LR Aerial Imagery Categorization by Transferring Cross-Resolution Perceptual Experiences

Yue Yu and Yi Li

***Abstract*—Hundreds of satellites orbiting at various altitudes capture an extensive array of aerial photographs daily. High-altitude satellites typically acquire low-resolution (LR) images that cover vast areas, whereas their low-altitude counterparts obtain high-resolution (HR) images detailing much smaller regions. The accurate interpretation of LR aerial imagery is crucial in the field of computer vision, yet it presents significant challenges, including the complexity of emulating human hierarchical visual perception and the daunting task of annotating enough data for effective training. To address these challenges, we introduce a cross-resolution perceptual knowledge propagation (CPKP) framework, which aims to leverage the visual perceptual insights gained from HR aerial imagery to enhance the categorization of LR images. This approach involves a novel low-rank model that segments each LR aerial photo into distinct visually and semantically significant foreground regions, alongside less pertinent background areas. This model is capable of generating a gaze shifting path (GSP) that reflects human gaze patterns and formulating a deep feature for each GSP. Subsequently, a kernel-induced feature selection algorithm is deployed to extract a concise yet powerful set of deep GSP features that are effective across both LR and HR aerial images. Utilizing these features, a linear classifier is collaboratively trained using labels from both LR and HR images, facilitating the categorization of LR images. Notably, the CPKP framework enhances the efficiency of training the linear classifier, given that HR photo labels are more readily available. Our comprehensive visualizations and comparative analysis underscore the effectiveness and superiority of this innovative approach.**

***Index Terms*—Aerial photo, cross resolution, gaze shifting, human perception, knowledge propagation.**

I. Introduction

SINCE the launch of the first earth observation satellite in October 1957, there has been significant advancement in carrier rocket technology, remote sensing, and satellite communication, resulting in the deployment of hundreds of satellites into orbit. Based on their orbital altitudes, these satellites can be classified into high-altitude (over 2000 km) and low-altitude (200–2000 km) categories. High-altitude satellites, in contrast to their low-altitude counterparts, span broader areas and have longer orbital periods, resulting in the captured aerial photos having lower resolutions than those taken from lower altitudes. The ability to interpret the semantics of these low-resolution (LR) aerial images plays a critical role in various computer vision applications. For instance, periodic analysis of LR aerial images can monitor the spatial distribution of wildlife, forests, and wetlands, offering valuable insights into biodiversity and aiding in the conservation of endangered species by managing their habitats effectively. Moreover, in the realm of autonomous logistics, understanding the semantic categories within each LR aerial image is crucial for optimizing the routing of long-haul driverless trucks by enabling the rapid and dynamic calculation of the most efficient paths between locations. In addition, semantic parsing of LR aerial images yields an understanding of the functional zones and topological layouts of urban areas, significantly benefiting urban planning tasks, such as 3-D architectural modeling and land-use strategy development.

In the field of computer vision, a considerable number of shallow and deep learning models for visual categorization and annotation have been developed to process aerial imagery with medium-to-high spatial resolutions (spatial resolution $\leq$ 10 m). Noteworthy contributions in this area include: 1) object localization in aerial photos employing multiple instance learning and convolutional neural networks (CNNs) with weakly labeled data [55], [56]; 2) the use of graphical models for semantic information propagation to parse aerial images [6], [8]; and 3) the development of sophisticated deep learning architectures specifically tailored for the semantic annotation of aerial imagery [9], [10], [11]. These methodologies have been validated through rigorous experimentation and have found practical application in commercial systems, demonstrating their effectiveness, user-friendliness, and scalability. Nevertheless, it has been observed that existing models face challenges in accurately representing LR aerial photographs, attributable to three primary limitations.

1) 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 nontrivial. Potential challenges include: a) determining the sequence of humans observing the extracted salient regions (e.g., the path displayed in Fig. 1); b) refining the contaminated labels of the training LR aerial photos; and c) transferring

Manuscript received 7 June 2023; revised 10 November 2023 and 1 December 2023; accepted 9 January 2024. Date of publication 23 February 2024; date of current version 18 March 2024. This work was supported by the intelligent precision manufacturing virtual simulation training base at Jinhua Polytechnic. *(Corresponding author: Yue Yu.)*

Yue Yu is with the Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321016, China (e-mail: yueyu.1778@gmail.com).

Yi Li is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325035, China.

