). These UHD aerial photos usually contain a rich set of small regions ($< 200 \times 200$) that are possibly contaminated. Thus we simply abandon these UHD aerial photos. Lastly, we order the entire UHD aerial photos by their file names. For each category, we use the first half UHD aerial photos for training while the rest samples are for testing.

In retrospect, one key advantage of our method is to robustly learn a categorization model from noisy image-level labels. To acquire the noisy labels for experimentation, for each category, we randomly use 60% UHD aerial photos to construct a training set. Based on this, we learn a multi-label categorization model, which is further leveraged to calculate the labels of the entire UHD aerial photos. In total, there are 11.3% mislabeled UHD aerial photos. They are combined with those correctly labeled ones to constitute our data set.

We observe that each UHD aerial photo in our data set typically takes up 200MB of storage space. Therefore, our 2.3 million UHD aerial photos will require a total of 460TB storage space. To optimally store such million-scale UHD aerial photos for fast I/O interface, we employ the Supermicro server solutions.⁴ More specifically, we adopt the 4U double-sided super storage platform. The platform is installed with 36 Toshiba HDD drivers, each of which has a 20TB storage space. Totally, the entire storage space of our platform is 720TB and it works in RAID 0 mode. Based on this, the average sequential data reading and writing speeds are respectively 1467MB/s and 862MB/s on our storage platform. That means on average, it takes 0.137s and 0.232s to load and update each UHD aerial photo respectively.

³They are graduate students from our computer science department. They are aged between 24 and 31 and are experienced in image processing and pattern recognition. There are 51 males and 31 females in total.

⁴www.supermicro.com

TABLE 1. Accuracies with standard errors of the 18 categorization models. (We repeat each experiment 20 times and report the average accuracies and each bold number represents the best result.)

