

RESEARCH ARTICLE

Categorizing Low-Resolution Aerial Photos by Hessian-Regularized Perceptual Feature Selection

GUIFENG WANG^{1,2}, JIANZHANG XIAO^{1,2}, AND YI YANG^{1,2}¹Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321017, China²College of Computer Sciences, Beijing Technology and Business University, Beijing 100048, China

Corresponding author: Guifeng Wang (guifengw@gmail.com)

This work was supported in part by the Science and Technology Program of Jinhua, China under Grant 2020-1-004a, in part by the Basic Public Welfare Research Project of Zhejiang Province under Grant LGG19E050010, and in part by the Special Project of Center for Scientific Research and Development in Higher Education Institutes of Ministry of Education under Grant ZJXF2022174.

ABSTRACT Amid advancements in aerospace technology and remote communication, a proliferation of Earth-observing satellites has been launched, creating a distinction between high- and low-altitude platforms. High-altitude satellites capture low-resolution (LR) aerial images, covering expansive areas, whereas their low-altitude counterparts provide high-resolution (HR) images of relatively confined spaces. The task of semantically categorizing LR aerial imagery is pivotal within numerous artificial intelligence (AI) systems but is encumbered by challenges including limited availability of labeled training data and the complexity of approximating human environmental perception through computational models. To address these challenges, this research proposes a novel strategy that marries active perception learning with Hessian-regularized feature selection (HRFS). This approach endeavors to procure perceptually and discriminatively potent visual representations for classifying LR aerial imagery. By emulating the human propensity to sequentially engage with salient regions within a visual field, an active learning paradigm is adopted to differentiate between salient and non-salient regions within LR aerial images. Theoretically, this methodology ensures that selected salient regions can reconstruct the aerial imagery in its entirety, thus mirroring the human visual system's perception. Following this, a pioneering HRFS technique is devised to extract premium features from these selectively identified salient regions, distinguished by its semi-supervised operation, the capability for concurrent linear classifier training, and the preservation of the geometric distribution of samples within the feature space. Empirical assessments underscore the resilience and efficacy of the proposed classification framework.


INDEX TERMS Perception learning, manifold regularizer, selecting features, low-resolution, active feature.

I. INTRODUCTION

Owing to advancements in aerospace engineering, remote sensing technology, and telecommunication, there has been a notable surge in the deployment of Earth observation satellites. These satellites can broadly be classified into two categories: high-altitude and low-altitude satellites. High-altitude satellites are characterized by their extensive coverage area, significantly surpassing that of their low-altitude counterparts. In contemporary applications, the precise semantic interpretation of low-resolution (LR) aerial

imagery has emerged as a crucial element within various intelligent systems.

The scholarly domain has seen the introduction of a multitude of visual categorization and annotation algorithms aimed at interpreting aerial imagery across a spectrum of resolutions. Prominent methodologies can be broadly classified into three distinct groups: 1) Multiple Instance Learning (MIL) and CNN-guided region localization utilizing weak supervision [43], [44]; 2) Semantically-aware graph models for aerial image parsing [3], [4]; and 3) Intricately designed hierarchical models for the annotation of aerial photographs [5], [6], [7]. Nonetheless, in the context of operational intelligent systems, these existing paradigms encounter

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

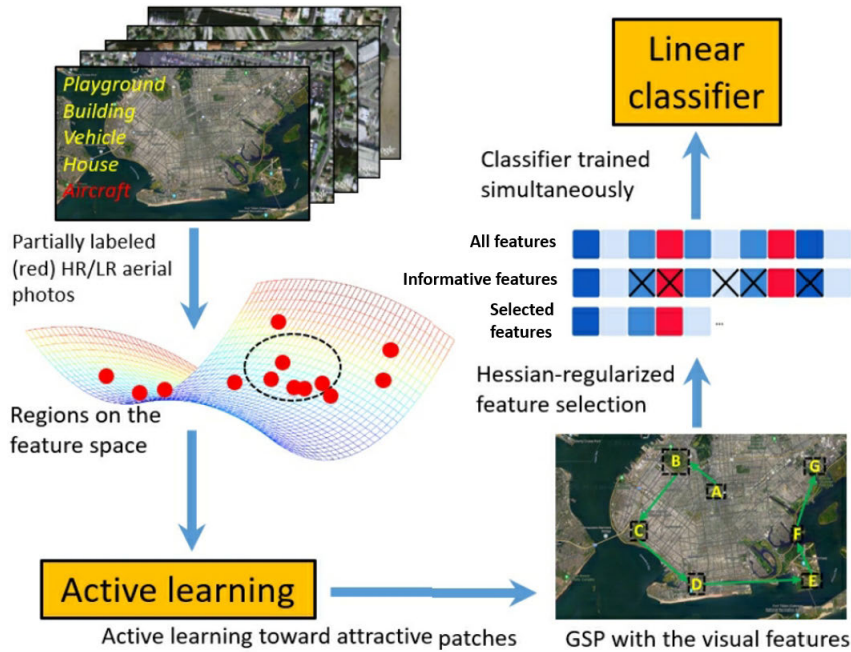


FIGURE 1. An overview of categorizing aerial photo with low resolution.

limitations in accurately representing low-resolution (LR) aerial images due to several critical challenges:

- The presence of numerous visually appealing objects or their components within an aerial image necessitates a biologically-inspired computational approach capable of simulating human perception of salient regions. Developing a deep learning algorithm that effectively identifies visually prominent areas while enhancing their representation poses significant hurdles. These include: i) determining the sequence of human gaze allocation towards attractive image segments (Gaze Shift Paths or GSPs), ii) mitigating the impact of inherent noisy labels within large-scale training datasets, and iii) translating image-level semantic labels to discrete patches within each LR aerial image;
- Differing from high-resolution aerial images, LR aerial imagery often exhibits reduced visual quality, influenced by external variables such as uncontrollable weather conditions. This results in a scarcity of annotated LR aerial images in contrast to their high-resolution counterparts, complicating the task of developing a feature selection algorithm trained on partially annotated LR datasets. Challenges include bridging the inherent relationship between LR and HR aerial images within a high-dimensional manifold space.

