IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 August 2023, accepted 6 September 2023, date of publication 10 October 2023, date of current version 17 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323317

RESEARCH ARTICLE

Perceptual Low-Rank Learning and Geometry-Preserving Feature Selection for Categorizing High-Resolution Aerial Photos

JUNWU ZHOU[1] AND FUJI REN[2], (Senior Member, IEEE)
[1]Higher Vocational and Technical College, Shanghai Dianji University, Shanghai 201306, China
[2]School of Computer Sciences, Anhui University, Hefei 230039, China
Corresponding author: Junwu Zhou (shzhjwu@163.com)

ABSTRACT Recognizing the multiple categories of an high-resolution (HR) aerial photos is an indispensable technique in geoscience and remote sensing. In this work, a perceptual low-rank algorithm combined with a geometry-preserving feature selection (FS) is proposed for categorizing HR aerial photos. In practice, the theory of human visual perception indicates that for each scenery, the background non-salient regions are highly correlated, whereas the foreground visually/semantically salient regions are almost uncorrelated. Motivated by this, we design a novel low-rank algorithm that seeks a sparse set of foreground visually/semantically salient image patches. These patches are sequentially linked into a so-called GSP (path reflecting gaze movement) to mimick human vision system. Afterward, a geometry-preserving FS algorithm is proposed to select highly discriminative features from the aforementioned gaze features, wherein a classifier can be trained simultaneously. Comprehensive experimental validation on our Internet-scale image set have shown its superiority.

INDEX TERMS Feature selection, geometry, high-resolution, low-rank.

I. INTRODUCTION

Thanks to the technology of delivering several satellites by a single rocket launch, many earth observation satellites have been launched in the past decades. These satellites capture HR aerial images containing ground objects with sophisticated spatial structures. Understandings the semantics of the ground objects by exploiting the inherent spatial structures becomes a useful tool in lots of artificial intelligence applications.

In image processing, plenty of image/video classification/parsing models were designed to encode aerial photos. Important work includes: 1) multiple instance learning/convolutional neural network-based object localization using weak labels; 2) graph model for semantically exploiting aerial photographs; and 3) well-designed deep models to semantically annotate aerial photographs. Nevertheless, as far as we know, the existing models are all sub-optimal characterize HR aerial photo because of the following reasons:

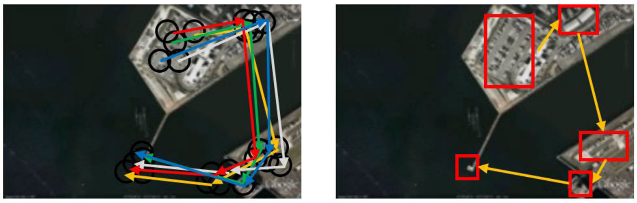


FIGURE 1. Left: human gaze shifting paths (GSPs) from five observers (arrows with different colors); right: a GSP calculated using the proposed method.

- Actually, each HR photo usually has many ground objects with complicated spatial layouts. Intelligently exploiting their underlying semantics is difficult. The inherent challenges include: i) discovering those ground visually/semantically salient objects according to human visual perception (as the circles exemplified on the left of Fig. 1), and ii) how to design a model that converts

The associate editor coordinating the review of this manuscript and approving it for publication was Mohammad Shorif Uddin.

the discovered salient objects into a fixed-length feature vector, which can be utilized for the subsequent feature classification;

- Biological studies have shown that humans sequentially perceive different regions within each scenery. As shown on the right of Fig. 1, humans will first attend to the upper central residential area, and then shift the gazes to the right one, and so on. In practice, the path reflecting human gaze movement is highly descriptive to categorize HR aerial photos. But designing a principled model extracting GSPs from HR aerial photos with different spatial layouts remains unsolved.

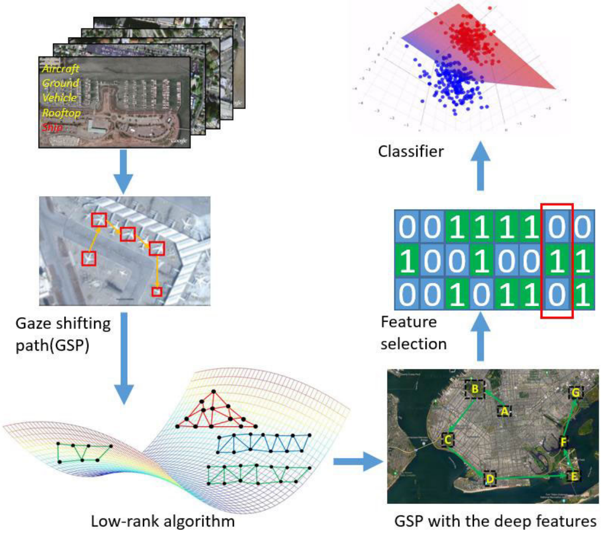


FIGURE 2. An overview of our proposed HR aerial photo categorization pipeline.

To handle these challenges, we propose a biologically-inspired HR aerial photo categorization framework. Our key contribute is a low-rank algorithm associated with a geometry-preserving feature selector that jointly: 1) extracts multiple visually/semantically salient image patches sequentially to constitute a GSP for each HR aerial photo, and 2) obtain a subset of highly discriminative GSP features for the subsequent visual categorization. More specifically, as elaborated in Fig. 2, by collecting a considerable quantity of HR aerial images, we first project their internal regions onto the feature space constructed by exploiting the visual and semantic channels collaboratively. Thereafter, to mimick human visual perception, a low-rank model is designed to decompose each HR aerial photo into a sequence of visually/semantically salient foreground image patches coupled with the non-salient background ones. Accordingly, the saliency value of each salient image patch can be calculated, which guides the GSP feature generation. Toward a subset of high quality GSP features, we further propose a geometry-preserving FS algorithm to obtain highly discriminative GSP features, coupled with a classifier trained from the selected GSP features. Noticeably, our designed feature selector can maximally preserve the sample distribution in the feature space during FS. This attribute is significant to ensure the discrimination of the selected features according to the manifold learning theory [44]. Such classifier is finally utilized to calculate the category labels of each HR aerial photo. Experimental comparison with over ten state-of-the-art shallow/deep categorization models has demonstrated the superiority of the proposed approach.

In summary, our method has two main contributions: 1) a novel low-rank algorithm that extracts many GSPs from each HR aerial photo and engineers GSP's visual feature simultaneously, and 2) a geometry-preserving FS algorithm to obtain highly discriminative GSP features and train the classifier for HR aerial photo categorization.

