IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 September 2023, accepted 26 September 2023, date of publication 6 November 2023,
date of current version 10 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3330241

RESEARCH ARTICLE

LR Aerial Photo Categorization by Semi-Supervised Perceptual Feature Selection

YINHAI LI AND YICHUAN SHENG
Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321007, China
Corresponding author: Yinhai Li (yinghaili@jhc.edu.cn)

ABSTRACT Recognizing the semantic categories of low-resolution (LR) aerial photos is an indispensable technique in geoscience and remote sensing. However, it is also a challenging task in practice. In this work, a semi-supervised perceptual feature selection (SPFS) pipeline is proposed for LR aerial photo categorization, focusing on selecting high quality perception-guided visual features. Specifically, by mimicking human vision system, a novel low-rank model is designed to decompose each LR aerial photo into multiple visually or semantically salient foreground regions coupled with the background non-salient regions. This model can: 1) produce the a gaze shifting path (GSP) simulating human gaze behavior; and 2) generate hierarchical deep representation for a GSP. Afterward, a semi-supervised feature selection (FS) is leveraged toward a succinct set of discriminative deep GSP features, wherein only labels of LR aerial photos are required. Based on the selected features, a classifier is trained for visual categorization. Comprehensive experimental results have validated our method's advantage.

INDEX TERMS Aerial imagery, semi-supervised, cross-resolution, human perception.

I. INTRODUCTION

Owing to the remarkable progress in carrier rocket, remote sensing, and satellite communication, hundreds of earth observation satellites have been launched since October 1957. According to the orbital altitudes, these satellites can be categorized into the high- (>2000km) and low-altitude ones (200∼2000km). Distinguished from low-altitude satellites, high-altitude ones cover a comparatively larger area with a longer orbital period. Thus resolutions of aerial photos captured by these high-altitude satellites are typically lower than the low-altitude ones. In practice, effectively understanding the semantic categories of these LR aerial photos is a useful technique in many computer vision tasks. For example, by periodically monitoring the geographical distribution of animals, forests, and swamps from an LR aerial photo, the biodiversity and wildlife trends can be well tracked. It is significant for keeping habitats inside their sanctuaries, especially for the endangered animals like pandas. Moreover, to optimize the planned path for long-haul driverless trucks, we have to accurately recognize the semantic categories of a variety of regions inside each LR aerial photo, based on which the shortest path between locations can be rapidly and dynamically calculated.

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

In computer vision, multiple categorization/annotation models have were designed to characterize aerial photos with mid/high resolutions (spatial resolution $\leq$10m). Representative work includes: 1) shallow/deep-learning-based object localization using weak labels [55], [56]; 2) graph models to enhance semantic propagation for aerial photo labeling [5], [6], [7]; and 3) carefully-designed hierarchical architectures for visual segmentation toward aerial photos [8], [9], [10]. As far as we know, however, the existing approaches cannot effectively encode LR aerial photos due to two reasons:

- Typically, there are tens of foreground objects within each LR aerial photo, as shown on the top of Fig.1. To calculate the semantics of an LR aerial photo, we expect a bionic model that simulates the process of human perceiving the foreground salient regions. Actually, building a deep model that can simultaneously extract the visually/semantically salient regions and engineer the deep features for these extracted regions is non-trivial. Potential challenges include: i) determining the sequence of humans observing the extracted salient regions (e.g., the path displayed in Fig.1, 2) refining the contaminated labels of the training LR aerial photos, and 3) transferring image-level semantic labels into multiple regions inside an LR aerial photo;
- Compared to HR aerial photos, LR ones are usually with an inferior image quality, as they are more sensitive

© 2023 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

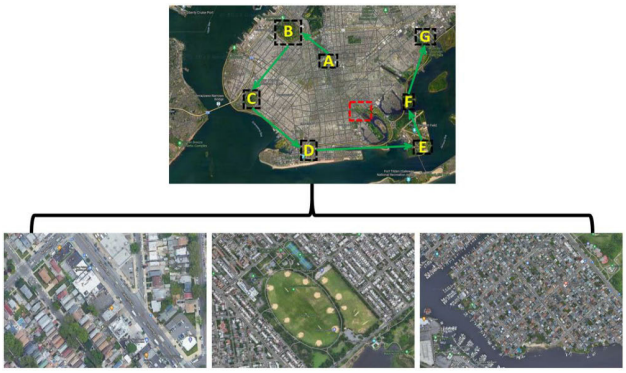

FIGURE 1. Top: salient aerial image patches sequentially observed by humans (marked by path $A \rightarrow \cdots \rightarrow G$) as well as the blurred playground (marked by red dashed box). Bottom: three HR aerial photos capture sub-regions of the LR aerial photo (the middle details the blurred playground inside the LR aerial photo).

to a variety of uncontrollable factors, e.g., the varying weather/lighting conditions and possibly communication interference. This brings a limited number of labeled LR aerial photos, coupled with a rich set of labeled HR ones. Thus, we expect a semi-supervised feature selector that is trained by partially-labeled LR aerial photos, which is an nontrivial task. Potential difficulties include how to uncover the underlying correlations among LR and HR aerial photos in the feature space.

In this work, a so-called SPFS framework is formulated that adopts the deeply-learned perceptual experiences from HR aerial photos to enhance LR one categorization. Given a considerable quantity of HR and LR aerial photos, part of which are unlabeled. We first project their internal regions onto the feature space constructed based on discovering the visual and semantic channels collaboratively. Afterward, to mimick human visual perception, a deep low-rank model is designed to decompose each LR aerial photo into a sequence of visually/semantically salient foreground regions, i.e., gaze shifting path (GSP) coupled with the non-salient backgrounds, wherein the deep representation for each GSP is calculated simultaneously. Aiming at a concise set of discriminative features shared between HR and LR aerial photos, a SPFS algorithm selects a concise set of high quality features shared between LR and HR aerial photos, wherein only a small faction of labeled samples are required. Besides, SPFS can optimally preserve the graph structure of LR/HR aerial photos during feature selection. Finally, the selected features are integrated into a kernel SVM for LR aerial photo categorization.