Category	[31]	[32]	[33]	[34]	[35]	[36]	[37]	SPP-CNN	CleanNet
Tall building	0.642±0.012	0.589±0.009	0.646±0.013	0.606±0.014	0.620±0.009	0.594±0.016	0.633±0.015	0.691±0.014	0.681±0.013
Residential	0.587±0.012	0.594±0.011	0.612±0.017	0.588±0.013	0.601±0.016	0.607±0.009	0.589±0.018	0.615±0.011	0.615±0.014
Intersection	0.703±0.014	0.715±0.012	0.694±0.016	0.685±0.014	0.716±0.019	0.684±0.017	0.721±0.010	0.684±0.013	0.695±0.012
Forest	0.684±0.013	0.673±0.014	0.698±0.013	0.664±0.014	0.682±0.012	0.658±0.014	0.685±0.012	0.713±0.011	0.705±0.014
Sea	0.674±0.013	0.647±0.015	0.684±0.015	0.633±0.013	0.665±0.017	0.646±0.018	0.673±0.013	0.662±0.013	0.686±0.010
Soccer field	0.546±0.014	0.565±0.016	0.587±0.013	0.577±0.016	0.583±0.009	0.562±0.014	0.584±0.012	0.570±0.021	0.583±0.018
Aircraft	0.732±0.012	0.704±0.014	0.721±0.013	0.695±0.015	0.705±0.013	0.718±0.017	0.685±0.015	0.716±0.014	0.705±0.013
Railway	0.621±0.014	0.613±0.016	0.635±0.013	0.643±0.015	0.607±0.015	0.596±0.016	0.605±0.014	0.614±0.017	0.616±0.015
Bridge	0.547±0.016	0.564±0.015	0.584±0.014	0.578±0.017	0.557±0.016	0.584±0.014	0.573±0.017	0.562±0.015	0.583±0.011
Road	0.613±0.013	0.624±0.012	0.635±0.014	0.615±0.016	0.625±0.013	0.621±0.014	0.605±0.016	0.616±0.013	0.627±0.014
River	0.721±0.015	0.708±0.017	0.716±0.010	0.716±0.014	0.726±0.013	0.699±0.013	0.702±0.015	0.709±0.016	0.715±0.019
Park	0.654±0.014	0.665±0.012	0.674±0.015	0.682±0.016	0.673±0.013	0.669±0.015	0.673±0.014	0.691±0.018	0.688±0.014
Palace	0.665±0.013	0.643±0.015	0.673±0.017	0.631±0.015	0.626±0.014	0.647±0.014	0.651±0.011	0.637±0.013	0.619±0.012
Factory	0.624±0.014	0.621±0.013	0.616±0.012	0.610±0.015	0.627±0.013	0.612±0.012	0.608±0.014	0.608±0.016	0.618±0.017
Farmland	0.604±0.013	0.602±0.016	0.608±0.012	0.598±0.016	0.584±0.014	0.614±0.013	0.592±0.015	0.609±0.018	0.611±0.016
Vehicle	0.685±0.009	0.674±0.013	0.658±0.011	0.694±0.015	0.653±0.012	0.668±0.014	0.670±0.016	0.684±0.014	0.671±0.014
Yacht	0.703±0.016	0.724±0.013	0.706±0.015	0.721±0.017	0.716±0.014	0.708±0.013	0.714±0.018	0.716±0.016	0.713±0.014
Swim. pool	0.654±0.014	0.636±0.012	0.641±0.015	0.652±0.013	0.633±0.016	0.665±0.011	0.673±0.015	0.631±0.013	0.636±0.018
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[38]	[39]	[40]	Ours
Tall building	0.625±0.013	0.664±0.014	0.659±0.012	0.682±0.016	0.673±0.014	0.625±0.014	0.642±0.016	0.647±0.014	0.704±0.009
Residential	0.594±0.014	0.614±0.013	0.618±0.012	0.624±0.015	0.613±0.014	0.576±0.015	0.597±0.016	0.588±0.014	0.667±0.008
Intersection	0.715±0.011	0.695±0.013	0.722±0.016	0.734±0.014	0.736±0.017	0.684±0.014	0.673±0.013	0.664±0.011	0.771±0.006
Forest	0.694±0.014	0.723±0.013	0.707±0.012	0.726±0.016	0.714±0.020	0.654±0.016	0.668±0.017	0.673±0.015	0.758±0.009
Sea	0.674±0.015	0.645±0.013	0.658±0.016	0.673±0.013	0.657±0.012	0.671±0.016	0.663±0.013	0.675±0.014	0.697±0.009
Soccer field	0.584±0.016	0.567±0.015	0.594±0.014	0.585±0.016	0.583±0.014	0.557±0.013	0.563±0.018	0.559±0.014	0.617±0.011
Aircraft	0.685±0.013	0.684±0.021	0.705±0.023	0.722±0.015	0.728±0.017	0.675±0.013	0.687±0.017	0.693±0.018	0.752±0.006
Railway	0.624±0.014	0.632±0.015	0.617±0.013	0.606±0.017	0.625±0.015	0.607±0.014	0.611±0.016	0.603±0.013	0.686±0.009
Bridge	0.564±0.015	0.547±0.017	0.568±0.013	0.574±0.016	0.536±0.017	0.530±0.014	0.543±0.013	0.532±0.016	0.595±0.007
Road	0.612±0.018	0.615±0.015	0.604±0.016	0.642±0.014	0.633±0.020	0.610±0.017	0.606±0.012	0.615±0.017	0.688±0.008
River	0.724±0.015	0.714±0.014	0.721±0.016	0.718±0.017	0.715±0.013	0.675±0.015	0.663±0.016	0.684±0.018	0.759±0.008
Park	0.674±0.015	0.663±0.017	0.690±0.018	0.684±0.014	0.684±0.016	0.694±0.017	0.682±0.015	0.683±0.014	0.704±0.006
Palace	0.624±0.011	0.631±0.023	0.614±0.018	0.621±0.019	0.635±0.017	0.596±0.016	0.604±0.015	0.609±0.014	0.685±0.007
Factory	0.614±0.015	0.608±0.016	0.612±0.012	0.607±0.017	0.614±0.015	0.603±0.017	0.615±0.013	0.613±0.012	0.665±0.005
Farmland	0.594±0.014	0.592±0.016	0.587±0.013	0.612±0.016	0.617±0.014	0.585±0.013	0.597±0.012	0.603±0.015	0.627±0.007
Vehicle	0.654±0.016	0.685±0.016	0.675±0.017	0.646±0.014	0.686±0.016	0.639±0.014	0.654±0.017	0.673±0.016	0.706±0.006
Yacht	0.724±0.015	0.720±0.021	0.716±0.018	0.714±0.016	0.718±0.017	0.709±0.014	0.706±0.018	0.724±0.013	0.778±0.006
Swim. pool	0.621±0.016	0.654±0.014	0.643±0.017	0.657±0.015	0.626±0.014	0.607±0.013	0.614±0.013	0.628±0.016	0.681±0.007