II. RELATED WORK

In the literature, dozens of computational aerial image models were developed to analyze aerial photos.[1] Some models are conducted at image-level. Zhang et al. [2] 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. [3] formulated a novel weak model which can semantically label HR aerial photos at image-level. Akar et al. [4] seamlessly combined the so-called rotation forest and object-level feature extractor to categorize a rich set of aerial images to different classes. The authors [5] developed a hierarchical deep architecture to recognize the multiple labels of HR aerial photos describing many downtown areas. In [6], researchers utilized a hierarchical and multi-layer deep model to classify HR aerial photos. A domain-specific scenic picture set is leveraged to fine tune the deep architecture. In [7], 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. Cai and Wei [8] proposed a cross-attention mechanism to learn the weights of aerial image features both horizontally and vertically. In [9], Bazi et al. formulated a vision transformer for aerial image classification, wherein the long-term contextual dependencies among regions can be intrinsically encoded. Although impressive performance have been achieved by the above methods, they cannot handle HR aerial photo categorization effectively because of three reasons: 1) the region-level visual features are particularly informative for HR aerial photo modeling, but they cannot be well encoded; 2) these methods cannot explicitly incorporate human gaze behavior into the categorization model. Thus, the predicted semantic labels might be inconsistent with human visual cognition; and 3) these methods are usually insufficiently fast since a rich number of highly time-consuming features have to be extracted.

For region-level modeling, the authors [10] designed an enhanced and multi-layer neural network to discover

[1] A more comprehensive survey of aerial photo understanding is illustrated in [1].

multi-scale attractive objects within an aerial image. In [11], a focal loss deep architecture is proposed that optimally discovers vehicles from aerial images. In [12], researchers developed a novel object localization algorithm toward remote sensing images. It intelligently extracts intersections as well as streets. In [13], Yu et al. integrated feature enhancement and soft label assignment into an anchor-independent object detector toward aerial images. In [14], Wang et al. proposed a deep rotation-invariant detector that effectively estimates the angles of multi-scale objects inside aerial images. In [15], 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, compared to image-level modeling, region-level models can exploit the regional features to facilitate HR aerial photo categorization. But there still some shortcomings: 1) the aforementioned region-level models are generally dataset-independent, which cannot be conveniently applied cross different datasets. Practically, however, we need a principled region-level image model that is applicable across multiple image sets; and 2) the human visual perception fails to be efficiently encoded by these models. Actually, we want an HR aerial photo processing system that can rapidly recognize each HR aerial photo.

In machine learning, the low-rank algorithm [16] has been pervasively used in seeking a succinct set of bases for representing a large-scale samples, *i.e.*, each sample can be represented by a linear combination of the bases. Low-rank algorithm can be used in applications like information retrieval, recommendation systems [17], and feature extraction [17]. In our work, we use low-rank approximation to represent the entire regions within each aerial image by a set of visually/semantically salient regions. This can be deemed as a novel visual feature extractor. Meanwhile, geometry-preserving feature selection (FS) attempts to obtain a few highly discriminative features from the original high-dimensional ones. During the FS process, the geometry distribution among samples is maximally preserved. This technique is widely used for face/speech recognition [18] and image retrieval [19]. Typically, geometry-preserving FS can significantly enhance AI systems' efficiency by reducing the number of extracted features.

III. OUR PROPOSED METHOD

A. LOW-RANK ALGORITHM FOR GSP LEARNING

In practice, there are multiple fine-grained objects inside each HR aerial photo. Biological studies [20], [21], [41] have shown that observers practically attend to a succinct set of salient objects. In our scenario, when humans perceive an LR aerial photo, their eye will first fix onto the ground attractive regions. Meanwhile, the unattractive background regions are kept almost unprocessed. Such human visual perceptual behavior is informative for categorizing HR aerial photos. Herein, we propose a low-rank algorithm that sequentially selects salient image patches to construct gaze shifting paths (GSPs). And the corresponding visual features can be jointly engineered.

The theory of human visual perception indicates the high correlation (self-representativeness) of the non-salient background image patches inside each scenery. Contrastively, the foreground salient image patches are almost uncorrelated. This observation motivates us to decompose the feature matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ of each HR aerial photo into the salient and non-salient parts,

$$\mathbf{X} = \mathbf{Y} + \mathbf{E}, \tag{1}$$

where N counts the image patches within each HR aerial photo and T its feature dimensionality. $\mathbf{Y} \in \mathbb{R}^{T \times N}$ preserves feature columns corresponding to the non-salient background image patches (the other columns are all zeros). $\mathbf{E} \in \mathbb{R}^{T \times N}$ represents feature columns corresponding to the salient image patches (the other columns are all zeros).

Aiming at a unique solution indicating the salient image patches, some criteria are proposed to constrain $\mathbf{Y}$ and $\mathbf{E}$. In our work, two observations are made. First, only a small fraction of image patches within each HR aerial photo are salient and will the detailedly processed by human vision system. This mathematically reflects that $\mathbf{E}$ is a sparse matrix. Second, the high correlation of the non-salient background image patches indicates that $\mathbf{Y}$ is a low-rank matrix. Based on these, we select the salient image patches by seamlessly integrating a sparsity and low-rankness constraint into (1):

$$\min_{\mathbf{Y},\Omega} ||\mathbf{Y}||_* + \alpha l_1(\mathbf{E}) + \beta l_2(\mathbf{Y}, f(\Gamma, \mathbf{X})) + \gamma \Omega(\Gamma), \tag{2}$$

where $||\cdot||_*$ is the matrix nuclear norm representing a convex approximation to matrix rank function, $l_1(\mathbf{E})$ quantizes the sparisty of $\mathbf{E}, f(\Gamma, \mathbf{X}))$ selects non-salient background image patches from each HR aerial photo and Γ contains the inherent parameters, and $l_2(\mathbf{Y}, f(\Gamma, \mathbf{X}))$ penalizes the loss of non-salient background image patches selection. $\Omega(\Gamma)$ serves as a regularizer. α, β, and γ are parameters measuring the importance of these terms. More concretely, to ensure a highly sparse $\mathbf{E}$, $l_1(\cdot)$ is defined as:

$$l_1(\mathbf{E}) = ||\mathbf{E}||_1, \tag{3}$$

Noticeably, each entity of $\mathbf{Y}$ is nonnegative. Herein, we set $l_2(a, b) = (a-b)^2/2$ to calculate the image patches selection error. Thereby, objective function (2) can be upgraded into:

$$\min_{\mathbf{Y},\Omega} ||\mathbf{Y}||_* + \alpha||\mathbf{E}||_1 + \beta||\mathbf{Y} - f(\Gamma, \mathbf{X})||_F^2 \\ + \gamma\Omega(\Gamma), \quad \mathbf{Y} \geq 0. \tag{4}$$

It is observable that (4) is a non-convex optimization over the entire variables. In our implementation, we follow the iterative algorithm in [42] to solve it. Thereafter, denoting $\mathbf{Y}^*$ as the optimal solution of (4), the saliency score of the i-th image patch in an HR aerial photo is calculated by:

$$s(\mathbf{X}_i) = ||\mathbf{E}^*(:, i)||_2, \tag{5}$$

where $\mathbf{E}^* = \mathbf{X} - \mathbf{Y}^*$, and $\mathbf{E}^*(:, i)$ denotes the i-th column of $\mathbf{E}^*$.