II. RELATED WORK

A. SEMANTICALLY MODELING AERIAL PHOTOS

Dozens of computational models were developed to analyze aerial photos. For visual modeling at image-level, Zhang et al. [57] 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 [59] formulated a weak model that semantically labels HR aerial photos at image-level. Akar et al. [60] proposed the so-called random forest and object-level feature extractor to classify each aerial image. The authors [62] developed a hierarchical CNN architecture for identifying the multiple labels of HR aerial photos describing many downtown areas. In [58], the authors utilized a deep model to classify remote sensing images. A domain-specific scenic picture set is leveraged to fine tune the deep architecture. In [43], 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. Researchers [11] designed a multi-resolution model to learn the weights of aerial image features both horizontally and vertically. In [64], 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. [4] proposed a deep learning model for discovering salient objects in each aerial image. In [1], a focal loss deep architecture is proposed that optimally discovers vehicles from aerial images. In [63], The authors developed a learning model toward aerial photos by intelligently extracting intersections and streets. In [18], Yu et al. integrated feature enhancement and soft label assignment into an anchor-independent object detector toward aerial images. In [19], Wang et al. proposed a deep rotation-invariant detector that effectively estimates the angles of multi-scale objects inside aerial images. In [54], 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.

B. SUPERVISED FEATURE SELECTION (FS)

In supervised FS, each feature's discrimination is quantized by its correlation with the labels. Nie et al. [15] formulated an effective FS algorithm by optimizing an objective function based on an $\boldsymbol{l_{12}}$-norm regularization. A fast and incremental FS framework particularly designed for high-dimensional features was formulated by [16]. Gui et al. proposed an attention-guided feature scoring algorithm in a supervised setting. Based on an elaborately-designed smooth hinge loss, a sparsity-regularized model was proposed to obtain a subset of discriminative features. In [17], an $\boldsymbol{l_{12}}$-norm coupled with an exclusive lasso was incorporated for FS, wherein the redundant and contaminated features can be optimally abandoned. An effective measure was proposed for identifying discriminative features. In [14], Ahadzadeh et al. proposed a double-stage FS based on particle swarm optimization toward high-dimensional features. Stage one globally removes low quality features while stage two locally searches the highly discriminative ones. Noticeably, the above FS handle features in the original space, whereas practically the samples may be distributed in the high-order

kernel space. Song et al. [30] designed a kernel-induced FS to maximize the correlation between the selected features and labels. In [31], Masaeli et al. proposed an HSIC-based implicit FS algorithm (a.k.a. feature transformation) using an l_1/l_∞-norm regularizer. Further, Yamada et al. [32] formulated the novel dual augmented Lagrangian in order to search for a global optimum. Researchers [33] proposed a kernel-induced feature selector that effectively acquires a subset of covariates that is most discriminative. In [34], Leng et al. extracted the features of both palmprints with 2D discrete cosine transform for constructing a dual-source space. And highly discriminative coefficients are optimally preserved for visual retrieval. Moreover, in [35], the standard cancelable palmprint coding is upgraded to 2D space. The so-called perpendicular orientation transposition and multi-orientation score level fusion collaboratively enhance the 2D cancelable palmprint codes.

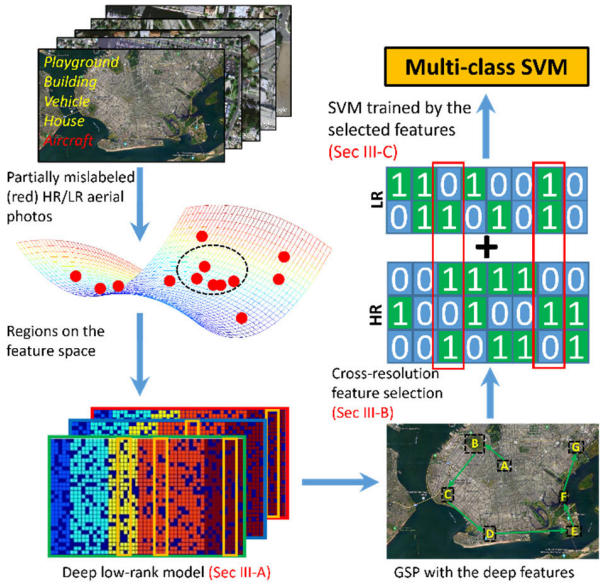

FIGURE 2. The pipeline of the LR aerial photo categorization by our designed SPFS framework. Our method first projects regions from HR/LR aerial photos into the feature space, based on which the deep low-rank algorithm is used to extract GSPs and generate the deep features accordingly. Then the SPFS is leveraged to select highly discriminative features, which are subsequently fed into the multi-class SVM for visual categorization.

III. OUR PROPOSED METHOD

An overview of our method is presented in Fig.2. Our method involves three key components: deep low-rank algorithm for GSP calculation, the semi-supervised perceptual feature selection, and the SVM training. The inter connection between these components are annotated by the blue arrows.

A. DEEP LOW-RANK ALGORITHM FOR GSP LEARNING

In practice, there are multiple fine-grained objects inside each LR aerial photo. Biological studies [2] have shown that humans usually attend to a few salient objects in the visual cognition process. In our scenario, to understand each LR aerial photo, we typcially first attend to the ground salient regions, wherein the background regions are kept almost unprocessed. Such human visual perceptual behavior is informative for categorizing LR aerial photos. Herein, we propose a deep low-rank algorithm that sequentially selects salient image patches to construct gaze shifting paths (GSPs). And the corresponding deep 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 LR 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 LR aerial photo and T its feature dimensionality. $\mathbf{Y} \in \mathbb{R}^{T\times N}$ preserves feature columns 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, multiple constrains 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 LR 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):

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

where $\|\cdot\|_*$ is the matrix nuclear norm representing a convex approximation to matrix rank function, $\boldsymbol{l_1}(\mathbf{E})$ quantizes the sparisty of $\mathbf{E}$, $\boldsymbol{f}(\boldsymbol{\Upsilon}, \mathbf{X})$ selects non-salient background image patches from each LR aerial photo, and $\boldsymbol{l_2}(\mathbf{Y}, \boldsymbol{f}(\boldsymbol{\Upsilon}, \mathbf{X}))$ penalizes the loss of non-salient background image patches selection. $\boldsymbol{\Omega}(\boldsymbol{\Upsilon})$ serves as a regularizer. $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ are positive coefficients balancing the trade-off among terms. More concretely, to ensure a highly sparse $\mathbf{E}$, $\boldsymbol{l_1}(\cdot)$ is defined as:

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

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

$$\min_{\mathbf{Y},\Omega} \|\mathbf{Y}\|_* + \alpha \|\mathbf{E}\|_1 + \beta \|\mathbf{Y} - \boldsymbol{f}(\Upsilon, \mathbf{X})\|_\mathbf{F}^2 + \gamma \boldsymbol{\Omega}(\boldsymbol{\Upsilon})), \tag{4}$$