1) ACCURACY COMPARISON

In this section, we evaluate our UHD aerial photo categorization framework by comparing its effectiveness and efficiency with a rich set of baseline recognition algorithms. We first test our algorithm by comparing it with deep aerial image classification models. Thereafter, we employ state-of-the-art deep generic visual categorization algorithms for comparison.

First of all, we compare our method with seven deep visual categorization models [31], [32], [33], [34], [35], [36], [37] that intrinsically incorporate some prior knowledge of different categories of aerial photos. We notice that the source codes of [31], [32], [35], and [36] are publicly available. Based on this, we conduct comparative study wherein the parameter settings are set as default. For [33], [34], and [37], we implement them since the source codes are unavailable. We have tried our best to make the re-implemented categorization models perform similarly to the results reported in their publications.

Meanwhile, many recent deep generic visual recognition models perform impressively on categorizing aerial images. Herein, we first compare our method with ten deep generic object classification algorithms. Moreover, since UHD aerial photo categorization can be considered as a sub-topic of scenery classification, we further conduct a comparative study between our method and three recently published scene classification models [38], [39], [40]. For the categorization models implemented by us, the experimental setups can be summarized in the following. For [33], we utilize the

ResDep-128 [41] to function as the backbone. This is further updated into the multi-label variant. Different from the fully-connected layer (unit number is set to 19), the rest deep layers are fixed by the above ResDep-128 [42]. The ResNet-108 [41] is employed as the backbone and the stochastic gradient descent optimizes the entire network. The learning ratio as well as the decay are respectively fixed to 0.001 and 0.05. The network loss is calculated by the mean squared error. For [38], we retrain the object bank [43] by leveraging our refined 18 UHD aerial photo categories, wherein the average-pooling strategy is applied. We employ the bilinear as the solution to the linear classifier, wherein the 7-fold cross evaluation is applied.

For the above 18 compared object/scene recognition algorithms, we repeatedly test each model 20 times. Accordingly, the averaged accuracies are displayed in Table 1. To quantify the stability of these categorization models, we report their standard errors simultaneously. We observe that the per-category standard errors produced by our method are significantly and consistently lower than its competitors. This demonstrated that our method is the most stable. In summary, the following conclusions can be made:

- Our method outperforms the other seven aerial photo categorization models remarkably due to three reasons. First, these compared methods typically characterize low/medium resolution aerial photos. To facilitate deep model training, they generally resize the original aerial photo to a fixed and much smaller size (e.g. 224×224) for the subsequent deep modeling. This

operation is negative to learning an effective UHD aerial photo categorization model since those tiny but discriminative visual details will be lost. Second, expect for our method, none of the seven counterparts can implicitly correct the noisy image-level labels, which will inevitably hurt the categorization model training. Third, only our method uses graphlets to explicitly capture the complicated spatial layouts of each UHD aerial photo. They are further incorporated by a deep hashing algorithm for calculating the discriminative image kernel. Comparatively, the seven counterparts only globally/locally characterize each UHD aerial photo, wherein the informative spatial layouts among multiple aerial photo regions are neglected.