The GSP learning is given as follows. A larger $s(\mathbf{X}_i)$ in Eq.(5) means that the i-th image patch is more visually/semantically salient. Given an HR aerial photo, we sequentially link the top P salient image patches to constitute its gaze shifting path (GSP). Accordingly, the visual feature of the GSP is obtained by sequentially concatenating the visual features of its constituent P image patches. In the following, the GSP feature from the i-th training HR aerial photo is denoted by g_i.

B. GEOMETRY-PRESERVING FEATURE SELECTION

1) SAMPLES GEOMETRY PRESERVATION[2]

Inspired by the recent progresses in manifold learning, the self-expressive model is leveraged to preserve the sample distribution during feature selection (FS). The self-expressive model hypothesizes that the entire samples are distributed on a combination of subspaces. Mathematically, each sample can be linearly represented by a constrained combination of the other samples, *i.e.* $\mathbf{G} = \mathbf{GT}$ and $\mathrm{diag}(\mathbf{T}) = 0$. Herein, $\mathbf{G} = [g_1, \cdots, g_N]$ is a matrix consisting of the GSP features from N training samples, $\mathbf{T}$ denotes the matrix containing the self-reconstruction parameters. Practically, we notice that the samples might be contaminated. Thus, the self-expressive model can be upgraded into: $\mathbf{G} = \mathbf{GT} + \mathbf{J}$, wherein $\mathbf{J}$ is the error matrix. Based on these, the general form self-expressive model is given as:

$$\min_{\mathbf{T},\mathbf{J}} \tau_1||\mathbf{T}||_u + \tau_2||\mathbf{J}||_v, \quad s.t. \ \text{constraints}(\mathbf{T},\mathbf{J}), \tag{6}$$

where $||\cdot||_u$ and $||\cdot||_v$ denote two pre-specified matrix norms, τ_1 and τ_2 are the two corresponding nonnegative weights, and $\text{cons}(\mathbf{T},\mathbf{J})$ represents the constrains on $\mathbf{T}$ and $\mathbf{J}$.

2) OBJECTIVE FUNCTION OF OUR FS

We denote $\mathbf{K}$ as a matrix projecting the original GSP features into the low-dimensional one. In practice, $\mathbf{K}$ is constrained to be a column-wise sparse matrix for FS. Then, we can assume that if sample g_i and g_j are from the same category, then the low-dimensional selected feature $\mathbf{K}g_i$ and $\mathbf{K}g_j$ should be close and the weight $\mathbf{H}_{ij}$ should be large. Herein, $\mathbf{H} = |\mathbf{T}| + |\mathbf{T}^T|$ denote the weight matrix measuring the entire samples. Meanwhile, if samples g_i and g_j are from different categories, then the distance between $\mathbf{K}g_i$ and $\mathbf{K}g_j$ should be far and the weight $\mathbf{H}_{ij}$ will be close to zero. Mathematically, the above observations can be formulated into the following objective function:

$$\min_{\mathbf{K},\mathbf{L},\mathbf{H}} \sum\nolimits_{i=1}^{N} \sum\nolimits_{j=1}^{N} \mathbf{H}_{ij}(\alpha||\mathbf{H}g_i - \mathbf{H}g_j||^2 + (1-\alpha)||l_i - l_j||^2)$$
$$= \min_{\mathbf{K},\mathbf{L},\mathbf{H}} ||\mathbf{H} \odot \Upsilon||_1 = \min_{\mathbf{K},\mathbf{L},\mathbf{T}} ||\mathbf{T} \odot \Upsilon||_1, \tag{7}$$

where $\odot$ is the Hadamard product, l_i denotes the i-th category labels to the i-th sample, $\mathbf{L}$ is a matrix comprising of category labels of the entire training samples. $\Upsilon_{ij} = \alpha||\mathbf{H}g_i - \mathbf{H}g_j||^2 + (1-\alpha)||l_i - l_j||^2$, and α is a trade-off parameter between zero and one.

By combining (6) and (7), the objective function of our FS and be reorganized into:

$$\min_{\mathbf{K},\mathbf{L},\mathbf{T},\mathbf{J}} \{\tau_1||\mathbf{T}||_u + \tau_2||\mathbf{J}||_v + ||\mathbf{T} \odot \Upsilon||_1 + \tau_3||\mathbf{K}||_{12}\},$$
$$s,t. \ \text{constraints}(\mathbf{T},\mathbf{J}), \tag{8}$$

In our method, the l_1-norm is employed for both $||\mathbf{T}||_u$ and $||\mathbf{J}||_v$. The l_{12}-norm ensures the column-wise sparsity of $\mathbf{K}$. Based on the constrains detailed above (6), objective function (8) can be updated into:

$$\min_{\mathbf{K},\mathbf{L},\mathbf{T},\mathbf{J}} \{\tau_1||\mathbf{T}||_1 + \tau_2||\mathbf{J}||_2 + ||\mathbf{T} \odot \Upsilon||_1 + \tau_3||\mathbf{K}||_{12}\},$$
$$s,t. \ \mathbf{G} = \mathbf{GT}, \mathrm{diag}(\mathbf{T}) = 0. \tag{9}$$

where τ_3 denotes another nonnegative weight. In our implementation, the solution is based on [43].

Based on the $\mathbf{H}$ calculated from $\mathbf{T}$ in (9), the category labels l^* of a new sample is derived by:

$$\arg\min_{l^*} \sum_{i=1}^{N+1} \sum_{i=1}^{N+1} \mathbf{H}_{ij}||l_i - l_j||_2^2 = \arg\min_{l^*} \mathrm{tr}(\mathbf{L}^*\mathbf{R}\mathbf{L}^{*T}), \tag{10}$$

where $\mathbf{R} = \mathrm{diag}(\mathbf{H}\mathbf{I}_{N+1}) - \mathbf{H}$ is the graph Laplacian matrix, $\mathbf{I}_{N+1}$ is an $(N+1) \times (N+1)$-sized identity matrix, and $\mathbf{L}^* = [\mathbf{L}, l^*]$.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

We validate the effectiveness and efficiency of our HR aerial photo categorization using four experiments. We first introduce our self-compiled data set, which includes million-scale LR(low resolution)&HR aerial photos collected from the top 100 metropolises from different continents. Subsequently, we compare our approach with 17 state-of-the-art deep categorization models from three perspectives: accuracy, stability, and time consumption. Then, we evaluate our categorization accuracy by adjusting the multiple inherent parameters, based on which the optimal parameters are suggested. Lastly, we design an ablation study to evaluate each key module in our HR aerial photo categorization pipeline.