To precisely select the non-salient background image patches inside each LR aerial photo, we formualte a deep semantic model $\boldsymbol{f}(\Upsilon, \mathbf{X})$. It includes $\boldsymbol{L}$ layers of linear/nonlinear transformations. The deep representation

from the top layer is denoted by $\boldsymbol{h}(\boldsymbol{x})$ and $\mathbf{X}_i$ is the T-dimensional column feature vector from the i-th image patch. Meanwhile, the current layer's output is utilized as the input of the next layer. Mathematically, this can be represented as:

$$\boldsymbol{h}(\mathbf{X}_i) = \boldsymbol{g}_L(\mathbf{X}_i), \tag{5}$$

$$\boldsymbol{g}_l(\mathbf{X}_i) = \phi(\boldsymbol{Z}_l \boldsymbol{h}_{l-1}(\mathbf{X}+\boldsymbol{\xi}_1)), \boldsymbol{l} = \mathbf{1}, \cdots, \boldsymbol{L}, \tag{6}$$

where $\boldsymbol{\phi}(\cdot)$ denotes the activation function and $\boldsymbol{g}_l(\cdot)$ the l-th layer's output. $\boldsymbol{Z}_l$ and $\boldsymbol{\xi}_l$ represent the transformation matrix and the bias corresponding to the l-th layer respectively. The first layer's input is $\mathbf{X}_i$, based on which the first layer's output is calculated as:

$$\boldsymbol{g}_l(\mathbf{X}_i) = \phi(\mathbf{Z}_l \mathbf{X}_i + \boldsymbol{\xi}_l), \boldsymbol{l} = \mathbf{1}, \cdots, \boldsymbol{L}, \tag{7}$$

We want the deeply-learned feature $\boldsymbol{h}(\mathbf{X}_i)$ sufficiently discriminative for selecting the non-salient background image patches. Without loss of generality, we adopt a linear mapping function to such selection process:

$$\boldsymbol{f}(\boldsymbol{\Upsilon}, \mathbf{X}) = \boldsymbol{Z}\boldsymbol{h}(\mathbf{X}), \tag{8}$$

where parameter set $\boldsymbol{\Upsilon} = \{\mathbf{Z}_1, \cdots, \mathbf{Z}_L, \boldsymbol{\xi}_1, \cdots, \boldsymbol{\xi}_L\}$.

To mitigate overfitting, we design a regularizer to penalize model complexity. Herein, the regularization function $\boldsymbol{\Omega}(\boldsymbol{\Upsilon})$ is given as:

$$\boldsymbol{\Omega}(\boldsymbol{\Upsilon}) = \frac{\mathbf{1}}{\mathbf{2}}(\|\mathbf{Z}\|_F^2 + \sum_{i=1}^{L}(\|\mathbf{Z}_l\|_F^2 + \|\xi_l\|_2^2)), \tag{9}$$

By leveraging the definition in (3,8,9), the objective function (4) can be upgraded into:

$$\min_{\mathbf{Y}\geq\mathbf{0}, \mathbf{Z}_1, \mathbf{Z}, \xi_1, \Omega} \|\mathbf{Y}\|_* + \alpha \|\mathbf{E}\|_1 + \beta \|\mathbf{Y} - \boldsymbol{Z}\boldsymbol{h}(\mathbf{X})\|_F^2 + \frac{\boldsymbol{\gamma}}{\mathbf{2}}(\|\mathbf{Z}\|_F^2 + \sum_{i=1}^{L}(\|\mathbf{Z}_l\|_F^2 + \|\boldsymbol{\xi}\|_2^2)), \tag{10}$$

The above optimization is non-convex over all the variables. In our implementation, we follow the iterative algorithm in [3] to solve it. Thereafter, denoting $\mathbf{Y}^*$ as (10)'s solution, the saliency score of the i-th image patch in an LR aerial photo is calculated by:

$$\boldsymbol{s}(\boldsymbol{X}_i) = \left\|\mathbf{E}^*(:, \boldsymbol{i})\right\|_2, \tag{11}$$

where $\mathbf{E}^* = \mathbf{X} - \mathbf{Y}^*$, and $\mathbf{E}^*(\boldsymbol{\mathcal{V}}, \boldsymbol{i})$ denotes the i-th column of $\mathbf{E}^*$. A larger $\boldsymbol{s}(\mathbf{X}_i)$ means that the i-th image patch is more visually/semantically salient. Given an LR aerial photo, we sequentially link the top $\boldsymbol{K}$ salient image patches to constitute its gaze shifting path (GSP). Thereby, the deep GSP representation is obtained by sequentially concatenating the deep features of its constituent $\boldsymbol{K}$ image patches.

During the deep low rank model learning, the loss function is the objective function in (10). The number of epochs is 200 and the learning rate is set to 0.005 and the entire deep network is pre-trained using the ResNet-152 [36].

B. SEMI-SUPERVISED PERCEPTUAL FEATURE SELECTION (SPFS)

Practically, the above deep GSP features might be inadequately discriminative. Herein, we expect to further obtain a subset of deep GSP features to enhance the subsequent visual categorization. Semi-supervised FS obtains high quality features by uncovering the binary relationships among labeled and unlabeled LR/HR aerial photos, which is suitable for our objective.

Without loss of generality, we assume that all the LR aerial photos are unlabeled while the entire HR ones are labeled. We denote as the feature matrix of the D-dimensional deep GSP features from both LR and HR aerial photos during training. The first $\boldsymbol{M}$ rows correspond to the M labeled HR aerial photos while the succeeding rows correpond to the unlabeled LR ones. $\boldsymbol{N}$ is the total number of training samples. Similarly, we denote $\boldsymbol{L} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_M, \boldsymbol{y}_{M+1}, \cdots, \boldsymbol{y}_N] \in \{\mathbf{0}, \mathbf{1}\}^{N \times C}$ as the label matrix of the training LR/HR aerial photos, wherein $\boldsymbol{C}$ counts the semantic categories. Herein, $\boldsymbol{y}_{ij}$ represents the j-th category label of $\boldsymbol{y}_i (1 \leq i \leq C)$. We set $\boldsymbol{y}_{ij} = \mathbf{1}$ if the i-th sample belonging to the j-th category, and $\boldsymbol{y}_{ij} = \mathbf{0}$ otherwise. Meanwhile, if the i-th sample is unlabeled, we simply set $\boldsymbol{y}_i$ as a $\boldsymbol{C}$-dimension row vector with all zeros.