Digital Object Identifier 10.1109/JSTARS.2024.3369413

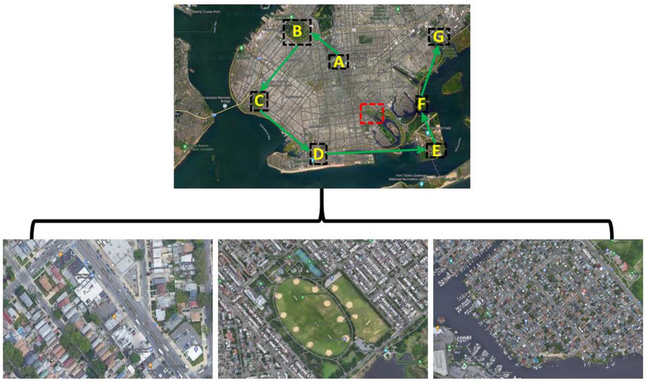


Fig. 1. Top: the visually/semantically salient regions sequentially observed by humans in an LR aerial photo (marked by path $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow F \rightarrow G$) and the blurred playground (marked by red dashed box). Bottom: three HR aerial photos capture subregions of the LR aerial photo (the middle one details the blurred playground inside the LR aerial photo).

image-level semantic labels into multiple regions inside an LR aerial photo.

2) Compared with HR aerial photos, LR ones are practically with an inferior image quality, as they are more sensitive to a variety of uncontrollable factors, e.g., the varying weather/lighting conditions and possibly communication interference. This makes the number of labeled LR aerial photos noticeably smaller than that of labeled HR ones. Thus, training an image model solely based on LR images is prone to overfitting. In practice, the HR and LR aerial photos are captured asynchronistically, based on which they characterize each area complementarily. As exemplified in Fig. 1, a playground is blurred in an LR aerial photo but clear in the HR one. To leverage HR aerial photos for image model upgradation, we have to discover the joint discriminative features shared between HR and LR aerial photos. However, designing a discriminative model that flexibly propagates human perceptual knowledge across multiple resolutions remains unsolved.

3) Toward an efficient and interpretable image model for semantic understanding, we want a set of highly discriminative and low redundant features shared between HR and LR aerial photos. However, instead of the original feature space, the shared discriminative features may be distributed in the high-order feature space, which may be unexpectedly high dimensional. This makes the conventional feature selection (FS) toward the high-order feature space computationally intractable. Besides, visualizing high-dimensional features on the high-order feature space remains unsolved due to the implicit feature mapping from the original feature space to the high-order one. Moreover, discriminative FS and feature classification are generally conducted separately. In theory, there is no guarantee that the selected features can maximize the classification performance. Ideally, we want a unified framework that jointly optimizes FS and feature classification.

To address the challenges highlighted, we introduce a cross-resolution perceptual knowledge propagation (CPKP)

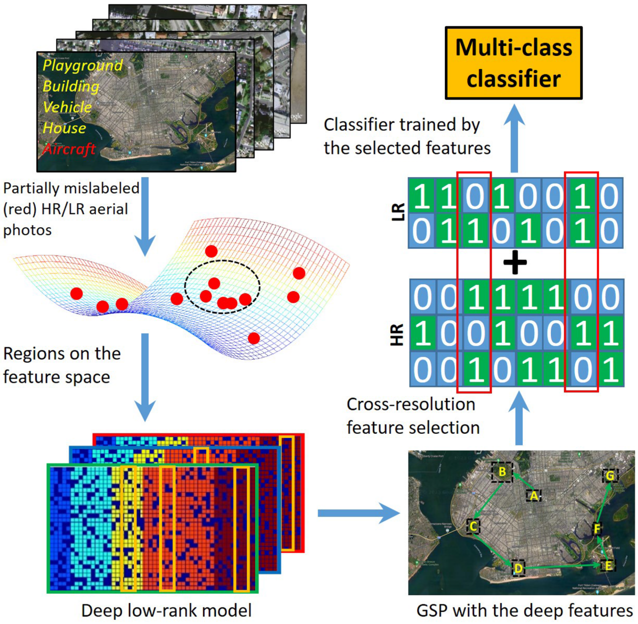

Playground
Building
Vehicle
House
Aircraft
Multi-class classifier
Classifier trained by the selected features
Partially mislabeled (red) HR/LR aerial photos
LR
1 1 0 1 0 0 1 0
0 1 1 0 1 0 1 0
+
HR
0 0 1 1 1 1 0 0
1 0 0 1 0 0 1 1
0 0 1 0 1 1 0 1
Regions on the feature space
Cross-resolution feature selection
B
A
G
C
F
D
E
Deep low-rank model
GSP with the deep features

Fig. 2. Pipeline of our LR aerial photo understanding framework based on CPKP.