To address these obstacles, we introduce a novel Hessian-Regularized Feature Selector (HRFS) that leverages actively learned human gaze behavior from HR aerial images to enhance the classification of LR aerial images. Figure 1 provides an overview of our proposed framework, which utilizes a comprehensive collection of HR and LR

aerial images (including partially unannotated samples). The algorithm maps regions from aerial images into a feature space and then, to mimic human cognitive processes in interpreting aerial imagery, employs an active learning strategy to segregate each image into a series of visually compelling patches alongside less attractive background patches. Concurrently, it computes the Gaze Shift Path (GSP) and its associated visual features. To distill a subset of highly representative GSP features across images of varying resolutions, HRFS is employed to identify discriminative features, leveraging a scenario where only a limited number of samples are labeled. Theoretically, HRFS preserves the geometric distribution of both LR and HR aerial images within the feature space and facilitates the concurrent learning of a linear classifier. A comparative analysis against 17 recognized visual categorization algorithms underscores the efficacy of our approach.

This research introduces two primary innovations: a) the deployment of an active learning algorithm for the sequential generation of GSPs from LR aerial images, and b) the HRFS, which selects high-quality features in a semi-supervised manner, ensuring the preservation of sample geometric distribution while enabling joint classifier training.

II. RELATED METHODS TO OURS

A plethora of computational visual models has been devised for the analysis of aerial imagery. To encapsulate the entirety of an image semantically, Lum et al. [46] introduced a topology-based visual schema delineating binary region-wise connections across various aerial images, facilitating a kernel-guided feature synthesis for comprehensive

global image capture and subsequent recognition. Xia and colleagues [48] advanced a framework utilizing weakly labeled training data to annotate high-resolution (HR) aerial imagery semantically. Akar et al. [50] ingeniously merged the renowned random forest algorithm with an object-oriented visual representation learner for the classification of remote sensing images. Sameen et al. [51] crafted a multi-tiered visual architecture for determining multiple labels across diverse HR urban aerial images. Chenggong and co-researchers [47] employed a predefined five-layer Convolutional Neural Network (CNN) for classifying high-definition remote sensing images, proposing a novel domain-level adjustment to refine the deep model for specific scenery. Danfeng et al. [28] conceptualized a multimodal learning algorithm for simultaneous annotation of HR aerial imagery. Cai Weiwei and team [8] developed an inter-visual-attention mechanism to weight the representations of aerial photographs.

Despite their effectiveness in classifying multi-resolution aerial images, these image-level visual models encounter limitations in optimally modeling low-resolution (LR) aerial images, primarily due to the blurring of tiny yet discriminative objects. To accurately detect and capture objects of various scales, a region-level modeling technique becomes essential for localizing small objects within each LR aerial image.

In related research, a group sparsity regularizer was introduced to robustly recognize human faces by proposing an upper-bounded function to enhance the l_1 -norm for sparsity, effectively addressing bias and outlier impacts [58]. A novel approach to incomplete multi-view clustering was formulated, upgrading incomplete similarity graphs and learning complete tensor representations [59]. Dehkordi et al. [53] developed a framework for generating remote sensing imagery to monitor surface water extent changes, leveraging Landsat images processed on the GEE cloud platform, with K-means initialized by Fmask. Another study evaluated post-processing schemes for object detection within aerial images, determining optimal strategies based on the average F-score metric [54].

For regional aerial image characterization, a multi-layer deep learning model was designed to detect visually salient foreground objects [2]. A hierarchical model based on focal loss was developed for precise car localization within LR and HR aerial photos [1]. A geographic object detection model tailored for HR aerial imagery was proposed, focusing on intelligent extraction of intersections and roads [52]. Moreover, a combination of feature engineering and soft-label computation was suggested for constructing effective visual detectors for aerial imagery analysis [9].

Our aerial image recognition methodology distinguishes itself by being biologically inspired and closely mirroring the human visual perceptual process. Although the aforementioned region-level image models proficiently utilize representative regions of multiple sizes from each LR aerial photo, they exhibit limitations such as domain-specific design

constraints, inability to explicitly identify visually or semantically salient regions for LR aerial photo representation, and lack of principled feature selection or explicit encoding of geometric sample structure during feature engineering.

III. PROPOSED WORK

A. ACTIVELY HUMAN GAZE MODELING

In the realm of low-resolution (LR) aerial photography, each photograph comprises numerous image patches that do not significantly contribute to understanding its semantic categorizations. These patches, typically constituting the less visually appealing background, fail to captivate human attention. To devise a proficient model for categorizing LR aerial photos, we employ an active learning algorithm to select a subset of semantically rich image patches from within an aerial photograph.

Theoretically, an optimally designed machine learning model should accurately discern the concealed distribution of samples. Given the semantic correlation among spatially adjacent image patches, it becomes possible to linearly represent each patch with the help of its neighboring patches. This premise allows for the calculation of reconstruction parameters, thereby facilitating a method to optimally capture and represent the underlying sample distribution through a refined selection of image patches that more accurately reflect the semantic content of the aerial photograph.

$$\begin{aligned} & \arg \min_{\mathbf{R}} \sum_{i=1}^N \left\| y_i - \sum_{j=1}^N \mathbf{T}_{ij} y_j \right\| \\ & s.t. \sum_{j=1}^N \mathbf{R}_{j=1}^N = 1, \mathbf{T}_{ij} = 0 \text{ if } x_j \notin \mathcal{B}(x_i), \end{aligned} \quad (1)$$

In this formulation, $\{x_1, x_2, \dots, x_N\}$ represents the visual features of N image patches, where \mathbf{T}_{ij} measures the significance of each image patch in reconstructing its spatially neighboring patches. Here, N signifies the total count of patches within an aerial photograph, and $\mathcal{B}(y_i)$ denotes the neighborhood of patches surrounding the i -th image patch.