A. DATA SET DESCRIPTION

To comprehensively evaluate our categorization model, we have to experiment on a massive-scale LR&HR aerial photo set from many categories. To our best knowledge, however, there is no such data set in the literature. In this work, we spent enormous efforts to compile a huge data set containing over 3.6 million LR&HR aerial photos. The sources of these LR&HR aerial photos are Google/Apple/Bing Maps, based on which we designed a crawler software that spent 4310 hours to search and download LR&HR aerial photo. Specifically, we use the name of 100 most popular metropolitan cities (as detailed

[2] For ease of expression, the samples denote the training/testing HR aerial photos in this article.

in Fig. 3) throughout the world as the keywords to search Google/Apple/Bing Maps. In total, there are 46 cities from North America, 38 from Europe, ten from Asia, four from Oceania, and two from South America. Subsequently, we crop LR&HR aerial photos from the cached maps, wherein the typical resolutions of HR aerial photos are between $5K \times 5K$ and $22K \times 22K$. In our implementation, we restrict the HR aerial photos' resolution upper bound to $22K \times 22K$. Meanwhile, the resolutions of LR aerial photos are between $0.35K \times 0.35K$ and $2K \times 2K$. We adopt these settings because: 1) we want to make each HR aerial photo associated with four categories mostly, 2) we enforce that there are maximally 5% overlapping areas between any pairwise LR&HR aerial photos, and 3) too few pixels inside an LR aerial photo will make it technically infeasible to perceive its semantics.

During our data set compilation, we notice that a few LR&HR aerial photos are blurred due to bad weathers or sensitive military regions, as exemplified in Fig. 4. Actually, our method focuses on discovering object patches with different scales and subsequently learn visual perceptual features for visual categorization. Practically, bad weathers will inevitably decrease the visibility of LR&HR aerial photos and in turn hurt the fairness of accuracy comparison. Therefore we abandon LR&HR aerial photos whose 20% pixels are unclear, wherein the clearness is measured by the blur estimation algorithm proposed by Tong et al. [22]. To quantitatively show the effectiveness of the above refining process, we use the IQA (image quality assessment) algorithm [23] to calculate the quality score of each LR&HR aerial photo in our data set. More specifically, the single image quality assessment module in [23] is adopted here to calculate the quality score (with normalization). Herein, the quality score used in our implementation is a normalized image quality score. We manually select K best quality HR aerial image with sharpness scores $\{Q_b^1, \cdots, Q_b^K\}$. Then, for a new HR aerial image with sharpness score Q_{new}, its quality score is calculated as:

$$Q = \frac{Q_{new}}{ave(Q_b^k)}, \tag{11}$$

As reported in Fig. 5, over 74% of our refined LR&HR aerial photos are scored over 0.7.

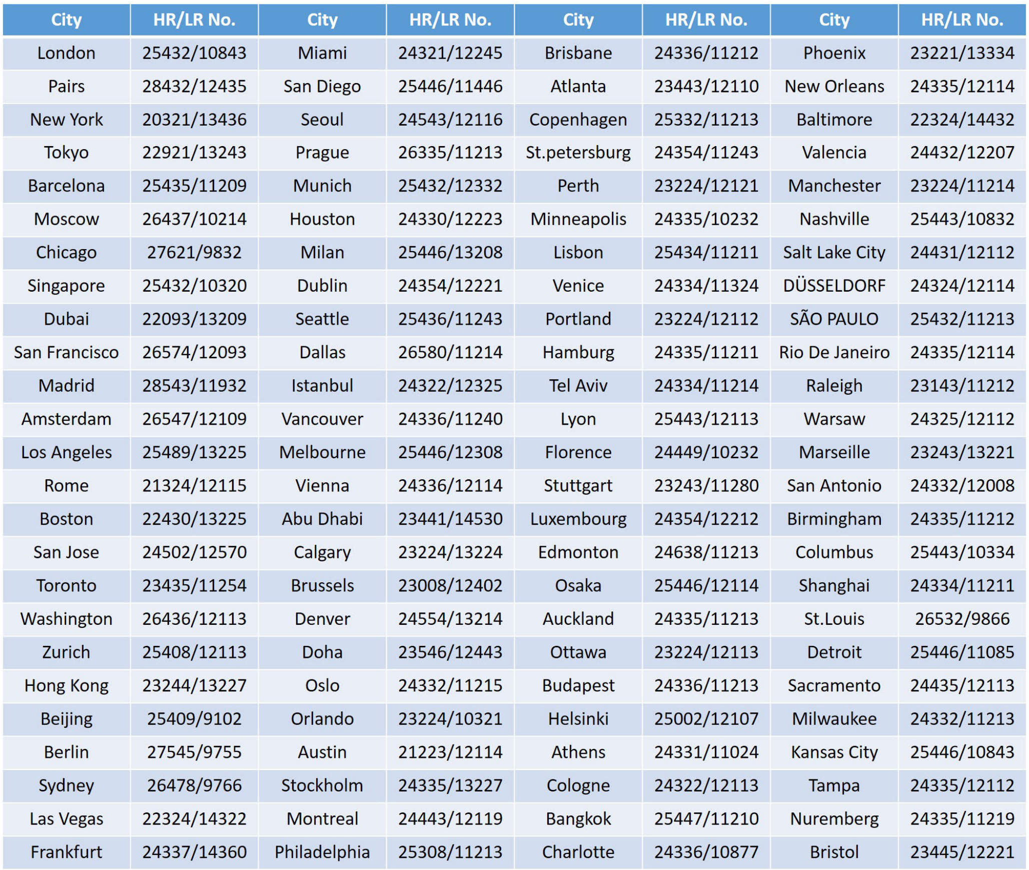
After collecting the million-scale LR&HR aerial photos, we have to annotate them to obtain the corresponding category labels. Herein, 106 volunteers[3] first manually annotate 23.8% HR aerial photos in each metropolitan city, wherein a total of 47 different category labels were utilized. Afterward, we train a multi-label SVM and employ it to annotate the category labels of the rest LR&HR aerial images. Then, the same 106 volunteers manually correct the labels calculated by SVM. It is noticeable that multiple category labels are associated with an intolerably small number of LR&HR aerial photos. This makes it infeasible to train a generalizable categorization model corresponding to these category labels. In our implementation, if the number of LR&HR aerial photos corresponding to a category label is smaller than 200,000, Then we abandon this label. In this way, we finally obtain 18 different category labels as detailed in Table 1. Thereafter, we notice that 99.983% LR&HR aerial photos have fewer than four category labels, while the rest very few LR&HR aerial photos have larger numbers of category labels (from five to 15). These LR&HR aerial photos usually contain a rich set of small regions ($< 200 \times 200$) that are possibly contaminated. Thus we simply abandon them. Lastly, we order the entire LR&HR aerial photos by their file names. The entire HR aerial photos are employed for training. For each category, the first half HR aerial photos constitute the training set while the rest are employed for testing. The entire LR aerial photos are employed for model validation, since manually detecting objects on LR aerial photos is much more convenient than the HR ones.