framework that leverages perceptual insights derived from high-resolution (HR) aerial imagery to improve the categorization of LR ones. An illustrative summary of our approach to understanding LR aerial photos is presented in Fig. 2. Starting with a substantial collection of HR and LR aerial images, including some that are unlabeled, we initially map the internal regions of these images onto a feature space that synergistically utilizes visual and semantic information. Following this, a deep low-rank model is employed to segment each LR aerial photo into sequences of visually and semantically significant foreground regions, or gaze shifting paths (GSPs), alongside less significant background areas, while concurrently computing the deep features for each GSP. Our goal is to identify a streamlined set of features that are discriminative across both HR and LR aerial imagery. To this end, we project the deep GSP features into a higher order kernel-induced feature space. In order to transfer the perceptual knowledge gleaned from HR aerial images to LR ones, we develop an FS algorithm that aims to: 1) minimize the distribution discrepancies between LR and HR aerial photos, both marginally and conditionally, and 2) optimize linear classification accuracy. The selected features are then used to train a classifier using both labeled HR and LR images, thereby addressing potential issues of sample inadequacy and preventing overfitting in LR aerial photo categorization. Our method is rigorously compared with 17 contemporary models in generic and aerial image categorization, showcasing its superiority. In addition, visualizations of the discriminative features common to both HR and LR aerial images offer further insights, facilitating discussions on the efficacy and innovation of our CPKP framework.

This article presents several significant advancements in the field of aerial photo categorization, highlighted by the following contributions: 1) the development of a deep low-rank model capable of identifying GSPs within LR aerial imagery and simultaneously deriving deep GSP features; 2) the introduction of a CPKP framework that efficiently identifies and selects highly discriminative yet minimally redundant features common to both HR and LR aerial photos. This approach effectively addresses and mitigates the issue of model overfitting; and 3) the implementation of a kernel-induced feature space mapping technique. This method projects the extracted deep GSP features into a higher order feature space, facilitating the simultaneous execution of FS and classifier training, thereby enhancing the categorization process.

II. Related Work

A. *Semantic Analysis of Aerial Imagery*

A variety of semantic models have been designed to interpret aerial imagery, offering insights into both image- and region-level characteristics. A detailed review of these deep learning approaches to aerial photo understanding is presented in [61].

1) Image-Level Semantic Modeling: Notable efforts in this domain include the work by Zhang et al. [57], who introduced a topological feature capturing inter-region connections within aerial images, utilizing a kernel-induced vector for image categorization. Xia et al. [59] explored a weakly supervised approach for semantically labeling HR aerial images. Akar [60] merged rotation forest techniques with object-level feature extraction for multicategory classification of aerial photos. Sameen et al. [62] constructed a hierarchical CNN to identify multiple labels of HR aerial images, particularly those depicting urban areas. Cheng et al. [58] applied a pretrained deep CNN, fine-tuned with a domain-specific dataset, for classifying HR aerial imagery. A cross-modality learning framework by Hong et al. [43] collaboratively utilized five deep models, integrating pixel- and spatial-level features for aerial image categorization. Cai and Wei [12] introduced a cross-attention mechanism for feature weight learning, while Bazi et al. [64] employed a vision transformer for capturing long-term contextual relationships among image regions.

2) Region-Level Semantic Modeling: Wang et al. [4] developed an end-to-end network for identifying multiscale salient objects within aerial photos. A focal-loss-based model by Yang et al. [1] efficiently locates vehicles, and Costea and Leordeanu [63] devised a geolocalization approach through the extraction of urban features like intersections and streets. Yu et al. [19] combined feature enhancement with soft label assignments for anchor-independent object detection in aerial images. Wang et al. [20] proposed a rotation-invariant detector for estimating object angles, and Chalavadi et al. [54] introduced mSODANet, a parallel deep model learning contextual features from objects across multiple scales and fields of view.

Distinctively, our approach draws inspiration from biotic mechanisms, closely replicating human gaze behavior for enhanced semantic interpretation of aerial photos. Additional innovative methodologies include a self-guided separating network for remote sensing imagery analysis by Wang et al. [35], addressing feature representation inconsistencies and background–target imbalance. A large kernel sparse deep model by Wang et al. [36] focuses on capturing extensive receptive fields, and a spatial–logical aggregation network by Zhang et al. [37] aims to reveal fine-grained morphological structures in hyperspectral images.