Denoting $\mathbf{Q} \in \mathbb{R}^D \times \boldsymbol{C}$ as the projection matrix for FS, a general FS can be formulated by minimizing the error:

$$\min_{\boldsymbol{Q}} \mathcal{L}(\boldsymbol{Q}) + \tau \mathcal{R}(\boldsymbol{Q}), \tag{12}$$

where the first term calculates the loss and the second one represents the regularizer.

We define the affinity graph E, wherein each entity $\mathbf{E}_{ij}$ indicates the similarity between h_i and h_j. Herein, we simply set $\mathbf{E}_{ij} = 1$ if h_i and h_j are the K nearest neighbors, and $\mathbf{E}_{ij} = 0$ otherwise. Herein, the K nearest neighbors is built upon the standard KNN algorithm [66]. KNN searches the K nearest samples to a reference one in the feature space, wherein K is determined by users. In our implementation, we set K=5. We set $\mathbf{F}$ as a diagonal matrix as $\mathbf{F}_{ii} = \sum_{j=1}^{N} \mathbf{E}_{ij}$. Afterward, we set $\mathbf{T} = \mathbf{F} - \mathbf{E}$ as the graph Laplacian.

To optimally exploit the entire samples, we define a predicted label matrix as $\mathbf{P} = [\boldsymbol{p}_1, \cdots \boldsymbol{p}_N] \in \mathbb{R}^N \times \boldsymbol{C}$ toward the entire training samples by leveraging the transductive classification [65], where $\boldsymbol{p}_i \in \mathbb{R}^C$ is the predicted category label of sample $\boldsymbol{x}_i$. In our solution, we enforce $\mathbf{P}$ maximally satisfy the smoothness of both ground-truth category label and the affinity graph. Mathematically, $\mathbf{P}$ is calculated using the following objective function:

$$\mathbf{argmin}_{\mathbf{P},\mathbf{Q}} \mathbf{tr}\left(\mathbf{P}^{\mathbf{T}}\mathbf{T}\mathbf{P}\right) + \mathbf{tr}\left((\mathbf{P} - \mathbf{Y})^{\mathbf{T}} \mathbf{V} (\mathbf{P} - \mathbf{Y})\right) + \sigma \left\|\mathbf{X}^{\mathbf{T}}\mathbf{Q} - \mathbf{P}\right\| + \tau \mathcal{R}(\mathbf{Q}), \tag{13}$$

where $\left\|\mathbf{X}^{\mathbf{T}}\mathbf{Q} - \mathbf{P}\right\|$ denotes the loss function and $\mathcal{R}(\boldsymbol{Q})$ functions as a regularizer penalizing the projection matrix $\mathbf{Q}$ for optimal FS; $\boldsymbol{\sigma} \in [\mathbf{0}, \mathbf{1}]$ and $\boldsymbol{\tau} \in [\mathbf{0}, \mathbf{1}]$ weight the loss function and regularizer respectively.

Due to the high sparsity and robustness of the non-convexity, the $l_{1,p}$-matrix norm is applied to define regularizer $\mathcal{R}(\mathcal{Q})$ in our SPFS framework ($p \in (0, 1]$). In this way, the regularizer can be formulated as:

$$\mathcal{R}(\mathcal{Q}) = \|\mathcal{Q}\|_{2,p} = \left(\sum_{i=1}^{D} \left\|\mathbf{Q}^i\right\|_2^p\right)^{1/p}, \tag{14}$$

In our implementation, we set p = 1/2. The solution of (14) is detailed in [21]. In practice, if we set p to a different value, then the optimization might be non-convex and we cannot obtain a global optimal solution.

C. KERNEL-INDUCED FEATURE VECTOR FOR CATEGORIZATION

It is noticeable that the selected deep GSP features may be distributed on the high-order kernel feature space. Herein, a kernel-induced quantization method is employed to calculate each LR aerial photo's representation. For an LR aerial image, the object patches are extracted to build its GSP, which are simultaneously converted into deep GSP features for SPFS. Then, the selected deep GSP feature from the i-th LR aerial photo is accumulated into a kernel-induced vector $\mathbf{v_i} = \{v_{i1}, v_{i2} \cdots, v_{1N},\}$ where N counts the training LR/HR aerial photos and N counts the testing LR aerial photos. The j-th element of $\mathbf{v_i}$ is calculated as:

$$\mathbf{v}_{ij} \propto \mathbf{exp}(-\mathbf{d_J}(\mathbf{b}_u, \mathbf{b}_v)), \tag{15}$$

where $-\mathbf{d_J}(\cdot, \cdot)$ computes the Euclidean distance between pairwise selected deep GSP features. Given N testing LR aerial photos, we can obtain an $N \times N$ kernel matrix at the training stage and an $N \times N^{'}$ kernel matrix at the testing stage. The first matrix is utilized to learn a classifier for LR aerial photo classification, while the second one is employed for testing.

IV. EXPERIMENTS

We validate our LR aerial photo categorization using four experiments. We first introduce our self-compiled image set, which includes >3.7 million LR/HR aerial photos collected from the top 100 metropolises from different continents. Based on this, we compare our approach with 17 state-of-the-art deep categorization models from three perspectives: accuracy, stability, and time consumption. Thereafter, 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 SPFS-based LR aerial photo categorization pipeline. Simultaneously, we visualize a set of attractive image patches selected by our SPFS-based FS.

A. KERNEL-INDUCED FEATURE VECTOR FOR CATEGORIZATION

To comprehensively evaluate the 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. We spend enormous efforts to compile a huge data set containing over 3.6 million

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 number of HR/LR aerial images selected by us.

LR/HR aerial photos. The sources of these LR/HR aerial photos are Google/Apple/Bing Maps, based on which we design a crawler software that spent 4310 hours to search and download LR/HR aerial photo. More specifically, we use the name of 100 most popular metropolitan cities (as detailed 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 impossible to perceive its semantics.

FIGURE 4. Foggy (left) and blurred sensitive military (right) regions.

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 deep perceptual features for visual categorization. Practically, bad weathers will decrease the visibility of LR/HR aerial photos and in turn hurt the fairness of categorization 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. [47]. To quantitatively show the effectiveness of the above refining process, we use the IQA (image quality assessment) algorithm [48] to calculate the quality scores of LR/HR aerial photos in our data set. As reported in Fig. 5, over 74% of our refined LR/HR aerial photos are scored over 0.7.