To evaluate the visual representativeness of the selected image patches, we introduce a reconstruction algorithm utilizing the parameters described above. An error metric is employed to assess the effectiveness of our selected image patches. Let $\{b_1, b_2, \dots, b_N\}$ represent the reconstructed image patches, which are derived through the application of the following objective function:

$$\begin{aligned} & \epsilon(b_1, b_2, \dots, b_N) \\ & = \sum_{i=1}^L \|a_{s_i} - b_{s_i}\|^2 + \mu \sum_{i=1}^N \left\| a_i - \sum_{j=1}^N \mathbf{R}_{ij} a_j \right\|^2, \end{aligned} \quad (2)$$

In this context, μ serves to determine the significance of our regularization term, while L specifies the quantity of selected visually appealing patches. The collection $\{a_{s_1}, a_{s_2}, \dots, a_{s_K}\}$ comprises the visually significant

patches, with $\{s_1, s_2, \dots, s_K\}$ indicating the indices of these chosen patches. Notably, the first part of the equation is responsible for quantifying the rectification cost of these chosen aesthetically appealing patches. Concurrently, the second part ensures that the reconstructed image patches retain the same distribution as the original input samples.

Consider $\mathbf{C} = [c_1, c_2, \dots, c_N]$ and $\mathbf{d} = [d_1, d_2, \dots, d_N]$ representing the original and reconstructed patch arrays, respectively. We define Υ as a diagonal matrix where each diagonal element is assigned a value of one if its index i belongs to $\{r_1, r_2, \dots, r_K\}$; otherwise, it is set to zero. Leveraging this setup, we proceed to reformulate the objective function accordingly.

$$\epsilon(\mathbf{D}) = \text{tr} \left((\mathbf{B} - \mathbf{C})^T \Upsilon (\mathbf{B} - \mathbf{C}) \right) + \mu \text{tr}(\mathbf{B}^T \mathbf{C} \mathbf{A}), \quad (3)$$

In this context, \mathbf{D} is defined as $(\mathbf{I} - \mathbf{R})^T (\mathbf{I} - \mathbf{R})$, where \mathbf{I} represents the identity matrix and \mathbf{R} the reconstruction coefficient matrix. Our goal is to minimize the objective function $\epsilon(\mathbf{A})$. To achieve this, we equate the gradient of $\epsilon(\mathbf{A})$ with respect to \mathbf{A} to zero, which allows us to derive the optimal values for \mathbf{A} .

$$\Upsilon (\mathbf{B} - \mathbf{C}) + \mu \mathbf{C} \mathbf{B} = 0. \quad (4)$$

Then, our patches are as follows:

$$\mathbf{A} = (\mu \mathbf{D} + \Upsilon)^{-1} \Upsilon \mathbf{B}, \quad (5)$$

In this way, we obtain:

$$\begin{aligned} \epsilon(b_{r_1}, \dots, b_{r_K}) &= \|\mathbf{C} - \mathbf{A}\|_F^2 = \left\| \mathbf{B} - (\mu \mathbf{D} + \Upsilon)^{-1} \Upsilon \mathbf{B} \right\|_F^2 \\ &= \|(\mu \mathbf{C} + \Upsilon) \mu \mathbf{D} \mathbf{X}\|_F^2, \end{aligned} \quad (6)$$

In this context, the Frobenius norm of a matrix is denoted by $\|\cdot\|_F^2$. Given the combinatorial nature of the problem, directly minimizing the objective function as presented may not be computationally feasible. To address this challenge, we employ a sequential selection process. We start by considering a subset of selected image patches within an aerial image, represented as $\{b_{s_1}, \dots, b_{s_{K'}}\}$. We then define Υ_n as a diagonal matrix of size $N \times N$, where N is the total number of patches, and Ψ_i , another $N \times N$ matrix, where diagonal elements are set to one, and off-diagonal elements are zero. The selection of the next image patch, $s_{K'+1}$, is determined by optimizing a specific criterion that takes these matrices into account.

$$r_{L'+1} = \arg \min_{i \notin \{r_1, \dots, r_{K'}\}} \left\| (\mu \mathbf{C} + \Upsilon_n + \Pi_i)^{-1} \mu \mathbf{D} \mathbf{X} \right\|_F^2. \quad (7)$$

Practically, the matrix \mathbf{D} in the context of optimizing the selection of the next image patch (as described in the previous equation) is sparse. To expedite the computation involved in inverting this matrix, we utilize the Sherman-Morrison formula, a well-established mathematical principle cited from the literature on matrix computations [49]. This approach significantly streamlines the calculation process, enabling

more efficient determination of subsequent image patch selections within the framework.

$$(\mu \mathbf{C} + \Upsilon_n + \Pi_i)^{-1} = \mathbf{J} - \frac{\mathbf{D}_{*i} \mathbf{D}_{i*}}{1 + \mathbf{D}_{ii}}, \quad (8)$$

In the mathematical formulation, \mathbf{J}_{*i} and \mathbf{J}_{i*} denote the i -th column and the i -th row of the matrix \mathbf{J} , respectively. These notations facilitate the manipulation of specific elements within the matrix, crucial for the computation of the updated objective function as outlined in equation (7).

$$\begin{aligned} & \left\| (\mu \mathbf{D} + \Upsilon_n + \Psi_i)^{-1} \mu \mathbf{D} \mathbf{B} \right\|_F^2 \\ &= \mu^2 \text{tr}(\mathbf{J} \mathbf{D} \mathbf{B} \mathbf{B}^T \mathbf{D} \mathbf{J}) \\ & \quad - \frac{2\mu^2 \mathbf{D} \mathbf{B} \mathbf{B}^T \mathbf{D} \mathbf{B}_{*i}}{1 + \mathbf{J}_{ii}} + \frac{\mu^2 \mathbf{J}_{i*} \mathbf{J}_{*i} \mathbf{D} \mathbf{B} \mathbf{B}^T \mathbf{D} \mathbf{B}_{*i}}{(1 + \mathbf{J}_{ii})^2}, \end{aligned} \quad (9)$$

By defining \mathbf{M} as $\mathbf{D} \mathbf{B} \mathbf{B}^T \mathbf{D}$, the optimization problem outlined in (7) is transformed.

$$\begin{aligned} r_{L'+1} &= \arg \min_{i \notin \{s_1, \dots, s_{K'}\}} \frac{1}{1 + \mathbf{C}_{ii}} \left(\frac{\mathbf{C}_{i*} \mathbf{J}_{*i} \mathbf{J}_{i*} \mathbf{M} \mathbf{J}_{*i}}{1 + \mathbf{J}_{ii}} \right. \\ & \quad \left. - 2 \mathbf{J}_{i*} \mathbf{M} \mathbf{J} \mathbf{J}_{*i} \right). \end{aligned} \quad (10)$$

Leveraging the formulation provided by (10), it becomes feasible to sequentially identify the K most visually compelling patches within each aerial image. These patches, when arranged sequentially, give rise to a Gaze Shift Path (GSP) that mirrors the natural progression of human visual attention across different segments of aerial imagery, a process illustrated in Fig. 1. To encapsulate the information contained within each GSP, 128-dimensional features are extracted from each image patch using a CNN architecture, as noted in [21]. These features are then amalgamated to form a comprehensive 128 K -dimensional vector, effectively capturing the essence of each GSP in a manner that aligns with human visual processing patterns.