B. *Supervised FS Techniques*

Supervised FS methods evaluate the importance of features based on their correlation with target labels. Nie et al. [16] developed a robust FS algorithm leveraging $l_{1,2}$-norm regularization to optimize FS criteria. An innovative and scalable FS framework suited for high-dimensional data was introduced by Gui et al. [17]. They crafted an attention-driven algorithm for scoring features under supervision, employing a smooth hinge loss to encourage sparsity and select discriminative features. The $l_{1,3}$-norm combined with an exclusive lasso approach for FS was explored in [18], effectively filtering out redundant and irrelevant features. An approach prioritizing the maximization of between-class variance to within-class variance, termed the worst case, was introduced as a criterion for discerning key features. Ahadzadeh et al. [15] presented a two-phase FS strategy using particle swarm optimization for high-dimensional datasets, initially removing low-quality features globally before finely identifying the most discriminative ones locally. These conventional FS methods operate in the original feature space, which may not adequately represent samples distributed in a higher order kernel space. Addressing this, Song et al. [30] proposed a kernel-based FS method utilizing the Hilbert–Schmidt independence criterion (HSIC) to enhance feature-label correlation. Masaeli et al. [31] developed an HSIC-based implicit FS technique using an l_1/l_∞-norm regularizer for feature transformation. The HSIC-LASSO by Yamada et al. [32] employs the dual augmented Lagrangian for global optimization. Chen et al. [33] introduced a kernel-induced feature selector focusing on identifying the most discriminative subset of features. Chen et al. [33] advanced multikernel FS by attributing an indefinite base kernel to each feature and deriving sparse kernel combination coefficients through an l_1-norm. Unlike these approaches, our CPKP-based feature selector uniquely enables: 1) the explicit and discriminative selection of deep GSP features within the kernel space; 2) the identification of features that distinctly categorize LR/HR aerial images across various classes; and 3) the utilization of cross-resolution perceptual insights to refine the FS process.

III. Our Proposed Method

A. *Deep Low-Rank Algorithm for GSP Learning*

Within each LR aerial photograph, a myriad of fine-grained objects can be identified. Biological research [2] suggests that human vision tends to focus on a limited subset of objects that are visually or semantically salient during the perception process. Specifically, when viewing an LR aerial photo, the human gaze is initially drawn to prominent ground features, while less conspicuous background areas remain largely unexamined. This pattern of human visual attention provides valuable insights for classifying LR aerial imagery. Consequently, we introduce a deep low-rank model designed to iteratively identify and select salient image patches, thereby forming GSPs, while concurrently deriving their deep features.

Human visual perception theory further highlights a significant self-representativeness or correlation among nonsalient background patches within each scene, in stark contrast to the salient foreground patches that exhibit minimal correlation. Inspired by this distinction, our approach involves segmenting the feature matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$ of each LR aerial photo into components representing salient and nonsalient regions. This methodology enables a more nuanced analysis of aerial imagery, leveraging the inherent contrast between foreground interest points and the background to enhance photo categorization

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

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 corresponding to the nonsalient background image patches (the other columns are all zeros). $\mathbf{E} \in \mathbb{R}^{T \times N}$ represents feature columns corresponding to the salient image patches (the other columns are all zeros).

Aiming at a unique solution indicating the salient image patches, some criteria are proposed to constrain $\mathbf{Y}$ and $\mathbf{E}$. In our work, two observations are made. First, only a small fraction of image patches within each LR aerial photo are salient and will be processed by the human vision system in detail. This mathematically reflects that $\mathbf{E}$ is a sparse matrix. Second, the high correlation of the nonsalient 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)

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

where $|| \cdot ||_*$ is the matrix nuclear norm representing a convex approximation to matrix rank function, $l_1(\mathbf{E})$ quantizes the sparsity of $\mathbf{E}$, $f(\Upsilon, \mathbf{X}))$ selects nonsalient background image patches from each LR aerial photo, and $l_2(\mathbf{Y}, f(\Upsilon, \mathbf{X}))$ penalizes the loss of nonsalient background image patch selection. $\Omega(\Upsilon)$ serves as a regularizer. α, β, and γ are nonnegative parameters balancing the tradeoff among the corresponding terms. More concretely, to ensure a highly sparse $\mathbf{E}$, $l_1(\cdot)$ is defined as

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

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

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

To precisely select the nonsalient background image patches inside each LR aerial photo, we propose to learn a deep semantic model $f(\Upsilon, \mathbf{X})$. It includes L layers of linear/nonlinear transformations. The deep representation from the top layer is denoted by $h(\mathbf{X}_i)$ and $\mathbf{X}_i$ is the T-dimensional column feature vector from the ith image patch. Meanwhile, the current layer's output is utilized as the input of the next layer. Mathematically, this can be represented as

$$h(\mathbf{X}_i) = g_L(\mathbf{X}_i) \tag{5}$$

$$g_l(\mathbf{X}_i) = \phi(\mathbf{Z}_l h_{l-1}(\mathbf{X}_i + \xi_l), \;\; l = 1, \ldots, L \tag{6}$$

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

$$g_1(\mathbf{X}_i) = \phi(\mathbf{Z}_1 \mathbf{X}_i + \xi_1), \;\; l = 1, \ldots, L. \tag{7}$$

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

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

where parameter set $\Upsilon = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_L, \mathbf{Z}, \xi_1, \ldots, \xi_L\}$.