B. PERFORMANCE COMPARISON

Herein, our method is compared with seven deep categorization models [24], [25], [26], [27], [28], [29], [30] that intrinsically encode some prior knowledge of different aerial photo categories. We notice that the source codes of [24], [25], [28], and [29] are publicly available. Thereby, we conduct a comparative study wherein the parameter settings are set as default. For [26], [27], and [30], the source codes are unavailable to our knowledge. Due to this reason, these baseline methods are re-implemented by software programmer. We tried our best so as to make the re-implemented models perform similarly to the results reported in their publications. Nowadays, many deep generic recognition models perform impressively on categorizing aerial photos. Herein, our method is first made a comparison with multiple deep generic object recognition models: the pyramid pool-CNN (S-CNN) [31], CNet [32], discrimination filtering bank algorithm (DFBA) [33], C-RNN [34], multi-label graph convolutional network (MLT) [35], semantic-specific graph model (SGM) [36] and multi-label transformer model (MTM) [37]. Furthermore, since HR aerial photo categorization can be deemed is a sub-topic of scenery classification, we additionally compare with three well-known scenery classification models [38], [39], [40].

For the above baseline object/scene recognition algorithms, each model is repeatedly tested multiple times and the results are displayed in Table 2. As shown, our method achieve the best per-category accuracies on the entire 18 categories. To quantify the stability of these categorization models, we report their standard derivations simultaneously. We observe that the per-category standard derivations produced by our method are significantly and consistently lower than its competitors. This demonstrated that our method is the most stable.

[3] They are graduate students from our computer science department. They are aged between 24 and 31 and experienced in image processing and pattern recognition. Totally, there are 57 males and 49 females.



City	HR/LR No.	City	HR/LR No.	City	HR/LR No.	City	HR/LR No.
London	25432/10843	Miami	24321/12245	Brisbane	24336/11212	Phoenix	23221/13334
Pairs	28432/12435	San Diego	25446/11446	Atlanta	23443/12110	New Orleans	24335/12114
New York	20321/13436	Seoul	24543/12116	Copenhagen	25332/11213	Baltimore	22324/14432
Tokyo	22921/13243	Prague	26335/11213	St.petersburg	24354/11243	Valencia	24432/12207
Barcelona	25435/11209	Munich	25432/12332	Perth	23224/12121	Manchester	23224/11214
Moscow	26437/10214	Houston	24330/12223	Minneapolis	24335/10232	Nashville	25443/10832
Chicago	27621/9832	Milan	25446/13208	Lisbon	25434/11211	Salt Lake City	24431/12112
Singapore	25432/10320	Dublin	24354/12221	Venice	24334/11324	DÜSSELDORF	24324/12114
Dubai	22093/13209	Seattle	25436/11243	Portland	23224/12112	SÃO PAULO	25432/11213
San Francisco	26574/12093	Dallas	26580/11214	Hamburg	24335/11211	Rio De Janeiro	24335/12114
Madrid	28543/11932	Istanbul	24322/12325	Tel Aviv	24334/11214	Raleigh	23143/11212
Amsterdam	26547/12109	Vancouver	24336/11240	Lyon	25443/12113	Warsaw	24325/12112
Los Angeles	25489/13225	Melbourne	25446/12308	Florence	24449/10232	Marseille	23243/13221
Rome	21324/12115	Vienna	24336/12114	Stuttgart	23243/11280	San Antonio	24332/12008
Boston	22430/13225	Abu Dhabi	23441/14530	Luxembourg	24354/12212	Birmingham	24335/11212
San Jose	24502/12570	Calgary	23224/13224	Edmonton	24638/11213	Columbus	25443/10334
Toronto	23435/11254	Brussels	23008/12402	Osaka	25446/12114	Shanghai	24334/11211
Washington	26436/12113	Denver	24554/13214	Auckland	24335/11213	St.Louis	26532/9866
Zurich	25408/12113	Doha	23546/12443	Ottawa	23224/12113	Detroit	25446/11085
Hong Kong	23244/13227	Oslo	24332/11215	Budapest	24336/11213	Sacramento	24435/12113
Beijing	25409/9102	Orlando	23224/10321	Helsinki	25002/12107	Milwaukee	24332/11213
Berlin	27545/9755	Austin	21223/12114	Athens	24331/11024	Kansas City	25446/10843
Sydney	26478/9766	Stockholm	24335/13227	Cologne	24322/12113	Tampa	24335/12112
Las Vegas	22324/14322	Montreal	24443/12119	Bangkok	25447/11210	Nuremberg	24335/11219
Frankfurt	24337/14360	Philadelphia	25308/11213	Charlotte	24336/10877	Bristol	23445/12221

FIGURE 3. The statistics of LR&HR aerial images collected from the 100 metropolitan cities.

TABLE 1. The selected 18 categories and the corresponding LR&HR aerial photo numbers.

Category	HR No.	LR No.	Category	HR No.	LR No.
Tall building	1,121,110	454,130	Residential	1,232,108	544,114
Forest	1,221,132	654,118	Sea	1,324,337	434,142
Aircraft	1,367,215	355,619	Railway	1,254,005	476,094
Road	1,556,540	453,884	River	1,324,337	435,093
Palace	1,375,547	546,881	Factory	1,443,672	509,448
Vehicle	1,325,443	621,214	Yacht	1,324,216	432,116
Intersection	1,414,214	315,446	Soccer field	1,116,436	454,338
Bridge	1,211,548	324,801	Park	1,325,658	342,556
Farmland	1,436,658	543,447	Swim. pool	1,213,008	376,643

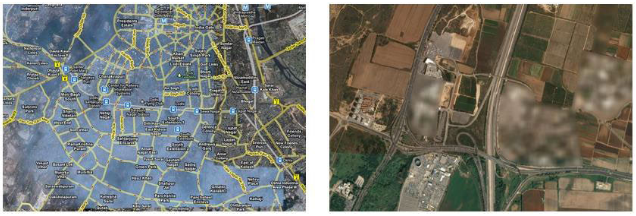


FIGURE 4. Examples of foggy (left) and blurred sensitive military (right) regions.