B. HESSIAN-REGULARIZED FEATURE SELECTION

Let us represent $\mathbf{G} = [g_1, \dots, g_N] \in \mathbb{R}^{N \times T}$ as the matrix holding the deep Gaze Shift Path (GSP) features derived from all training samples. By convention, the first L aerial images are labeled, denoted as $\mathbf{G}_L = [g_1, \dots, g_L]$, whereas the remaining images are unlabeled, represented as $\mathbf{G}_U = [g_{L+1}, \dots, g_N]$. Additionally, $\mathbf{L} = [l_1, \dots, l_L]$ signifies the label matrix for the L labeled instances.

The matrix $\mathbf{P} \in \mathbb{R}^{T \times C}$ is introduced to delineate the selection of features, with C indicating the total number of aerial image categories. Based on this setup, the objective function for our feature selection mechanism is formulated as follows:

$$\min_{\mathbf{P}} \epsilon(\mathbf{P}) + \varphi \cdot \eta(\mathbf{P}), \quad (11)$$

In this formulation, $\epsilon(\mathbf{P})$ signifies the designated loss function, with $\varphi \cdot \eta(\mathbf{P})$ representing the regularization term, where φ is the regularization parameter.

An affinity graph is denoted as \mathbf{W} , where each element \mathbf{W}_{ij} measures the similarity between samples i and j within the graph. We assign $\mathbf{W}_{ij} = 1$ if samples i and j are adjacent in the feature space, signifying a spatial or semantic closeness; otherwise, $\mathbf{W}_{ij} = 0$ for non-neighboring samples. Moreover, \mathbf{D} is introduced as a diagonal matrix where $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, leading to the construction of the graph Laplacian matrix $\mathbf{K} = \mathbf{D} - \mathbf{W}$. Preserving the structure of \mathbf{K} during the feature selection process ensures the maintenance of the geometric relationships among the samples, as illustrated in Fig. 2.

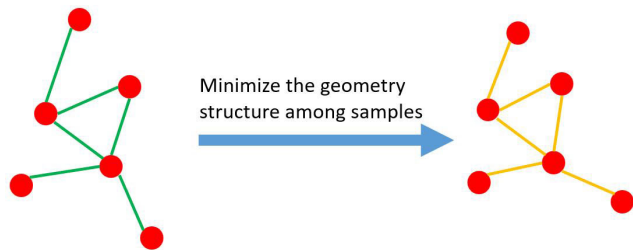


FIGURE 2. An example of preserving the relative positions among samples during feature selection.

Our approach aims to simultaneously account for both labeled and unlabeled samples, employing transductive learning principles [11] to predict labels for the LR aerial images. Denoting $\mathbf{J} = [j_1, j_2, \dots, j_N]^T \in \mathbb{R}^{N \times C}$ as the matrix of predicted labels for all training samples, where j_i represents the predicted label vector for the i -th sample, the optimal \mathbf{J} should align well with the true labels \mathbf{L} for labeled samples and adhere to the structural constraints imposed by the affinity graph \mathbf{W} . Thus, the goal is to optimize an objective function that reflects both label fidelity and graph consistency, allowing \mathbf{J} to accurately predict the label distribution across all training samples, leveraging the framework established in [11].

$$\arg \min_{\mathbf{J}} \text{tr}(\mathbf{J}^T \mathbf{K} \mathbf{J}) + \text{tr}((\mathbf{J} - \mathbf{L})^T \mathbf{V} (\mathbf{J} - \mathbf{L})), \quad (12)$$

The matrix \mathbf{V} , defined as diagonal, is established based on a specific decision criterion relating to the labeling status of samples. For labeled samples, \mathbf{V} is assigned a value approximating infinity (implemented as 1×10^{10}), while for unlabeled samples, \mathbf{V} is set to 1. This criterion is devised to ensure that the predicted labels closely match the actual ground truth by emphasizing the significance of labeled samples in the optimization process.

In aiming to minimize the discrepancy between predicted and actual labels, especially for unlabeled aerial images, the process involves adapting feature selection to account for the semi-supervised context provided by our affinity graph. This approach allows for the nuanced incorporation of both labeled and unlabeled data, facilitating a more informed and accurate prediction of labels across the dataset.

$$\arg \min_{\mathbf{J}, \mathbf{G}} \text{tr}(\mathbf{J}^T \mathbf{K} \mathbf{J}) + \text{tr}((\mathbf{J} - \mathbf{L})^T \mathbf{V} (\mathbf{J} - \mathbf{L})) + \beta \|\mathbf{G}^T \mathbf{P}\|_F^2 + \varphi \cdot \eta(\mathbf{P}), \quad (13)$$

Algorithm 1 LR Aerial Image Classification Framework Training and Testing

- 1: **Training Phase:**
- 2: **Input:** Set of N partially labeled LR and HR aerial images, K , μ , β , and φ .
- 3: **Output:** Trained LRAL model, selected feature indices, and parameters for a linear SVM classifier.
- 4: Train the active learning algorithm as per equation (6), identify GSPs, and compute deep GSP features for each LR aerial image.
- 5: Determine the indices for feature selection using solution to equation (13).
- 6: Derive the linear SVM classifier parameters, denoted as \mathbf{P} .
- 7: **Testing Phase:**
- 8: **Input:** A test LR aerial image, the pre-trained active learning model, feature selection indices, and the linear SVM classifier.
- 9: **Output:** Predicted category labels for the test image.
- 10: Generate the deep GSP feature for the test image using the pre-trained active learning model.
- 11: Apply feature selection with \mathbf{P} on the deep GSP features and predict category labels.