To mitigate the overfitting problem, we design a regularization term to penalize model complexity. Herein, the regularization function $\Omega(\Upsilon)$ is given as

$$\Omega(\Upsilon) = \frac{1}{2}\left(||\mathbf{Z}||_F^2 + \sum_{i=1}^{L}(||\mathbf{Z}_l||_F^2 + ||\xi||_2^2)\right). \tag{9}$$

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

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

We observe that (10) is a nonconvex optimization over the entire variables. In our implementation, we follow the iterative algorithm in [3] to solve it. Thereafter, denoting $\mathbf{Y}^*$ as the optimal solution of (10), the saliency score of the ith image patch in an LR aerial photo is calculated by

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

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

B. CPKP for FS

The deep GSP features derived from LR aerial imagery typically reside within a complex high-dimensional space. Given the limited availability of labeled LR aerial photos, this scenario poses a challenge known as the "curse of dimensionality," which can adversely affect the categorization of LR aerial images. To address this issue, we introduce a CPKP framework designed to identify and select a concise yet highly discriminative set of features common to both HR and LR aerial images. This process enables the use of selected features from both HR and LR images to collaboratively train the categorization model, effectively bridging the gap between different resolutions. Essentially, the CPKP approach serves to not only diminish the feature space dimensionality but also to augment the pool of training samples, thereby significantly alleviating the challenges posed by high dimensionality.

1) Feature Mapping by Approximating Polynomial Kernel: The polynomial kernel can be mathematically represented as

$$\varphi(\mathbf{u}, \mathbf{v}) = (\tau \mathbf{u}^T \mathbf{v} + \kappa)^Q \tag{12}$$

where Q denotes the degree. Such kernel is comprised of features whose monomial's degree is smaller than Q. This can be further represented as

$$\varphi_{q,\mathbf{e}}(\mathbf{u}) = \sqrt{C_Q^q \cdot \kappa^{Q-q}} \prod\nolimits_{j=1}^{q} \mathbf{u}_{e_j}, i = 0, \ldots, Q \tag{13}$$

where $\mathbf{e} \in \{1, \ldots, TK\}^q$ enumerates the entire selections of q-dimensional coordinates in $\mathbf{u}$, and TK is the dimensionality of deep GSP feature. By leveraging the multinomial theorem, (13) can be reorganized into

$$\varphi(\mathbf{u}) = \cup_{q=1}^{Q} \{\varphi_{q, \mathbf{e} \in \{1, \ldots, TK\}^q}(\mathbf{u})\}. \tag{14}$$

This reflects that for degree Q, there are a total of $S = C_{TK+Q}^{Q}$ candidate features for FS, where operator C_i^j counts the combinations of selecting j features from i features. Noticeably, our FS also supports other kernels like linear kernel and radial basis function (RBF).

2) Objective Function of FS: By leveraging the above explicit feature map, deep GSP features engineered from HR and LR aerial photos are represented by $\{(\varphi(\mathbf{u}_i) \in \mathbb{R}^S), r_i^H\}_{i=1}^{M^H}$ and $\{(\varphi(\mathbf{u}_i) \in \mathbb{R}^S), r_i^L\}_{i=1}^{M^L}$, respectively, where M^H and M^L count the HR and LR aerial photos, respectively. r_i^H and r_i^L denote the category labels of the HR and LR aerial photos, respectively. It is worth emphasizing that, for some LR aerial photos, these labels might be absent. As shown in (14), the number of explicit features increases exponentially with Q and TK. Herein, a novel FS algorithm is proposed to select features discriminative to both HR and LR aerial photos.