- The seven generic object recognition algorithms perform inferiorly than our method because of three reasons. First, these generic recognition models generally handle medium sized images typically containing under ten million pixels. They can hardly discover the tiny but discriminative regions from the hundreds of object components inside an UHD aerial photo with over 100 million pixels. This case is particularly worse when the image-level labels are contaminated. Second, our method can conveniently incorporate some prior knowledge of UHD aerial photo set, *e.g.*, the maximum graphlet size and the category-specific object patches. Contrastively, the seven generic object recognition models cannot encode the domain knowledge reflecting UHD aerial photos. Third, by leveraging our noise-tolerant hashing algorithm, only our method allows a fast and accurate comparison of many discriminative object parts between UHD aerial photos. Nevertheless, the seven generic object recognition models simply convert each UHD aerial photo into a long feature vector for deep classification. They cannot achieve such precise region-to-region comparison like ours.
- The three scene categorization models perform unsatisfactorily on UHD aerial photos. This is because they deeply and implicitly learn a descriptive set of scene-aware semantic categories, such as “birds” and “tables”, which usually infrequently appear in our UHD aerial photo set. Moreover, the three categorization methods can successfully handle sceneries captured at horizontal view angles. But our UHD aerial photos are captured at overhead view angles. Apparently, such view angle gap will largely hurt the categorization accuracy.

A. ABLATION STUDY

The two key modules in our work are the DMCMF and kernel-induced feature quantization. Herein, the effectiveness of the two modules are evaluated in our designed categorization pipeline. Each module is changed into a degraded and the performances are recorded accordingly. Meanwhile, insights are provided to elaborate the underlying reasons for the received results.

TABLE 2. Categorization accuracy drop (“-”) and improvement (“+”) in our ablation study.

	S2	S3
O1	-6.546%	-7.446%
O2	-3.432%	-2.354%
O3	-4.522%	-1.765%
O4	-6.412%	Unavailable

First of all, we test our key theoretical contribution, the proposed DMCMF. Specifically, we analyze the four functional components as formulated in (7). The label noise refinement component is first abandoned (S11). Mathematically, the term $\nu\|\mathbf{M} - \mathbf{U}\|_1$ is removed and we update \mathbf{L} into \mathbf{T} . Afterward, the data graph updating term $\frac{\beta}{2}\|\mathbf{N} - \mathbf{N}_0\|_F^2$ is abandoned, wherein the remaining components keep intact (S12). Then, the binary hash codes constraint is removed and we maintain the rest terms unchanged (S13). Last but not least, the hierarchical feature engineering term is reduced to a flat one (S24) by setting $F = 1$. The results in Table 2 have shown that, label noise refinement and hierarchical feature learning models play the most important roles. This is because removing each will cause an $> 6.4\%$ classification accuracy drop. Moreover, abandoning the limitation of binary codes will bring a 4.522% accuracy decrement. Even worse, the time consumed at the test stage significantly increased by over seven times. This clearly shows the effectiveness and efficiency of adopting binary codes to characterize UHD aerial photos.

Lastly, to demonstrate the usefulness of the kernel-based quantized vector calculated from each UHD aerial photo, the following experimental setups are applied. We first use the aggregation-based deep network that accumulates the predicted category labels corresponding to the entire graphlets within an UHD aerial photo. These labels are subsequently combined into the final image-level category label (S31). Thereafter, we replace our adopted linear kernel by polynomial kernel (S32) and Gaussian radial basis function (RBF) (S33) respectively. As shown in Table 2, aggregating the graphlet-level category label severely hurts the categorization accuracy. This is because calculating the category label at graphlet-level is sometimes obscure and misleading. In practice, each graphlet occupies very few regions within each UHD aerial photo, and some regions correspond to the background areas irrelevant to a particular category. Besides, both polynomial and RBF kernels perform inferiorly than our linear kernel. This observation demonstrates that projecting the quantized vectors onto a linear space can better separate UHD aerial photos from different categories.