In this framework, β serves as the weighting factor for the regularization term. The initial segments of the objective function aim to ensure that, within the semi-supervised learning framework, the predicted labels for aerial images, represented by \mathbf{J} , align closely with both the ground truth and the constructed affinity graph. The regularization term, $\varphi \cdot \eta(\mathbf{P})$, is strategically implemented to induce a high degree of sparsity in the feature selection matrix \mathbf{P} . Additionally, the term $\beta \|\mathbf{G}^T \mathbf{P}\|_F^2$ introduces a penalty for label prediction errors, facilitating the simultaneous optimization of the linear classifier (i.e., \mathbf{P}) and the predicted labels (i.e., \mathbf{J}).

In the realm of machine learning, the design of feature selection algorithms often incorporates a variety of regularizers to optimally identify high-quality features, with a common choice being the $l_{2,p}$ -norm to enforce feature sparsity. Extensive empirical analyses, as referenced in [19], demonstrate that applying the $l_{2,1/2}$ -norm significantly enhances the robustness and discriminative power of the selected features. Theoretical considerations designate the $l_{2,1/2}$ -norm as a Hessian regularizer when $p = 1/2$, leading to a refined formulation of the objective function to leverage this advanced regularization approach for feature selection.

$$\arg \min_{\mathbf{J}, \mathbf{G}} \text{tr}(\mathbf{J}^T \mathbf{K} \mathbf{J}) + \text{tr}((\mathbf{J} - \mathbf{L})^T \mathbf{V} (\mathbf{J} - \mathbf{L})) + \beta \|\mathbf{G}^T \mathbf{P}\|_F^2 + \varphi \cdot \|\mathbf{P}\|_{2,1/2}^{1/2}, \quad (14)$$

The optimization of the $l_{2,1/2}$ -norm within the context of the Hessian regularizer introduces non-convexity into the problem. To address this, we employ an iterative solution approach, as outlined in [45], to effectively navigate the

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. Statistical overview of high-resolution and low-resolution aerial photographs in our dataset.

challenges posed by the non-convex landscape. This iterative method allows for the gradual refinement of solutions, ensuring that the feature selection process under the Hessian regularized framework is both feasible and practical. Consequently, the entire process of categorizing aerial images via our Hessian-regularized Feature Selection (HRFS) approach is encapsulated in Algorithm 1, providing a structured methodological outline for implementation.

IV. EMPIRICAL EVALUATION

A. IMAGE COLLECTION

In our endeavor to semantically annotate a vast collection of Low-Resolution (LR) and High-Resolution (HR) aerial images gathered from the Internet (as detailed in the statistical analysis presented in Fig. 3), we enlisted the help of 82 volunteers to manually label 14.7% of the LR aerial images for each major city within our dataset. This process utilized a total of 47 distinct image-level labels. Subsequently, a multi-class classification model, such as Support Vector Machine (SVM), was trained and applied to infer the image-level labels for the unlabeled portion of the dataset. These automatically assigned labels were then meticulously reviewed by the volunteers for accuracy. It was observed that several image-level labels were associated

with a relatively small number of aerial images, posing a challenge for the development of a robust classification algorithm. Consequently, labels representing fewer than 220,000 LR aerial images were excluded from further analysis, resulting in a final selection of 18 label categories. Following this curation, 99.973% of the aerial images, across both resolutions, were found to contain fewer than 4 labels, with the remainder bearing more complex label associations. These images often featured numerous small patches (less than 210×210 pixels) that could introduce noise into the dataset; hence, they were omitted from the study. The remaining LR and HR aerial images were systematically organized, with the dataset split such that the first half of the images for each label category was used for training our model, while the latter half was reserved for evaluation purposes.

1) COMPARATIVE ACCURACIES

In assembling our extensive dataset of Low-Resolution (LR) and High-Resolution (HR) aerial images (statistics illustrated in Fig. 3), it becomes crucial to semantically annotate these images. Eighty-two volunteers manually labeled 14.7% of LR aerial images for each major city in our dataset, utilizing a total of 47 distinct image-level labels. Subsequently, a multi-class classifier such as SVM is trained with these labels

TABLE 1. Comparison result with many techniques.

Category	[12]	[13]	[14]	[15]	[16]	[17]	[18]	SPP+CNN	CleNet
average	0.654±0.014	0.635±0.011	0.644±0.015	0.656±0.015	0.635±0.011	0.68±0.013	0.672±0.011	0.632±0.013	0.655±0.011
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[10]	[27]	[40]	Ours
Class 18	0.621±0.013	0.652±0.012	0.641±0.015	0.659±0.013	0.629±0.012	0.609±0.012	0.613±0.013	0.623±0.011	0.693±0.008

TABLE 2. Evaluation of computational efficiency for comparative recognition algorithms (top performances highlighted)).

	[12]	[13]	[14]	[15]	[16]	[17]	[18]	SPP-CNN	CleanNet
Tr	24h14m	32h1m	43h16m	34h17m	30h32m	40h9m	33h51m	15h0m	34h1m
Te	2.231s	2.331s	2.145s	1.875s	3.324s	2.127s	2.317s	1.225s	1.768s
	DFB	ML-CRNN	ML-GCN	SSG	MLT	[10]	[27]	[40]	Ours
Tr	31h2m	20h58m	27h18m	40h48m	25h24m	28h44m	32h5m	29h14m	22h11m
Te	1.461s	1.142s	2.531s	1.584s	1.974s	2.242s	2.351s	1.831s	0.553s

and used to automatically assign labels to the remaining unlabeled images. These automatically generated labels were then meticulously verified by the volunteers. Notably, some image-level labels were associated with a very small number of images, posing a challenge for developing an effective classification algorithm. In this study, labels with fewer than 220,000 associated LR aerial images were excluded, resulting in 18 remaining label categories. Almost all (99.973%) of the aerial images across resolutions had fewer than four labels, while a minority had more. Images typically containing small and potentially noisy patches (less than 210×210) were excluded from the dataset. The final step involved organizing the LR and HR aerial images by labels, using half for training our model and the remaining half for evaluation.