Without loss of generality, we assume that all the HR aerial photos are labeled, while the LR ones are unlabeled. We denote the HR aerial photos as $\{\mathbf{u}_i^H, r_i^H \subseteq \{1, \ldots, B\}\}_{i=1}^{M^H}$, where $\mathbf{u}_i^H$ denotes the TK-dimensional deep GSP feature and r_i^H the corresponding category labels. We denote $\mathbf{U}^H = \{\mathbf{u}_i\}_{i=1}^{M^H}$ as deep GSP features from the entire HR aerial photos. Similarly, we denote the unlabeled target data as $\mathbf{U}^L = \{\mathbf{u}_i^L\}_{i=1}^{M^L}$, where $\mathbf{U}^L$ are deep GSP features from the LR aerial photos. Let $p^H(\mathbf{U}^H)$ and $p^H(\mathbf{U}^H)$ be the marginal distributions of $\mathbf{U}^H$ and $\mathbf{U}^L$, respectively, and $q^H(\mathbf{U}^H)$ and $q^H(\mathbf{U}^H)$ be the conditional distributions of $\mathbf{U}^H$ and $\mathbf{U}^L$, respectively. The objective of our FS is to select an optimal feature set that predicts labels $\{r_i^L\}_{i=1}^{M^L}$ using the input LR aerial photos $\{\mathbf{u}_i^L\}_{i=1}^{M^L}$ under assumptions $p^H(\mathbf{u}^H) \neq p^H(\mathbf{u}^H)$ and $q^H(r^H|\mathbf{u}^H) \neq q^H(r^L|\mathbf{u}^L)$.

It is reasonable to assume that there exists a binary indicator $\mathbf{s} \in \{0,1\}^S$, such that $p(\varphi(\mathbf{u}^H) \odot \mathbf{s}) \approx p(\varphi(\mathbf{u}^L) \odot \mathbf{s})$ and $p(\mathbf{r}^H|\varphi(\mathbf{u}^H) \odot \mathbf{s}) \approx p(\varphi(\mathbf{r}^L|\mathbf{u}^L) \odot \mathbf{s})$. Our target is to learn the indicator $\mathbf{s}$. Since we practically have insufficient LR aerial photos, $\mathbf{s}$ cannot be effectively learned due to the overfitting problem. In this way, we propose to learn binary indicator $\mathbf{s}$ and a linear classifier $\mathbf{H}$ jointly, in order to satisfy the following three criteria: 1) the distance between the marginal distributions $p(\varphi(\mathbf{u}^H) \odot \mathbf{s})$ and $p(\varphi(\mathbf{u}^L) \odot \mathbf{s})$ is sufficiently small; 2) $\varphi(\mathbf{u}^H) \odot \mathbf{s}$ and $\varphi(\mathbf{u}^L) \odot \mathbf{s}$ preserve the discriminative features of deep GSP features $\varphi(\mathbf{U}^H)$ and $\varphi(\mathbf{U}^L)$, based on which $p(\mathbf{r}^H|\varphi(\mathbf{u}^H) \odot \mathbf{s}) \approx p(\varphi(\mathbf{r}^L|\mathbf{u}^L) \odot \mathbf{s})$; and 3) the learned classifier $\mathcal{C}(\mathbf{u}^L) = (\varphi(\mathbf{u}^L) \odot \mathbf{s})\mathbf{H}$ can optimally categorize the training LR aerial photos $\varphi(\mathbf{u}^L)$. These criteria can be mathematically represented in the following subsections.

a) Marginal distribution discrepancy minimization: Given the polynomial-kernel-based feature mapping $\varphi(\mathbf{u})$ induced by (14), we aim to minimize the marginal distribution discrepancy by FS. This can be formulated as

$$\begin{aligned} &\min_{\mathbf{s} \in \mathbf{S}} \eta_1(\mathbf{s}) \\ &= \left\| \frac{1}{M^H} \sum_{\mathbf{u}^H \in \mathbf{U}^H} \varphi(\mathbf{u}^H) \odot \mathbf{s} - \frac{1}{M^L} \sum_{\mathbf{u}^L \in \mathbf{U}^L} \varphi(\mathbf{u}^L) \odot \mathbf{s} \right\|_F^2 \end{aligned} \tag{15}$$

where $|| \cdot ||_F^2$ denotes the squared Frobenius norm, the binary indicator's domain is represented by $\mathbf{S} = \{\mathbf{s}|\mathbf{s} \in \{0,1\}^S, ||\mathbf{s}||_0 \leq A\}$, and A is the maximum number of selected features. Apparently, only minimizing the marginal distribution discrepancy cannot maximally reduce the conditional distribution discrepancy. Herein, the conditional distribution discrepancy minimization is formulated as follows.

b) Conditional distribution discrepancy minimization: We notice that $q^L(r^L|\mathbf{u}^L)$ of the LR aerial photos is unknown. Thus, it is infeasible to directly compare the conditional distributions. Researchers proposed to compare pairwise conditional distributions by predicting their kernel densities. Nevertheless, this method requires the prespecified labels of the LR aerial photos, which are usually unavailable in practice.