1) TRAINING/TESTING TIME COMPARISON

It is generally acknowledged that time consumption is a key criterion reflecting the performance of a categorization model. Herein, we report the training and testing time of the aforementioned 18 aerial photo categorization models. As shown in Table 3, during training, only two baseline models are faster than our pipeline. This is because the architectures of [31] and [35] are much simpler than ours. Meanwhile, we observe that the per-category accuracies of [31] and [35] are noticeably lower than ours. For the testing time comparison, our method can be conducted at a significantly faster speed than the baseline methods. Notably, distinguished from model training that can be conducted offline, outstanding testing time is comparably more valuable to many time-sensitive AI systems, such as weather forecasting and automatic navigation.

Our HR aerial photo categorization pipeline involves three key modules: 1) GSP learning using the low-rank algorithm, 2) geometry-preserving FS. During training, the time consumed for each module is: 9h12m (module 1) and 1h51m (module 2). During testing, the time cost of each module is: 71ms (module 1) and 11ms (module 2). We

TABLE 2. Performances with deviations of the aforementioned image recognizers (The highest accuracies are shown in bold numbers).

	[24]	[25]	[26]	[27]	[28]	[29]	[30]	S-CNN [31]	CNet [32]
Class1	0.633± 0.011	0.576±0.012	0.643±0.009	0.593±0.013	0.625±0.011	0.588 ±0.007	0.622±0.009	0.657±0.008	0.661±0.008
Class2	0.591±0.009	0.572±0.011	0.605±0.011	0.567±0.009	0.618±0.014	0.612±0.011	0.545±0.011	0.609±0.013	0.588±0.009
Class3	0.711±0.005	0.707±0.009	0.679±0.011	0.663±0.009	0.711±0.008	0.652±0.011	0.705±0.007	0.661±0.007	0.672±0.010
Class4	0.666±0.011	0.663±0.009	0.661±0.008	0.643±0.011	0.676±0.011	0.631±0.009	0.681±0.014	0.695±0.007	0.683±0.011
Class5	0.663±0.011	0.643±0.009	0.651±0.009	0.630±0.007	0.636±0.011	0.615±0.012	0.658±0.009	0.631±0.013	0.673±0.012
Class6	0.548±0.009	0.548±0.008	0.561±0.010	0.551±0.008	0.579±0.007	0.546±0.011	0.575±0.013	0.536±0.009	0.561±0.011
Class7	0.731±0.014	0.682±0.011	0.718±0.011	0.671±0.009	0.708±0.009	0.707±0.011	0.671±0.008	0.707±0.009	0.678±0.011
Class8	0.632±0.009	0.605±0.014	0.615±0.009	0.631±0.011	0.608±0.011	0.575±0.011	0.572±0.011	0.591±0.008	0.581±0.007
Class9	0.555±0.011	0.547±0.012	0.567±0.007	0.553±0.012	0.544±0.013	0.572±0.013	0.550±0.009	0.543±0.010	0.575±0.009
Class10	0.617±0.009	0.618±0.009	0.618±0.011	0.602±0.009	0.628±0.011	0.610±0.011	0.584±0.011	0.611±0.009	0.611±0.009
Class11	0.712±0.012	0.681±0.011	0.706±0.009	0.693±0.009	0.722±0.011	0.695±0.008	0.670±0.011	0.682±0.008	0.706±0.010
Class12	0.658±0.014	0.641±0.012	0.651±0.011	0.657±0.011	0.670±0.011	0.682±0.010	0.652±0.008	0.671±0.008	0.662±0.007
Class13	0.668±0.011	0.623±0.012	0.645±0.009	0.611±0.008	0.623±0.012	0.643±0.012	0.632±0.008	0.621±0.008	0.603±0.009
Class14	0.631±0.011	0.616±0.009	0.581±0.009	0.607±0.009	0.623±0.011	0.615±0.010	0.582±0.009	0.581±0.009	0.611±0.009
Class15	0.611±0.009	0.575±0.011	0.591±0.008	0.582±0.012	0.581±0.012	0.612±0.011	0.578±0.010	0.586±0.011	0.607±0.014
Class16	0.668±0.008	0.641±0.009	0.646±0.011	0.682±0.010	0.641±0.010	0.663±0.011	0.652±0.011	0.652±0.013	0.651±0.011
Class17	0.686±0.011	0.702±0.010	0.692±0.008	0.712±0.008	0.707±0.008	0.702±0.011	0.708±0.011	0.682±0.008	0.688±0.009
Class18	0.654±0.009	0.615±0.010	0.631±0.013	0.654±0.011	0.619±0.015	0.658±0.009	0.652±0.010	0.609±0.011	0.618±0.009

	DFBA [33]	C-RNN [34]	MTM [35]	SGM [36]	MLT [35]	[38]	[39]	[40]	Proposed
Class1	0.612±0.012	0.645±0.015	0.643±0.009	0.681±0.012	0.671±0.012	0.614±0.009	0.626±0.014	0.643±0.017	**0.721±0.011**
Class2	0.584±0.011	0.614±0.014	0.617±0.014	0.641±0.013	0.616±0.012	0.565±0.009	0.597±0.015	0.588±0.016	**0.676±0.009**
Class3	0.701±0.013	0.673±0.011	0.708±0.011	0.732±0.009	0.738±0.012	0.679±0.012	0.658±0.011	0.671±0.009	**0.778±0.004**
Class4	0.645±0.011	0.712±0.014	0.721±0.010	0.725±0.009	0.708±0.012	0.656±0.015	0.664±0.013	0.657±0.014	**0.752±0.012**
Class5	0.657±0.010	0.632±0.011	0.639±0.011	0.672±0.009	0.654±0.011	0.661±0.011	0.653±0.009	0.665±0.009	**0.698±0.009**
Class6	0.578±0.009	0.548±0.014	0.578±0.016	0.576±0.013	0.587±0.018	0.566±0.015	0.547±0.013	0.533±0.012	**0.626±0.013**
Class7	0.663±0.013	0.674±0.012	0.671±0.012	0.723±0.009	0.716±0.012	0.621±0.009	0.665±0.014	0.681±0.015	**0.776±0.009**
Class8	0.622±0.015	0.621±0.015	0.631±0.013	0.621±0.014	0.617±0.015	0.616±0.015	0.615±0.013	0.608±0.013	**0.689±0.011**
Class9	0.558±0.014	0.544±0.012	0.578±0.013	0.585±0.013	0.531±0.014	0.532±0.011	0.544±0.012	0.513±0.013	**0.589±0.003**
Class10	0.613±0.011	0.621±0.009	0.596±0.015	0.654±0.014	0.631±0.010	0.612±0.011	0.612±0.013	0.614±0.010	**0.695±0.012**
Class11	0.722±0.009	0.709±0.009	0.709±0.009	0.717±0.012	0.703±0.011	0.664±0.009	0.658±0.016	0.660±0.017	**0.749±0.005**
Class12	0.649±0.011	0.644±0.019	0.665±0.013	0.693±0.012	0.684±0.012	0.683±0.014	0.658±0.013	0.672±0.011	**0.718±0.010**
Class13	0.611±0.012	0.629±0.013	0.617±0.016	0.623±0.014	0.627±0.013	0.587±0.014	0.594±0.013	0.572±0.015	**0.683±0.004**
Class14	0.588±0.016	0.608±0.015	0.613±0.014	0.616±0.013	0.615±0.015	0.616±0.018	0.612±0.019	0.622±0.015	**0.674±0.006**
Class15	0.593±0.013	0.592±0.014	0.571±0.009	0.609±0.015	0.611±0.013	0.594±0.015	0.571±0.016	0.614±0.017	**0.634±0.004**
Class16	0.648±0.015	0.681±0.015	0.669±0.015	0.647±0.013	0.677±0.015	0.641±0.015	0.642±0.014	0.646±0.014	**0.687±0.006**
Class17	0.715±0.014	0.713±0.013	0.712±0.016	0.714±0.016	0.722±0.011	0.682±0.011	0.627±0.012	0.717±0.015	**0.791±0.005**
Class18	0.611±0.014	0.638±0.014	0.636±0.017	0.658±0.014	0.626±0.013	0.609±0.014	0.619±0.014	0.622±0.011	**0.702±0.003**