This experiment evaluates our classification framework against several leading visual classification models, including deep learning-based aerial image recognition architectures and generic visual classification models. Our method was first benchmarked against seven state-of-the-art deep learning architectures for aerial image recognition [12], [13], [14], [15], [16], [17], [18] known for effectively capturing domain-specific knowledge across different aerial image categories. Publicly available implementations from [12], [13], [16], [17] were directly used in our comparative analysis without modification. For [14], [15], [18], lacking accessible source codes, we developed our implementations aiming to match the performance reported in their original publications as closely as possible.

Additionally, we compared our algorithm against various cutting-edge deep object classification models including Spatial Pyramid Pooling CNN (SPP-CNN) [55], CleanNet [56], Discriminative Filter Bank (DFB) [57], Multi-layer CNN-RNN (ML-CRNN) [36], Multi-label Graph Convolutional Network (ML-GCN) [37], Semantic-specific Graph (SSG) [38], and Multi-label Transformer (MLT) [39]. Given that LR aerial image classification can be regarded as a subset of scene categorization, we also conducted an in-depth comparison with three contemporary scene recognition methods [10], [27], [40].

For the algorithms we implemented, configurations were as follows: For [14], we utilized ResDep-192 [20] as the backbone, modifying it for multi-label classification. Unlike

TABLE 3. Comparative evaluation of average classification accuracies across six feature selection algorithms.

ITFS	CNNR	FSL	PCAR	CNDE	DPFS
0.543	0.587	0.616	0.605	0.614	0.661

the fully connected layer settings, we adapted the rest of the architecture based on ResDep-128 [42], with ResNet-128 [20] serving as the core framework. The learning rate and decay were set to 0.002 and 0.06, respectively, with network loss computed using mean squared error. For [10], the established Object Bank [41] was employed for 18 selected LR aerial image classes, utilizing an average-pooling approach. For this experiment, we evaluated our Hessian-Regularized Feature Selection (HRFS) framework against several prevalent feature selection algorithms in the domain of aerial photography classification, including Information Theory Feature Selection (ITFS) [53], CNN Feature Reduction (CNNFR) [54], Feature Selection for Land Cover Classification (FSLC) [55], PCA Feature Reduction (PCAFR) [56], and CNN-based Dimensionality Reduction (CNNDR) [57]. The comparative analysis focused on average classification accuracies, as summarized in Table 3. Our HRFS method demonstrated superior performance among the competitors. This advantage is attributed to HRFS's ability to effectively leverage the intrinsic geometric distribution of samples within the high-dimensional feature space, where aerial image features are likely situated, hence optimizing the feature selection process for aerial image classification tasks.

2) COMPARATIVE COMPUTATIONAL COST

In the evaluation of visual classification methods, computational efficiency during both training and testing phases is a critical metric of effectiveness. Comparative analysis of time consumption, as detailed in Table 2, reveals that during the training phase, two algorithms outperform our model, attributed to the less complex and more efficient architectures of [33], [37]. However, it's noted that these models underperform by approximately 4.1% in per-class accuracy compared to our approach. Importantly, our proposed algorithm demonstrates superior speed during the testing phase, underscoring the advantage of optimizing for

faster evaluation times, given that training is typically an offline process.

Reflecting on our Low-Resolution (LR) aerial image classification framework, it incorporates three pivotal components: 1) a deep low-rank model for Gaze Shift Path (GSP) generation, 2) the Hessian-Regularized Feature Selection (HRFS), and 3) a kernelized classifier for label prediction. The time invested in each component during the learning phase is reported as: 11 hours and 4 minutes for the GSP generation, 3 hours and 54 minutes for HRFS, and 7 hours and 21 minutes for the kernelized classifier. During the evaluation phase, the respective durations are 203 ms for the GSP generation, 321 ms for HRFS, and 32 ms for the kernelized classifier. It is noteworthy that the bulk of the time is consumed in the first component during training, which can be significantly reduced in practical AI applications through the employment of Nvidia GPUs, potentially achieving a tenfold speed increase via parallel processing techniques.

B. STEP-BY-STEP PERFORMANCE VALIDATION

In this study, we meticulously assess each constituent of our Low-Resolution (LR) aerial image categorization framework to ascertain its overall efficiency. Initially, we explore the effectiveness of our chosen active learning strategy by excluding it and instead opting for a random selection of K image patches (denoted as S11). Alternatively, we simulate human visual preference by selecting central K patches within each aerial image (denoted as S12), given the propensity of human vision to focus on central image areas. Results, indicated in the second column of Table 4, reveal significant performance degradation under both alternate conditions, underscoring the vital role of emulating human visual perception in accurately representing LR aerial imagery.

Further, to validate the importance of maintaining the geometric distribution of samples during the feature selection process, we juxtapose our Hessian-Regularized Feature Selection (HRFS) approach against four renowned feature selection techniques (referenced as S21, S22, S23, S24) prevalent in the field. Notably, these comparative methods do not inherently preserve sample distribution. The outcomes, detailed in Table 4, illustrate that omitting this consideration in feature selection invariably leads to a minimum 4

Lastly, we examine the performance of our HRFS-enhanced classifier in the classification of each LR aerial image. This examination encompasses three scenarios: S31 employs an aggregation-guided multi-layer Convolutional Neural Network (CNN) to accumulate labels from all patches within an aerial image for final image label determination; S32 and S33 respectively substitute our linear kernelized feature representation with polynomial and radial basis function kernels. The variations in classification accuracy, as tabulated in Table 4, reveal that the aggregation approach (S31) markedly diminishes classification efficacy, highlighting the pivotal contribution of our HRFS and kernelized classifier in the categorization process.

TABLE 4. Optimization and decline in performance through module modification.