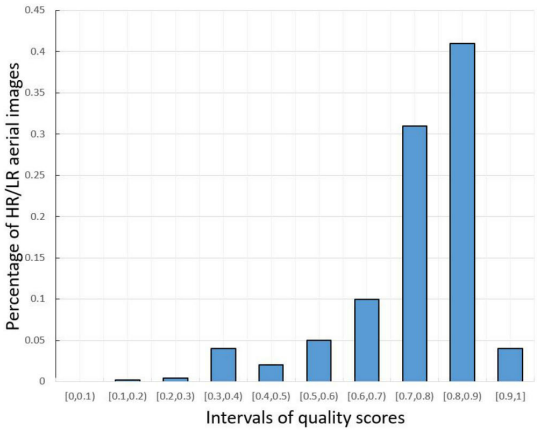

FIGURE 5. Statistics of LR/HR aerial photos with different quality scores in our complied LR/HR aerial photo set.

After collecting the million-scale LR/HR aerial photos, we have to annotate them to obtain the corresponding category labels. Herein, 106 volunteers 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 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 LR aerial photos constitute the training set while the rest are employed for testing.

B. COMPARATIVE STUDY

1) ACCURACY COMPARISON

Herein, we test our LR aerial image classification by evaluating its effectiveness and efficiency with a bunch of competitors.

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

	HR No.	LR No.		HR No.	LR No.
Tall building	1,121,240	454,165	Road	1,235,223	544,432
Residential	1,221,542	654,436	River	1,332,652	434,332
Intersection	1,367,765	355,874	Park	1,256,325	476,657
Forest	1,556,566	455,325	Palace	1,324,657	435,435
Sea	1,375,543	547,657	Factory	1,447,113	509,657
Soccer field	1,325,432	621,657	Farmland	1,324,430	432,335
Aircraft	1,417,222	317,658	Vehicle	1,118,446	454,644
Railway	1,232,532	327,628	Yacht	1,325,336	342,772
Bridge	14332,611	547,557	Swim.pool	1,254,660	376,843

We first conduct a comparative study with seven deep categorization models [23], [24], [25], [26], [27], [28], [29] that intrinsically encode some prior knowledge of different aerial photo categories. We notice that the source codes of [23], [24], [27], and [28] are publicly available. Thereby, we conduct comparative study wherein the parameter settings are set as default. For [25], [26], and [29], the source codes are unavailable to our knowledge. In this way, we re-implement them. We have tried our best 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, we launch a comparative study between our method and ten deep generic object classification models: the SPP-CNN [52], CleanNet [12], discriminative filter bank (DFB) [13], multi-layer CNN-RNN (ML-CRNN) [20], multi-label graph convolutional network (ML-GCN) [45], semantic-specific graph (SSG) [46] and multi-label transformer (MLT) [49]. Furthermore, since LR aerial photo categorization can be deemed as a sub-topic of scenery classification, we additionally experiment using three well-known scenery understanding models [22], [42], [44]. For these models, only the source codes of [22] are unavailable. Thus we re-implement them using C++.

Moreover, we compare our method with [67] and [68]. We observe that our method outperforms those aerial image classifiers not specifically designed for LR aerial photo categorization. Besides, [67] and [68] cannot encode auxiliary information from HR aerial photos. Thus their performances are inferior.

For the categorization models implemented by ourselves, the experimental setups are briefed as follows. In [25], we utilize the ResNet-152 [36] as the backbone, which is subsequently upgraded into a multi-label variant. Except for the last fully-connected layer (unit number is fixed at 13), while the remaining layers are pre-retained using the ResNet learned from ImageNet [53]. For [26], the weights in the 1536-D LSTM layer are calculated using a random number. For [29], the domain adaptation is implemented from the RSSCN7 [28] to our compiled LR\& HR aerial photo set. The ResNet101V2 [36] is employed as the backbone and the stochastic gradient descent optimizes the entire deep model. The network loss is calculated by the mean squared

TABLE 2. Accuracies with Standard Errors of the 19 Categorization Models (Experiments are repeated 20 times).