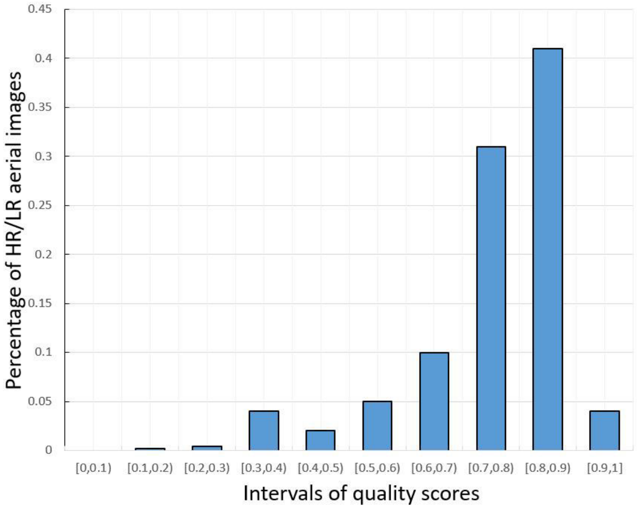


FIGURE 5. Statistics of LR&HR aerial photos with different quality scores in our complied LR&HR aerial photo set.

observe that most of the training time is spent for module 1 and practically this can be accelerated by Nvidia GPUs.

C. PARAMETER ANALYSIS

We first evaluate three weights in our low-rank algorithm. Parameter α, β, γ and L's default values are fixed to 0.3, 0.1, and 0.15 respectively. In our implementation, the default values are determined by 10-fold cross validation. The validation set contains 54000 samples, which is constituted by selecting 3000 HR aerial photos from each of the 18 categories. More concretely, we tune each of α, β, and γ from zero to one. And all the possible parameter combinations are enumeratively employed to test the HR aerial photo categorization. The parameter combination receiving the highest categorization accuracy is reported as the default values. Based on this, we adjust one of the three parameters while keep the others unchanged. Each parameter is increased from zero to one. We then report the accuracy accordingly. As the three curves displayed on the left of Fig. 6, the best performances are achieved when $\alpha = 0.1$, $\beta = 0.15$, and $\gamma = 0.3$.

To evaluate the influences of τ_1, τ_2, and τ_3 in our geometry-preserving FS, we set $\tau_1 = \tau_2$ and then tune τ_1 and τ_3. Then we follow the experimental settings described above. τ_1 and τ_3's initial values are both set to 0.45. As the two curves shown in Fig. 6, the highest performance is observed if $\tau_1 = 0.4$ and $\tau_3 = 0.6$.

D. ABLATION STUDY

As aforementioned, our method is comprised of two key modules: 1) GSP learning using the low-rank algorithm, 2) geometry-preserving FS. Herein, we test the importances of these modules in our HR aerial photo categorization pipeline. Specifically, each module is replaced by a different one. Then the performance decrement/increment is presented. Also, insights are provided

TABLE 3. Training/testing time of the 18 categorization models.

	[24]	[25]	[26]	[27]	[28]	[29]	[30]	S-CNN [31]	CNet [32]
Train	31h7m	43h14m	52h21m	39h23m	36h43m	46h13m	41h32m	6h33m	38h22m
Test	1.143s	1.774s	1.846s	1.564s	2.437s	1.463s	1.675s	0.893s	1.660s
	DFBA [33]	C-RNN [34]	MTM [35]	SGM [36]	MLT [35]	[38]	[39]	[40]	Proposed
Train	40h23m	25h25m	32h15m	44h16m	10h6m	32h14m	35h44m	32h12m	**11h3m**
Test	1.213s	1.002s	1.875s	0.983s	1.436s	1.774s	1.983s	1.546s	**0.782s**

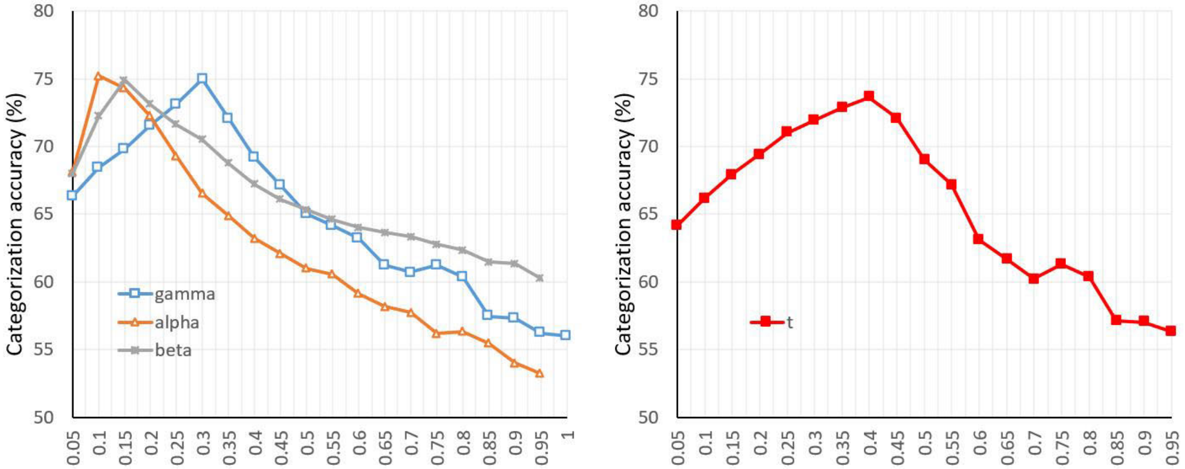


FIGURE 6. Categorization performance variation when tuning α, β, γ and τ_1, τ_2, τ_3 respectively.

to elaborate the underlying reasons for the observed results.