B. PERFORMANCE BY VARYING PARAMETERS

In our work, we have multiple tunable parameters that will be evaluated. The first set denotes the weights balancing different clues in the DMCMF framework. The second set contains parameters influencing deep topological feature engineering. In this experiment, we test the UHD aerial image classification accuracy using different parameter settings.

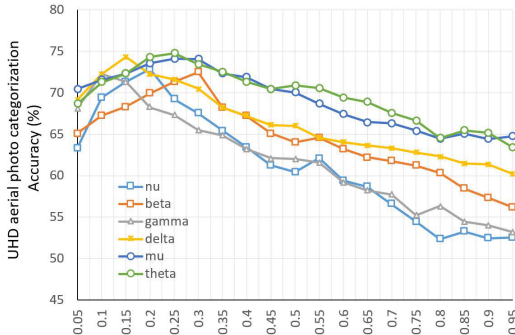


FIGURE 3. UHD aerial photo categorization accuracies by varying the six parameters in the first set.

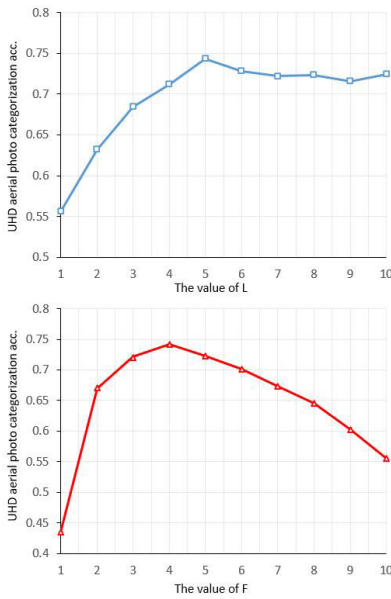


FIGURE 4. Recognition accuracies by varying L (top) and F (bottom) respectively.

To analyze the first parameter set, the default values of ν , β , γ , δ , μ and θ are set to 0.2, 0.3, 0.1, 0.15, 0.3, 0.3 respectively. In our implementation, the default values are determined by 10-fold cross validation. Herein, the validation set contains 18000 samples, which is constituted by selecting 1000 UHD aerial photos from each category. More specifically, we tune each parameter from 0.05 to one with a step of 0.05. And all the possible parameter combinations are enumeratively employed to test the UHD aerial photo categorization accuracy. The parameter combination receiving the highest categorization accuracy is preserved as the default values. Based on this, we adjust one of the five parameters while keep the others unchanged. Each parameter is increased from 0.05 to one with step of 0.01, wherein the corresponding performance is reported. As the six curves displayed in Fig. 3, the six parameters consistently increase stably and then peak. Afterward, they all decrease to a low level. Such monotonicity properties indicate the feasibility to tune the six parameters toward an optimal level in practice.

Next, we evaluate the UHD aerial photo categorization by changing the L (maximum size of graphlet) and F (deep

layer number). For both L and F , we tune them from one to ten and record the corresponding recognition performance. In Fig. 4, when increasing L , the categorization accuracy increases sharply if $L \in [1, 5]$ and then keeps stable when $L > 5$. Meanwhile, we notice that when L goes up, the time and storage costs increase dramatically since more graphlets will be generated. Toward an efficient and effective UHD aerial photo categorization system, we set $L = 5$. Moreover, we observe that the highest categorization accuracy is achieved when there are four deep layers. To our best knowledge, too few deep layers will make the deeply-learned binary hash codes insufficiently discriminative. Meanwhile, too many deep layers will increase the number of deep model parameters, which inevitably causes deep model overfitting.