	S1	S2	S3
O1	-7.215%	-6.112%	-5.166%
O2	-5.442%	-5.565%	-3.213%
O3	Unavailable	-5.324%	-2.874%
O4	Unavailable	-4.230%	Unavailable

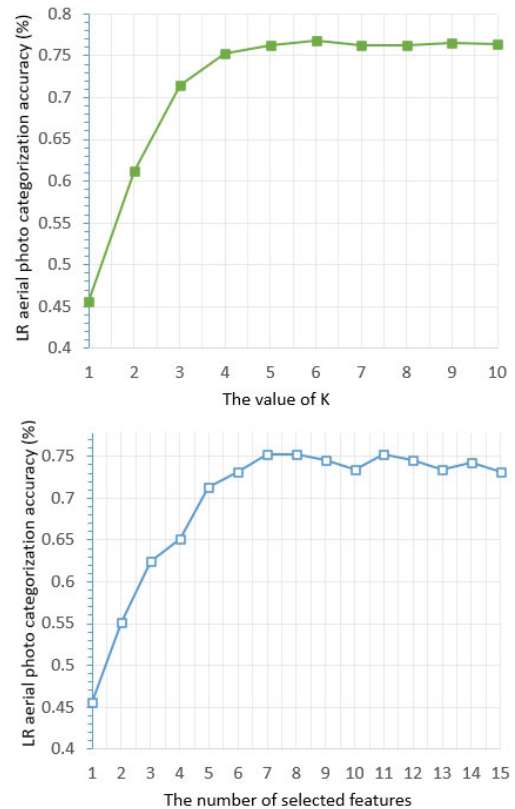


FIGURE 4. Change of classifying accuracy by changing K .

C. CATEGORIZATION BY ADJUSTING PARAMETERS

In our aerial image classification framework, two critical parameters require optimization: the quantity of attractive image patches selected via active learning and the number of features chosen for classification. This part of our study explores how varying these parameters influences the performance of Low-Resolution (LR) aerial image classification.

The first parameter under examination is K , representing the number of actively selected, visually compelling image patches. We incrementally adjust K from one to ten, holding other parameters constant at their default settings. These default values are established through 5-fold cross-validation on a dataset comprising 12,000 aerial images. According to the performance graph shown in Fig. 4, an increase in K initially leads to a rise in classification accuracy, achieving a peak before gradually declining. This trend underscores the significance of carefully calibrating the number of selected image patches to optimize classification effectiveness. In a

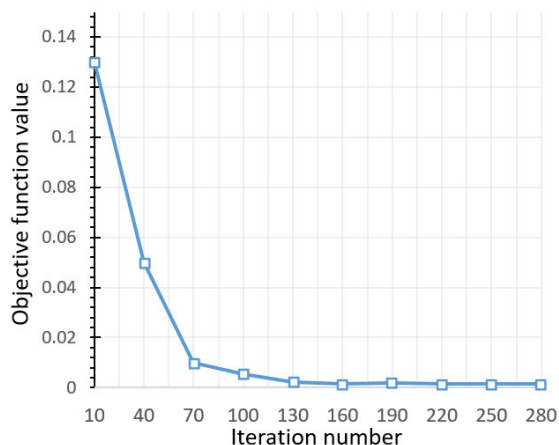


FIGURE 5. Change of objective function value by tuning the number of iterations.

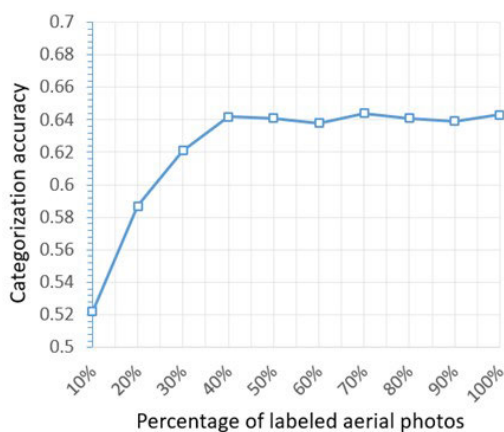


FIGURE 6. Variation in objective function value with iteration number adjustments.

subsequent experiment, we investigate the impact of varying the number of selected features on visual categorization accuracy. Observations, as detailed in the lower portion of Fig. 4, indicate a marked improvement in categorization performance when the number of selected features ranges from one to five. Beyond this range, accuracy levels off, suggesting no significant benefit from selecting additional features. Consequently, for an efficient and effective Low-Resolution (LR) aerial image classification framework, we establish the optimal number of features to be five.

Subsequently, we evaluate the performance of our active learning algorithm by varying the number of iterations. The results, depicted in Fig. 5, demonstrate a consistent and gradual decrease in the objective function value with an increase in iteration numbers. The objective function stabilizes upon reaching 130 iterations, indicating minimal benefit from further iterations. To balance efficiency and efficacy, we thus set the iteration count to 130 for our experiments.

In a final series of tests, we explore the impact of varying the percentage of labeled aerial photographs from 10% to

100%, in increments of 10%. These labeled samples are chosen randomly, and the process is repeated 20 times to derive average categorization accuracies. Results, illustrated in Fig. 6, reveal that our methodology effectively manages aerial photo categorization with a minimum of 40% labeled photographs, implying it can accommodate up to 60% unlabeled Low-Resolution (LR) aerial photos. This capability is deemed highly beneficial for practical applications in LR aerial photo categorization.

V. SUMMARY

The task of categorizing low-resolution (LR) aerial images is pivotal in the development of intelligent systems, drawing significant attention in recent research efforts [22], [23], [24], [25], [26]. This study introduces an innovative framework for LR aerial image recognition, leveraging deep Gaze Shift Path (GSP)-based visual representations, enhanced through the analysis of high-resolution (HR) aerial imagery. Our comprehensive approach includes: 1) deploying an active learning paradigm for the generation of GSPs across various resolutions of aerial imagery, and 2) implementing a novel Hessian-regularized feature selection (HRFS) strategy to isolate the most informative features. The efficacy of our framework is substantiated through a series of rigorous experimental validations.

Constructing an effective categorization system for real-world LR aerial photographs presents a myriad of challenges, which are acknowledged but not fully addressed within the scope of this paper. These challenges include managing the temporal, spatial, and spectral variability of aerial images, adapting to diverse weather conditions, integrating human visual perception insights, and overcoming the limitations in human resources for annotating extensive datasets. To tackle the scarcity of labeled samples, our research introduces a semi-supervised HRFS. Additionally, we incorporate GSPs to embed aspects of human visual perception into the aerial image categorization process. To mitigate the effects of temporal and spectral variances in our experimental dataset, preliminary image processing and selective manual sample curation were employed. It is important to note, however, that this study does not address spectral discrepancies among aerial images directly. Future work will aim to develop a holistic and integrative system for understanding LR aerial photographs, equipped to address the aforementioned challenges through dedicated modules for each specific issue.