In the first place, to evaluate the effectiveness of the low-rank algorithm, two experimental settings are deployed. We first abandon the sparse constraint term $||\mathbf{E}||_1$ in (4) (marked by "S11"). Afterward, we abandon the regularizer $||\mathbf{Z}||_F^2 + \sum_{i=1}^{L}(||\mathbf{Z}_i||_F^2 + ||\xi||_2^2)$ in (4) (marked by "S12"). We report the variation of categorization accuracy in Table 4. Herein, the intersection of column "Si" and row "Oj" denotes the setup "Sij". Noticeably, a shallow feature engineering module will cause a performance decrement. Also, removing the regularizer will greatly decrease the accuracy. This observation shows the necessity to mitigate the overfitting of our designed low-rank algorithm. Next, to evaluate the performance of the geometry-preserving FS, we remove $\tau_1||\mathbf{T}||_1$, $\tau_1||\mathbf{J}||_2$, and $\tau_1||\mathbf{K}||_{12}$ respectively. As shown in 4, abandoning the geometry-preserving term causes the largest categorization accuracy drop. This demonstrates the importance of maintain sample distribution in FS.

TABLE 4. HR aerial photo categorization accuracy variation.

	S1	S2
O1	-2.115%	-2.325%
O2	-4.554%	-2.032%
O3	N/A	-6.665%

V. CONCLUSION

Recognizing aerial images is an indispensable task in geoscience and remote sensing [45], [46], [47], [48], [49]. We proposed a novel HR aerial photo categorization model, wherein the key is a low-rank algorithm as well as a geometry-preserving FS. The comparative study on our complied million-level HR aerial photo set has shown the competitiveness of our method.

One limitation of our work is the low-rank algorithm has a shallow architecture. Currently deep models have been pervasively used in visual categorization since they can produce more descriptive features. In the future, we plan to upgrade our low-rank algorithm into a deep architecture toward a more descriptor feature extractor.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] O. Akar, "The rotation forest algorithm and object-based classification method for land use mapping through UAV images," *Geocarto Int.*, vol. 33, no. 5, pp. 538–553, 2017.

[5] 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, Jun. 2018, Art. no. 7195432.

[6] 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.

[7] 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, May 2021.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] D. Costea and M. Leordeanu, ''Aerial image geolocalization from recognition and matching of roads and intersections,'' 2016, *arXiv:1605.08323*.

[13] 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. 5624216.

[14] 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. 4707013.

[15] V. Chalavadi, P. Jeripothula, R. Datla, S. B. Ch, and M. C. Krishna, ''mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions,'' *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108548.

[16] P. Wang, Z. He, K. Xie, J. Gao, M. Antolovich, and B. Tan, ''A hybrid algorithm for low-rank approximation of nonnegative matrix factorization,'' *Neurocomputing*, vol. 364, pp. 129–137, Oct. 2019.

[17] Z. Kang, C. Peng, and Q. Cheng, ''Top-N recommender system via matrix completion,'' in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 179–185.

[18] X. He, ''Laplacianfaces,'' *Scholarpedia*, vol. 5, no. 8, p. 9324, 2010.

[19] Z. Li, J. Liu, Y. Jiang, J. Tang, and H. Lu, ''Low rank metric learning for social image retrieval,'' in *Proc. 20th ACM Int. Conf. Multimedia*, Oct. 2012, pp. 853–856.

[20] Y. Luo, Y. Wong, and Q. Zhao, ''Label Consistent Quadratic Surrogate model for visual saliency prediction,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5060–5069.

[21] L. Xiao, Z. Zhu, H. Liu, C. Li, and W. Fu, ''Gaze prediction based on long short-term memory convolution with associated features of video frames,'' *Comput. Electr. Eng.*, vol. 107, Apr. 2023, Art. no. 108625.

[22] H. Tong, M. Li, H. Zhang, and C. Zhang, ''Blur detection for digital images using wavelet transform,'' in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, vol. 1, Jun. 2004, pp. 17–20.

[23] H. Zhang, B. Li, J. Zhang, and F. Xu, ''Aerial image series quality assessment,'' *IOP Conf. Ser., Earth Environ. Sci.*, vol. 17, no. 1, Mar. 2014, Art. no. 012183.

[24] C. Kyrkou and T. Theocharides, ''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, pp. 1687–1699, 2020.

[25] C. Kyrkou and T. Theocharides, ''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.

[26] 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.

[27] 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.

[28] M. D. Pritt and G. Chern, ''Satellite image classification with deep learning,'' 2020, *arXiv:2010.06497*.

[29] 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.

[30] 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.

[31] K. He, X. Zhang, S. Ren, and J. Sun, ''Spatial pyramid pooling in deep convolutional networks for visual recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[32] K.-H. Lee, X. He, L. Zhang, and L. Yang, ''CleanNet: Transfer learning for scalable image classifier training with label noise,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5447–5456.

[33] Y. Wang, V. I. Morariu, and L. S. Davis, ''Learning a discriminative filter bank within a CNN for fine-grained recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4148–4157.

[34] A. Caglayan and A. B. Can, ''Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition,'' in *Proc. ECCV Workshops*, 2018, pp. 675–688.

[35] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, ''Multi-label image recognition with graph convolutional networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5172–5181.

[36] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, ''Learning semantic-specific graph representation for multi-label image recognition,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 522–531.

[37] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, ''General multi-label image classification with transformers,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16473–16483.

[38] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, ''Unsupervised learning of semantics of object detections for scene categorization,'' in *Pattern Recognition Applications and Methods*. IEEE, 2015.

[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. ICCV*, 2017, pp. 5746–5754.

[41] 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. 133b, Sep. 2019.

[42] Z. Li, J. Tang, L. Zhang, and J. Yang, ''Weakly-supervised semantic guided hashing for social image retrieval,'' *Int. J. Comput. Vis.*, vol. 128, nos. 8–9, pp. 2265–2278, Sep. 2020.

[43] M. Fan, X. Zhang, J. Hu, N. Gu, and D. Tao, ''Adaptive data structure regularized multiclass discriminative feature selection,'' *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5859–5872, Oct. 2022.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] V. Damminsed, 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.

[49] 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 the Shanghai Danji University. His research interests include image processing, computer vision, and pattern recognition.

FUJI REN (Senior Member, IEEE) is currently a Researcher with Anhui University.

• • •