Category	EgNet [22]	AIDER [23]	MLAIC [24]	CAConvBi [25]	DLSIC [26]	GLNet [27]	DARS [28]	SPP-CNN [50]	CleanNet [12]	MSFF [67]
Tall building	0.612± 0.013	0.565±0.011	0.631±0.011	0.589±0.012	0.620±0.009	0.584 ±0.012	0.625±0.012	0.654±0.010	0.665±0.012	0.674±0.010
Residential	0.593±0.011	0.579±0.009	0.602±0.014	0.573±0.011	0.614±0.012	0.607±0.009	0.562±0.012	0.611±0.011	0.586±0.013	0.591±0.008
Intersection	0.708±0.009	0.703±0.011	0.677±0.012	0.665±0.012	0.709±0.009	0.655±0.012	0.702±0.009	0.664±0.009	0.678±0.011±	0.682±0.009
Forest	0.675±0.012	0.666±0.012	0.664±0.012	0.646±0.012	0.682±0.012	0.634±0.012	0.685±0.012	0.698±0.011	0.687±0.012	0.692±0.011
Sea	0.674±0.013	0.653±0.012	0.657±0.013	0.621±0.009	0.632±0.014	0.621±0.011	0.662±0.011	0.635±0.011	0.676±0.008	0.679±0.010
Soccer field	0.553±0.011	0.556±0.011	0.564±0.012	0.554±0.013	0.583±0.009	0.532±0.012	0.572±0.011	0.532±0.011	0.567±0.013	0.566±0.007
Aircraft	0.734±0.016	0.684±0.013	0.713±0.012	0.673±0.013	0.705±0.013	0.702±0.012	0.674±0.012	0.704±0.011	0.683±0.012	0.686±0.007
Railway	0.634±0.007	0.602±0.011	0.612±0.008	0.627±0.013	0.607±0.012	0.577±0.013	0.564±0.012	0.597±0.012	0.586±0.012	0.592±0.008
Bridge	0.557±0.012	0.552±0.013	0.563±0.009	0.558±0.014	0.548±0.012	0.565±0.012	0.552±0.012	0.546±0.012	0.577±0.012	0.583±0.010
Road	0.621±0.012	0.612±0.010	0.616±0.012	0.601±0.007	0.625±0.013	0.608±0.012	0.587±0.012	0.613±0.011	0.612±0.011	0.615±0.009
River	0.716±0.013	0.685±0.012	0.708±0.011	0.698±0.011	0.726±0.013	0.699±0.013	0.674±0.012	0.688±0.010	0.706±0.013	0.712±0.011
Park	0.661±0.017	0.644±0.015	0.654±0.013	0.676±0.012	0.673±0.013	0.685±0.011	0.654±0.010	0.675±0.012	0.668±0.010	0.666±0.011
Palace	0.671±0.012	0.626±0.013	0.654±0.013	0.613±0.013	0.626±0.014	0.647±0.014	0.636±0.009	0.623±0.011	0.605±0.011	0.609±0.006
Factory	0.632±0.013	0.612±0.012	0.586±0.010	0.602±0.011	0.627±0.013	0.612±0.012	0.587±0.012	0.586±0.012	0.608±0.012	0.604±0.009
Farmland	0.612±0.011	0.588±0.014	0.596±0.011	0.587±0.009	0.584±0.014	0.614±0.013	0.584±0.012	0.588±0.013	0.603±0.12	0.606±0.008
Vehicle	0.672±0.010	0.645±0.011	0.644±0.012	0.687±0.012	0.643±0.011	0.668±0.014	0.656±0.013	0.656±0.011	0.654±0.012	0.661±0.009
Yacht	0.693±0.012	0.706±0.013	0.696±0.010	0.719±0.012	0.703±0.011	0.708±0.013	0.705±0.012	0.688±0.011	0.697±0.012	0.694±0.005
Swim. pool	0.659±0.013	0.613±0.009	0.634±0.012	0.652±0.013	0.624±0.013	0.665±0.011	0.656±0.009	0.612±0.012	0.622±0.013	0.627±0.011
Category	DFB [13]	ML-CRNN [20]	ML-GCN [43]	SSG [44]	MLT [47]	ULSOD [21]	CNNSR [40]	MFAFVNet [42]	JADA [68]	Ours
Tall building	0.604±0.011	0.651±0.011	0.632±0.010	0.687±0.010	0.673±0.014	0.618±0.011	0.621±0.012	0.654±0.012	0.674±0.010	0.716±0.007
Residential	0.578±0.012	0.605±0.012	0.613±0.012	0.634±0.011	0.613±0.014	0.573±0.012	0.593±0.011	0.594±0.013	0.608±0.009	0.664±0.009
Intersection	0.704±0.009	0.677±0.014	0.711±0.012	0.734±0.011	0.733±0.013	0.684±0.014	0.665±0.012	0.672±0.010	0.676±0.008	0.768±0.008
Forest	0.682±0.012	0.714±0.011	0.701±0.011	0.722±0.012	0.705±0.014	0.652±0.012	0.667±0.012	0.657±0.012	0.666±0.007	0.759±0.011
Sea	0.661±0.011	0.634±0.013	0.642±0.012	0.675±0.011	0.657±0.012	0.663±0.013	0.654±0.011	0.672±0.011	0.677±0.009	0.698±0.008
Soccer field	0.574±0.010	0.543±0.012	0.573±0.011	0.573±0.011	0.583±0.014	0.562±0.014	0.543±0.010	0.536±0.009	0.541±0.007	0.617±0.010
Aircraft	0.663±0.011	0.671±0.014	0.675±0.013	0.728±0.011	0.721±0.011	0.623±0.012	0.675±0.011	0.685±0.013	0.674±0.012	0.759±0.007
Railway	0.618±0.012	0.618±0.012	0.626±0.011	0.617±0.012	0.614±0.012	0.613±0.013	0.606±0.012	0.596±0.011	0.593±0.007	0.685±0.011
Bridge	0.554±0.011	0.532±0.013	0.574±0.010	0.579±0.011	0.524±0.012	0.526±0.012	0.547±0.010	0.517±0.012	0.522±0.010	0.598±0.008
Road	0.604±0.013	0.611±0.012	0.588±0.012	0.648±0.012	0.627±0.012	0.614±0.013	0.613±0.009	0.612±0.013	0.608±0.011	0.684±0.007
River	0.713±0.011	0.706±0.010	0.713±0.013	0.714±0.011	0.705±0.013	0.672±0.012	0.654±0.013	0.665±0.013	0.662±0.008	0.748±0.008
Park	0.654±0.012	0.647±0.012	0.677±0.012	0.687±0.011	0.687±0.013	0.688±0.012	0.665±0.011	0.674±0.012	0.671±0.011	0.703±0.008
Palace	0.612±0.009	0.631±0.012	0.611±0.013	0.625±0.010	0.632±0.012	0.593±0.011	0.596±0.012	0.576±0.013	0.571±0.010	0.672±0.006
Factory	0.597±0.012	0.601±0.014	0.609±0.011	0.613±0.011	0.612±0.012	0.612±0.012	0.609±0.011	0.632±0.012	0.638±0.014	0.662±0.009
Farmland	0.582±0.011	0.587±0.012	0.576±0.012	0.616±0.012	0.613±0.011	0.585±0.013	0.565±0.011	0.612±0.013	0.617±0.008	0.627±0.010
Vehicle	0.643±0.012	0.675±0.013	0.664±0.013	0.643±0.014	0.672±0.012	0.634±0.012	0.639±0.012	0.643±0.012	0.648±0.009	0.699±0.012
Yacht	0.714±0.012	0.709±0.014	0.703±0.012	0.711±0.012	0.714±0.012	0.685±0.010	0.625±0.013	0.712±0.011	0.715±0.006	0.779±0.007
Swim. pool	0.606±0.012	0.632±0.012	0.631±0.013	0.653±0.012	0.621±0.011	0.605±0.011	0.608±0.010	0.618±0.012	0.624±0.010	0.680±0.011

error. For [22], we retrain the object bank [51] based on our refined 18 LR/ HR aerial photo categories, wherein the average-pooling strategy is applied. We employ the liblinear to solve the SVM classifier, wherein the 7-fold cross validation is utilized.