V. CONCLUSION

Aerial image understanding is an indispensable technique in pattern recognition [44], [45], [46], [47], [48]. We propose a novel deep matrix factorization that optimally fuzes multiple clues into a solvable optimization for multi-label UHD aerial photo categorization. We first extract the BING [25]-based patches describing objects or their parts. Then, multiple graphlets are built to capture the spatial configurations of the ground salient objects that are visually/semantically salient. Afterward, we propose a so-called DMCMF that effectively encodes image labels to improve our binary hashing. Lastly, the binary feature vectors are integrated into a kernel SVM to label each UHD aerial image to multiple categories. Comprehensive experiments on the collected UHD aerial image set reflected the our algorithm’s advantage.

REFERENCES

- [1] S. Zhou, J. Irvin, Z. Wang, E. Zhang, J. Aljbran, W. Deadrick, R. Rajagopal, and A. Ng, “DeepWind: Weakly supervised localization of wind turbines in satellite imagery,” in *Proc. CVPR*, 2009, pp. 1–5.
- [2] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, and R. Ji, “Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning,” *Pattern Recognit.*, vol. 64, pp. 417–424, Apr. 2017.
- [3] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, “Joint inference of groups, events and human roles in aerial videos,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4576–4584.
- [4] J. Porway, Q. Wang, and S. C. Zhu, “A hierarchical and contextual model for aerial image parsing,” *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 254–283, 2010.
- [5] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, “Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4095–4104.
- [6] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, “Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.
- [7] R. Kemker, C. Salvaggio, and C. Kanan, “Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning,” *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.
- [8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.
- [9] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, and Q. Tian, “Discovering discriminative graphlets for aerial image categories recognition,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5071–5084, Dec. 2013.