Motivated by Pan et al. [5], a linear classifier $\mathcal{C}(\mathbf{u}) = (\varphi(\mathbf{u}) \odot \mathbf{s})\mathbf{H}$ is learned from the labeled HR aerial photos to calculate the category labels of the LR aerial photos. Practically, the posterior probabilities $q^H(r^H|\mathbf{u}^H)$ and $q^H(r^L|\mathbf{u}^L)$ have complicated forms. Instead, we utilize the class-conditional distributions $q^H(\mathbf{u}^H|r^H = b)$ and $q^L(\mathbf{u}^L|r^L = b)$. More specifically, we first calculate the conditional distribution distance between HR and LR aerial photos labeled by $b \subseteq \{1, \ldots, B\}$. Thereafter, we attempt to minimize the conditional distribution discrepancy, i.e.,

$$\begin{aligned} &\min_{\mathbf{s} \in \mathbf{S}} \eta_2(\mathbf{s}) \\ &= \left\| \frac{1}{M_b^H} \sum_{\mathbf{u}^H \in \mathbf{U}_b^H} \varphi(\mathbf{u}^H) \odot \mathbf{s} - \frac{1}{M_b^L} \sum_{\mathbf{u}^L \in \mathbf{U}_b^L} \varphi(\mathbf{u}^L) \odot \mathbf{s} \right\|_F^2 \end{aligned} \tag{16}$$

where $\mathbf{U}_b^H$ denotes the HR aerial photos with category label b and M_b^H counts their number. $\mathbf{U}_b^L$ represents the LR aerial photos with pseudo label b and M_b^L denotes their number.

c) Empirical error minimization: As we mentioned, we expect that the selected features not only minimize the distribution difference but also are sufficiently discriminative for visual categorization. Toward a succinct set of discriminative features, the third criterion is to minimize the empirical error. In our implementation, One-vs-All coding of error-correcting output codes (ECOCs) [5] is employed. Mathematically, if $r_i^H = b$, then the ECOC is a vector, where the bth entry is one while others are zero. Since we hypothesize that LR aerial photos are unlabeled, the empirical error of the HR ones will be minimized, i.e.,

$$\min_{\mathbf{s}\in\mathbf{S}} \min_{\mathbf{H}} \eta_3(\mathbf{s},\mathbf{H}) = \sum_{\mathbf{u}_i\in\mathbf{U}^H} \frac{1}{2}||\epsilon_i||_F^2 + \frac{\psi}{2}||\mathbf{H}||_F^2$$
$$\text{s.t. } \epsilon_i = (\varphi(\mathbf{u}_i)\odot\mathbf{s})\mathbf{H} - \mathbf{r}_i,\ i = 1,\ldots,M^H. \tag{17}$$

In our work, the pseudolabels of LR aerial photos are calculated by

$$r_*^L = \arg\max_{b\in\{1,\ldots,B\}} \mathcal{C}_b(\mathbf{u}^L) \tag{18}$$

where $\mathcal{C}_b$ denotes the bth entity of $\mathcal{C}_b^L(\mathbf{u}) = (\varphi(\mathbf{u}^L)\odot\mathbf{s})\mathbf{H}$. Practically, although many pseudolabels may be mistaken, it is suitable to compare the conditional distributions. This is because, different from density estimation, the conditional distributions are compared by exploiting sample statistics. In this way, the classifier $\mathcal{C}$ can be updated progressively along with the designed iterative optimization.

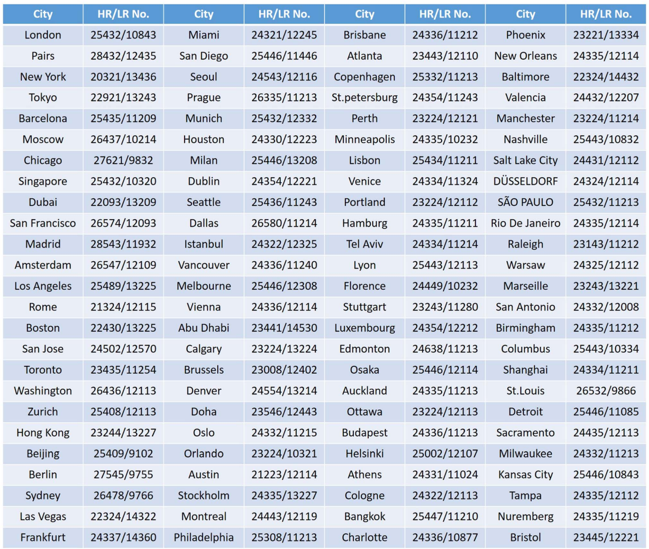
By combining the above criteria, the final objective function is given as