For the above 18 compared object/scene classification algorithms, we repeatedly test each model ten and the results are displayed in Table 2. 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:

1) Our method outperforms the other aerial photo categorization models remarkably due to three reasons. First, to facilitate deep model training, our competitors typically resize each original aerial photo to a fixed and much smaller size (e.g., $\mathbf{128 \times 128}$) for the subsequent hierarchical feature engineering. This hurts the learning of an LR aerial photo categorization model since many tiny but discriminative visual details will be lost. Second, expect for our method, none of the seven counterparts can select high quality features by leveraging discriminative information from HR aerial photos. Third, only our method generates GSPs sequentially capturing the semantics of LR aerial photos perceived by humans. They are further incorporated into a CPKP-based FS for calculating category labels. Comparatively, the seven counterparts only globally/locally characterize each LR aerial photo, wherein the perceptual visual features are neglected.
2) The seven generic object recognition algorithms perform inferiorly than ours because of three reasons. First, these generic recognition models generally handle medium-sized images typically containing tens of salient objects. They can hardly discover the tiny but discriminative regions inside each LR aerial photo. Second, our method can flexibly incorporate the prior knowledge of HR aerial photos. Contrastively, the seven generic object recognition models cannot encode such information. Third, by leveraging our CPKP-based FS, our method can dynamically abandon those indiscriminative regions. But the seven generic object recognition models do not have this function.
3) The three scene categorization models perform unsatisfactorily on LR aerial photos. This is because they deeply and implicitly learn a descriptive set of scene-aware semantic categories, such as ''birds'' and ''tables'', which infrequently appear on our LR aerial photo set. Moreover, the three categorization methods can successfully handle sceneries captured at horizontal view angles. But our collected LR aerial photos are captured at overhead view angles. Apparently, such view angle gap will decrease the categorization accuracy.

TABLE 3. Training/testing Time of the 18 Categorization Models (Each Bold Number Represents the Best Result).

	[23]	[24]	[25]	[26]	[27]	[28]	[29]	SPP-CNN	CleanNet
Train	31h7m	43h14m	52h21m	39h23m	36h43m	46h13m	41h32m	26h33m	38h22m
Test	1.143s	1.774s	1.846s	1.564s	2.437s	1.463s	1.675s	0.893s	1.660s
	DFB	ML-CRNN	ML-GCN	SSG	MLT	[22]	[42]	[42]	Ours
Train	40h23m	25h25m	32h15m	44h16m	31h16m	32h14m	35h44m	32h12m	27h21m
Test	1.213s	1.002s	1.875s	0.983s	1.436s	1.774s	1.983s	1.546s	0.477s

2) TRAINING/TESTING TIME COMPARISON

It is generally acknowledged that time consumption is a key criterion reflecting the performance of a classification algorithm. Then, 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 [45], [52] are much simpler than ours. Meanwhile, we observe that the per-category accuracies of [45], [52] are noticeably lower than ours. For the testing time comparison, our method can be conducted at a significantly faster speed than all 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 LR aerial photo categorization pipeline involves three key modules: 1) GSP learning using the deep low-rank algorithm, 2) CPKP-based FS, and 3) feature classification for category labels. During training, the time consumed for each module is: 9h12m (m1), 10h11m (m2), and 3h58m (m3). During testing, the time cost of each module is: 77ms (m1), 3ms (m2), and 12ms (m3). We observe that most of the training time is spent for module 1 and practically this can be accelerated by Nvidia GPUs.

C. EVALUATION BY TUNING PARAMETERS

There are two sets of tunable parameters to be evaluated. The first set denotes the weights balancing multiple attributes in the low-rank algorithm, i.e., α, β, and γ, as well as the deep layer number L. The second set includes the polynomial kernel degree Q and the target dimensionality for CPKP-based FS V. Herein, we report the LR aerial photo categorization accuracy by adjusting the two sets of parameters.

To analyze the first set of parameters, we set the default values of α, β, γ and L to 0.3, 0.1, 0.15, and 7 respectively. In our implementation, the default values are determined by 10-fold cross validation. Herein, the validation set contains 54000 samples, which is constituted by selecting 3000 LR aerial photos from each category. More concretely, we tune each of α, β, and γ from 0.05 to one with a step of 0.05. And all the possible parameter combinations are enumeratively employed to test the LR 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 0.05 to one with step of 0.01, wherein the corresponding categorization accuracy is reported. As the three curves displayed on the top of Fig. 6, the three parameters consistently increase stably and then peak. Afterward, they all decrease to a low level. Such monotonicity properties indicate the feasibility to tune the three parameters toward an optimal level in practice. As shown at the bottom of Fig.6, the accuracy increases stably when L is increased from one to 7. Thereafter, the accuracy maintains stably. We notice that a deeper categorization model indicates more parameters to be learned, which may cause model overfitting practically. Thus we set $L = 7$.

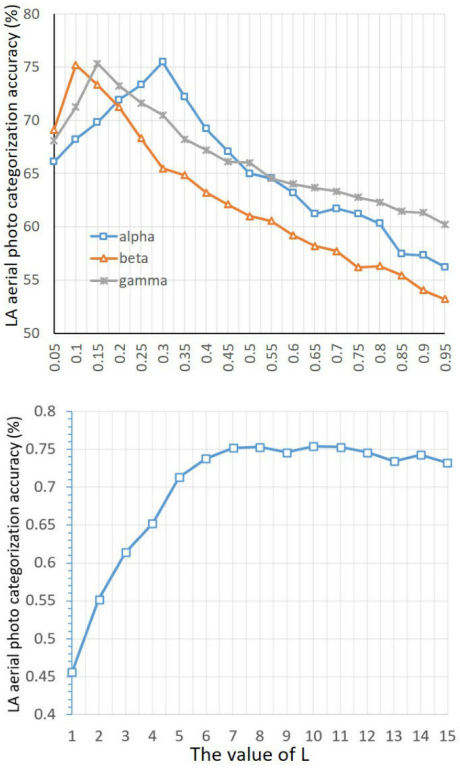

FIGURE 6. LR aerial photo categorization accuracies by varying α, β, and γ (top) and L (bottom).

D. ABLATION STUDY

As aforementioned, our method is comprised of two key modules: 1) GSP learning using the low-rank algorithm, 2) semi-supervised 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 to elaborate the underlying reasons for the observed results.

TABLE 4. LR aerial photo categorization accuracy increment/decrement.

	S1	S2
O1	-2.332%	-3.339%
O2	-5.430%	-2.332%
O3	n/a	-1.875%

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 (10) (marked by "S11"). Afterward, we abandon the regularizer $\|\boldsymbol{Z}\|_F^2 + \sum_{i=1}^{L}(\|\boldsymbol{Z}_l\|_F^2 + \|\boldsymbol{\xi}\|_2^2)$ in (10) (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, abandon the regularizer converts the deep feature learning to a shallow one. And a shallow feature engineering module will cause a sharp performance decrement. Also, removing the sparse constraint 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 such function and use the full feature set for LR aerial image categorization (S21). Then, we remove the two terms $\sigma \left\|\mathrm{X}^{\mathrm{T}}\mathrm{Q} - \mathrm{P}\right\|$ (S22) and $\tau\mathcal{R}(Q)$ (S23) respectively and report the categorization accuracies. As shown in 4, abandoning the FS module causes the largest categorization accuracy drop. This demonstrates the importance of feature selection in LR aerial image categorization.