- [10] Y. Xia, L. Zhang, Z. Liu, L. Nie, and X. Li, "Weakly supervised multimodal kernel for categorizing aerial photographs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3748–3758, Aug. 2017.
- [11] Ö. Akar, "Mapping land use with using rotation forest algorithm from UAV images," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 269–279, Jan. 2017.
- [12] M. I. Sameen, B. Pradhan, and O. S. Aziz, "Classification of very high resolution aerial photos using spectral-spatial convolutional neural networks," *J. Sensors*, vol. 2018, pp. 1–12, Jun. 2018.
- [13] G. Cheng, C. Ma, P. Zhou, X. Yao, and J. Han, "Scene classification of high resolution remote sensing images using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 767–770.
- [14] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, "More diverse means better: Multimodal deep learning meets remote-sensing imagery classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, Aug. 2020.
- [15] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
- [16] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, p. 516, Feb. 2021.
- [17] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 310–314, Feb. 2019.
- [18] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Deep learning for vehicle detection in aerial images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3079–3083.
- [19] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," 2016, *arXiv:1605.08323*.
- [20] Y. Yu, X. Yang, J. Li, and X. Gao, "Object detection for aerial images with feature enhancement and soft label assignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3177255.
- [21] J. Wang, F. Li, and H. Bi, "Gaussian focal loss: Learning distribution polarized angle prediction for rotated object detection in aerial images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 3175520.
- [22] V. Chalavadi, P. Jeripothula, R. Datla, S. B. Ch, and K. M. C., "MSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108548.
- [23] J. Kang, S. Tariq, H. Oh, and S. S. Woo, "A survey of deep learning-based object detection methods and datasets for overhead imagery," *IEEE Access*, vol. 10, pp. 20118–20134, 2022.
- [24] F. van Ede, S. R. Chekroud, and A. C. Nobre, "Human gaze tracks the focusing of attention within the internal space of visual working memory," *J. Vis.*, vol. 19, no. 10, p. 133, Sep. 2019.
- [25] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. S. Torr, "BING: Binarized normed gradients for objectness estimation at 300fps," *Comput. Vis. Media*, vol. 5, no. 1, pp. 3–20, Mar. 2019.
- [26] L. Zhang, C. Chen, J. Bu, D. Cai, X. He, and T. S. Huang, "Active learning based on locally linear reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 10, pp. 2026–2038, Oct. 2011.
- [27] R. Diestel, *Graph Theory*. Berlin, Germany: Springer-Verlag, 2005.
- [28] M. Stricker and M. Orengo, "Similarity of color images," *Storage Retr. Image Video Databases*, vol. 2420, pp. 381–392, 1995.
- [29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05)*, Jun. 2005, pp. 886–893.
- [30] Z. Li and J. Tang, "Weakly-supervised deep nonnegative low-rank model for social image tag refinement and assignment," in *Proc. Conf. Artif. Intell. (AAAI)*, Feb. 2017, pp. 4154–4160.
- [31] C. Kyrkou and T. Theodoridis, "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 7, pp. 1687–1699, Mar. 2020.
- [32] C. Kyrkou and T. Theodoridis, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 517–525.
- [33] Y. Hua, S. Lobry, L. Mou, D. Tuia, and X. X. Zhu, "Learning multi-label aerial image classification under label noise: A regularization approach using word embeddings," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 525–528.
- [34] Y. Hua, L. Mou, and X. X. Zhu, "Multi-label aerial image classification using a bidirectional class-wise attention network," in *Proc. Joint Urban Remote Sens. Event (JURSE)*, May 2019, pp. 1–4.
- [35] M. D. Pritt and G. Chern, "Satellite image classification with deep learning," 2020, *arXiv:2010.06497*.
- [36] H. Sun, Y. Lin, Q. Zou, S. Song, J. Fang, and H. Yu, "Convolutional neural networks based remote sensing scene classification under clear and cloudy environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 713–720.
- [37] S. Song, H. Yu, Z. Miao, Q. Zhang, Y. Lin, and S. Wang, "Domain adaptation for convolutional neural networks-based remote sensing scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1324–1328, Aug. 2019.
- [38] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorizations," in *Proc. PRAM*, 2015, pp. 209–224.
- [39] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: Objects, scales and dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 571–579.
- [40] Y. Li, M. Dixit, and N. Vasconcelos, "Deep scene image classification with the MFAFVNet," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5757–5765.
- [41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [43] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1–9.
- [44] Z. He and Z. Xiong, "Research on pattern matching of dynamic sustainable procurement decision-making for agricultural machinery equipment parts," *IEEE Access*, vol. 11, pp. 1–17, 2023.
- [45] Y. Shimizu, "Efficiency optimization design that considers control of interior permanent magnet synchronous motors based on machine learning for automotive application," *IEEE Access*, vol. 11, pp. 41–49, 2023.
- [46] H. Zhang, C. Ma, Z. Jiang, and J. Lian, "Image caption generation using contextual information fusion with Bi-LSTM-s," *IEEE Access*, vol. 11, pp. 134–143, 2023.
- [47] V. Dammins, W. Panup, and R. Wangkeeree, "Laplacian twin support vector machine with pinball loss for semi-supervised classification," *IEEE Access*, vol. 11, pp. 31399–31416, 2023.
- [48] W. Mu and B. Liu, "Voice activity detection optimized by adaptive attention span transformer," *IEEE Access*, vol. 11, pp. 31238–31243, 2023.

JUNWU ZHOU is currently a Lecturer with Shanghai Dianji University. His research interests include big data, machine learning, computer vision, and image processing.

GUIFENG WANG is currently an Associate Professor with Shanghai Dianji University. His research interests include robotics, mechanical manufacturing, image parsing, and image rendering.

FUJI REN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Engineering, Hokkaido University, Sapporo, Japan, in 1991. From 1991 to 1994, he was a Chief Researcher with CSK. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, Hiroshima, Japan, as an Associate Professor. Since 2001, he has been a Professor with the Faculty of Engineering, Tokushima University, Tokushima, Japan.

•••