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] 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.
- [3] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4576–4584.

- [4] J. Porway, Q. Wang, and S. C. Zhu, "A hierarchical and contextual model for aerial image parsing," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 254–283, Jun. 2010.
- [5] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4095–4104.
- [6] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.
- [7] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, Nov. 2018.
- [8] 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. 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. 11.
- [10] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorization," in *Proc. PRAM*, 2015, pp. 209–224.
- [11] X. Zhu, Z. Ghahramani, and J. D. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proc. ICML*, 2003, pp. 1–7.
- [12] C. Kyrkou and T. Theodoridis, "EmergencyNet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1687–1699, 2020.
- [13] C. Kyrkou and T. Theodoridis, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 517–525.
- [14] 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.*, Sep. 2020, pp. 525–528.
- [15] 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.
- [16] M. Pritt and G. Chern, "Satellite image classification with deep learning," 2020, *arXiv:2010.06497*.
- [17] 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.*, Oct. 2021, pp. 713–720.
- [18] 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.
- [19] L. Wang and S. Chen, " $l_{2,p}$ matrix norm and its application in feature selection," 2013, *arXiv:1303.3987*.
- [20] 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.
- [21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 22–24.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] W. Mu and B. Liu, "Voice activity detection optimized by adaptive attention span transformer," *IEEE Access*, vol. 11, pp. 31238–31243, 2023.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] Z. Xu, I. King, M. R. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1033–1047, Jul. 2010.
- [31] A. T. Dehkordi, M. J. V. Zoej, H. Ghasemi, E. Ghaderpour, and Q. K. Hassan, "A new clustering method to generate training samples for supervised monitoring of long-term water surface dynamics using Landsat data through Google Earth Engine," *Sustainability*, vol. 14, no. 13, p. 8046, Jun. 2022.
- [32] X. Gao, S. Ram, R. C. Philip, J. J. Rodríguez, J. Szep, S. Shao, P. Satam, J. Pacheco, and S. Hariri, "Selecting post-processing schemes for accurate detection of small objects in low-resolution wide-area aerial imagery," *Remote Sens.*, vol. 14, no. 2, p. 255, Jan. 2022.
- [33] 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.
- [34] 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.
- [35] 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.
- [36] A. Caglayan and A. B. Can, "Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition," in *Proc. ECCV Workshops*, Jun. 2018, pp. 43–65.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] Y. Li, M. Dixit, and N. Vasconcelos, "Deep scene image classification with the MFAFVNet," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5757–5765.
- [41] L.-J. Li, H. Su, L. Fei-Fei, and E. Xing, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. NIPS*, 2010, pp. 33–54.
- [42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
- [43] S. Zhou, J. Irvin, Z. Wang, E. Zhang, J. Aljbran, W. Deadrick, R. Rajagopal, and A. Ng, "DeepWind: Weakly supervised localization of wind turbines in satellite imagery," in *Proc. CVPR*, vol. 4, 2009, pp. 23–32.
- [44] 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.
- [45] C. Shi, Q. Ruan, G. An, and R. Zhao, "Hessian semi-supervised sparse feature selection based on $L_{2,1/2}$ -matrix norm," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 16–28, Jan. 2015.
- [46] 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.
- [47] 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.

- [48] 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.
- [49] G. H. Golub and C. F. van Loan, *Matrix Computation*. Baltimore, MD, USA: Johns Hopkins University, 1996.
- [50] Ö. Akar, "Mapping land use with using rotation forest algorithm from UAV images," *Geocarto Int.*, vol. 33, no. 5, pp. 538–553, Jan. 2017.
- [51] 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.
- [52] D. Costea and M. Leordeanu, "Aerial image geolocation from recognition and matching of roads and intersections," 2016, *arXiv:1605.08323*.
- [53] M. P. Uddin, M. A. Mamun, M. I. Afjal, and M. A. Hossain, "Information-theoretic feature selection with segmentation-based folded principal component analysis (PCA) for hyperspectral image classification," *Int. J. Remote Sens.*, vol. 42, no. 1, pp. 286–321, Jan. 2021.
- [54] F. Özyurt, E. Avci, and E. Sert, "UC-merced image classification with CNN feature reduction using wavelet entropy optimized with genetic algorithm," *Traitement Signal*, vol. 37, no. 3, pp. 347–353, Jun. 2020.
- [55] O. Stromann, A. Nascetti, O. Yousif, and Y. Ban, "Dimensionality reduction and feature selection for object-based land cover classification based on Sentinel-1 and Sentinel-2 time series using Google Earth Engine," *Remote Sens.*, vol. 12, no. 1, p. 76, Dec. 2019.
- [56] M. P. Uddin, M. A. Mamun, and M. A. Hossain, "PCA-based feature reduction for hyperspectral remote sensing image classification," *IETE Tech. Rev.*, vol. 38, no. 4, pp. 377–396, Jul. 2021.
- [57] M. Ramamurthy, Y. H. Robinson, S. Vimal, and A. Suresh, "RETRACTED: Auto encoder based dimensionality reduction and classification using convolutional neural networks for hyperspectral images," *Microprocessors Microsyst.*, vol. 79, Nov. 2020, Art. no. 103280.
- [58] C. Zhang, H. Li, C. Chen, Y. Qian, and X. Zhou, "Enhanced group sparse regularized nonconvex regression for face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2438–2452, May 2022.
- [59] C. Zhang, H. Li, W. Lv, Z. Huang, Y. Gao, and C. Chen, "Enhanced tensor low-rank and sparse representation recovery for incomplete multi-view clustering," in *Proc. AAAI*, 2023, pp. 234–253.
- [60] D. Ming and C. Ding, "Robust flexible feature selection via exclusive L21 regularization," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 1182–1189.
- [61] 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.
- [62] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Proc. NIPS*, 2017, pp. 1–10.

GUIFENG WANG is a Faculty with the Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua, China. His research topics are computer vision, image processing, and artificial intelligence.

JIANZHANG XIAO is an Associate Professor with Jinhua Polytechnic, Jinhua, China. His research topics are multimedia and computer vision.

YI YANG is currently a Faculty Member of the College of Computer Sciences, Beijing Technology and Business University. His research interest includes artificial intelligence.

...