V. CONCLUSION

Recognizing aerial images is an indispensable application in deep neural networks [9], [37], [38], [39], [40], [41], [62],. We proposed a novel LR aerial photo categorization pipeline, wherein deep perceptual features are extracted and refined by propagating the prior knowledge of HR aerial photos into LR ones. Our work includes three key modules: 1) a deep low-rank algorithm that learns deep features from LR/HR aerial images; 2) a novel SPFS-based FS that selects high quality features on the high-order feature space, and 3) a kernel SVM classifier that is learned from the selected features. Experiments shown the competitiveness of our approach.

One shortcoming of our categorization pipeline is that the feature selection is conducted in the original feature space, whereas practically the deep GSP features might be distributed in the nonlinear high-order feature space. In the future, we plan to design a high-order feature selection algorithm to further enhance the quality of the selected features.

REFERENCES

[1] 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.
[2] 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.
[3] 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.
[4] 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.
[5] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.
[6] 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.
[7] 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, Jun. 2010.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] B. Ahadzadeh, M. Abdar, F. Safara, A. Khosravi, M. B. Menhaj, and P. N. Suganthan, "SFE: A simple, fast and efficient feature selection algorithm for high-dimensional data," *IEEE Trans. Evol. Comput.*, early access, Jan. 23, 2023, doi: 10.1109/TEVC.2023.3238420.
[15] J. Lu, S. Yi, J. Zhao, Y. Liang, and W. Liu, "Interpretable robust feature selection via joint $\ell_{1,2}$-norms minimization," in *Proc. 13th Int. Conf. Mach. Learn. Comput.*, Feb. 2021, pp. 1–9.
[16] N. Gui, D. Ge, and Z. Hu, "AFS: An attention-based mechanism for supervised feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 3705–3713.
[17] D. Ming and C. Ding, "Robust flexible feature selection via exclusive $\ell_{1,2}$ regularization," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3158–3164.
[18] 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.
[19] 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.
[20] A. Caglayan and A. Can, "Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 675–688.
[21] C. Shi, Q. Ruan, G. An, and R. Zhao, "Hessian semi-supervised sparse feature selection based on $\ell_{2,1/2}$-matrix norm," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 16–28, Jan. 2015.
[22] 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. 1–16.
[23] 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.
[24] 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.

[25] 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.
[26] 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.
[27] M. Pritt and G. Chern, "Satellite image classification with deep learning," in *Proc. IEEE Appl. Imag. Pattern Recognit. Workshop (AIPR)*, Oct. 2017, pp. 1–7.
[28] 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.
[29] 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.
[30] L. Song, A. Smola, A. Gretton, J. Bedo, and K. Borgwardt, "Feature selection via dependence maximization," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 1393–1434, Jan. 2012.
[31] M. Masaeli, G. Fung, and J. Dy, "From transformation-based dimensionality reduction to feature selection," in *Proc. Int. Conf. Mach. Learn.*, 2010, pp. 751–758.
[32] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural Comput.*, vol. 26, no. 1, pp. 185–207, Jan. 2014.
[33] J. Chen, M. Stern, M. Wainwright, and M. Jordan, "Kernel feature selection via conditional covariance minimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–10.
[34] L. Leng, M. Li, C. Kim, and X. Bi, "Dual-source discrimination power analysis for multi-instance contactless palmprint recognition," *Multimedia Tools Appl.*, vol. 76, no. 1, pp. 333–354, Jan. 2017.
[35] L. Leng and J. Zhang, "PalmHash code vs. PalmPhasor code," *Neurocomputing*, vol. 108, pp. 1–12, May 2013.
[36] 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.
[37] W. Mu and B. Liu, "Voice activity detection optimized by adaptive attention span transformer," *IEEE Access*, vol. 11, pp. 31238–31243, 2023.
[38] 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.
[39] 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.
[40] 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.
[41] 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.
[42] 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.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] 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)*, Jun. 2004, pp. 17–20.
[48] H. Zhang, B. Li, J. Zhang, and F. Xu, "Aerial image series quality assessment," *IOP Conf. Ser., Earth Environ. Sci.*, vol. 17, Mar. 2014, Art. no. 012183.
[49] 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.
[50] R. Diestel, *Graph Theory*. New York, NY, USA: Springer-Velag, 2005.
[51] L. Li, H. Su, E. Xing, and F. Li, "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.
[52] 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.
[53] 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.
[54] V. Chalavadi, P. Jeripothula, R. Datla, S. B. Ch, and C. K. Mohan, "mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognit.*, vol. 126, Jun. 2022, Art. no. 108548.
[55] S. Zhou, J. Irvin, Z. Wang, E. Zhang, J. Aljubran, W. Deadrick, R. Rajagopal, and A. Ng, "DeepWind: Weakly supervised localization of wind turbines in satellite imagery," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2009, pp. 1–5.
[56] 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.
[57] 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.
[58] 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.
[59] 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.
[60] 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.
[61] 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.
[62] 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.
[63] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," 2016, *arXiv:1605.08323*.
[64] 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.
[65] F. Xiao, L. Pang, Y. Lan, Y. Wang, H. Shen, and X. Cheng, "Transductive learning for unsupervised text style transfer," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 2510–2521.
[66] E. Fix and J. Hodges, "Discriminatory analysis nonparametric discrimination: Consistency properties," *Int. Statistical Rev./Revue Internationale Statistique*, vol. 57, no. 3, pp. 238–247, 1989.
[67] J. Chen, J. Yi, A. Chen, and Z. Jin, "EFCOMFF-Net: A multiscale feature fusion architecture with enhanced feature correlation for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604917.
[68] X. Tang, C. Li, and Y. Peng, "Unsupervised joint adversarial domain adaptation for cross-scene hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5536415.

YINHAI LI is currently an Associate Professor with Jinhua Polytechnic. His research interests include visual modeling and image processing.

YICHUAN SHENG is currently a Lecturer with Jinhua Polytechnic. His research interests include computer vision and image processing.