$$\min_{\mathbf{s}\in\mathbf{S}} \min_{\mathbf{H}} \eta(\mathbf{s},\mathbf{H}) = \eta_1(\mathbf{s}) + \eta_2(\mathbf{s}) + \eta_3(\mathbf{s},\mathbf{H})$$
$$\text{s.t. } \epsilon_i = (\varphi(\mathbf{u}_i)\odot\mathbf{s})\mathbf{H} - \mathbf{r}_i,\ i = 1,\ldots,M^H. \tag{19}$$

This objective function is NP-hard due to the combinatorial integral constraints on $\mathbf{s}$. Herein, an efficient solution is elaborated online.[1]

IV. Experimental Evaluation and Insights

Our study assesses the performance of LR aerial photo categorization through a series of four detailed experiments. Initially, we introduce a meticulously curated dataset comprising over 3.7 million LR and HR aerial images sourced from the top 100 metropolitan areas across various continents. Utilizing this extensive image set, we benchmark our proposed method against 17 leading deep categorization models, examining aspects such as accuracy, stability, and computational efficiency. This comparative analysis sheds light on the superior performance and competitive edge of our approach. Subsequently, we fine-tune our model by systematically adjusting its intrinsic parameters, aiming to identify the optimal configuration for enhanced categorization performance. Furthermore, an ablation study is conducted to dissect and evaluate the contribution of each critical component within our CPKP-based categorization framework. In parallel, we present visualizations of the visually and semantically significant regions identified through our CPKP-based FS, offering additional insights into the model's operational dynamics.



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

Fig. 3. Number of HR/LR aerial photos crawled from the 100 metropolitan cities selected by us.

All experiments were executed on a high-performance computing cluster equipped with 16 Intel i9-12900K CPUs, 512 GB of RAM, and four Nvidia GeForce RTX 3090Ti graphics cards. The baseline models for comparison were implemented using C++ and Python, ensuring a rigorous and fair evaluation of our categorization methodology.

A. *Dataset Description*

To rigorously assess our aerial photo categorization model, we embarked on creating a comprehensive dataset, recognizing the absence of a suitable large-scale collection of LR and HR aerial images across diverse categories in existing research. This endeavor required substantial effort, culminating in the assembly of a dataset exceeding 3.6 million LR and HR aerial images. These images were sourced from Google, Apple, and Bing Maps, with a dedicated crawler software developed for this purpose, which operated for a total of 4310 h to search for and download the necessary imagery. Specifically, we utilized the names of the 100 most renowned metropolitan areas worldwide as search keywords on these mapping services, as illustrated in Fig. 3. The dataset encompasses 46 cities from North America, 38 from Europe, ten from Asia, four from Oceania, and two from South America. The LR and HR aerial images extracted from the cached maps vary significantly in resolution, with HR images ranging between 5K × 5K and $22K \times 22K$ pixels. To

[1] https://docs.google.com/document/d/1JtCVkH3vXc8KgRf1JID8-mG60-91jj7O/edit?usp=sharing&ouid=101578137679720572579&rtpof=true&sd=true

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

maintain consistency, we capped the resolution of HR aerial photos at a maximum of 22K × 22K pixels. Conversely, the LR images feature resolutions from 0.35K × 0.35K to 2K × 2K pixels. This specific range of resolutions was chosen based on several criteria: 1) to ensure that each HR aerial photo could be broadly associated with up to four distinct categories; 2) to limit the overlap between any two LR/HR aerial photos to a maximum of 5%; and 3) to avoid the issue where too few pixels in an LR aerial image would render it virtually indecipherable. This meticulous dataset compilation allows us to conduct a thorough and nuanced evaluation of our categorization model, bridging a significant gap in the field of aerial imagery analysis.

In the process of assembling our dataset, we encountered instances where some LR and HR aerial images appeared blurred, primarily due to adverse weather conditions or the photography of sensitive military zones, as illustrated in Fig. 4. Our approach is geared toward identifying object patches across varying scales to extract deep perceptual features for visual categorization. However, poor visibility in LR/HR aerial photos caused by inclement weather can compromise the integrity of categorization accuracy comparisons. To address this, we opted to exclude LR/HR aerial images if more than 20% of their pixels were deemed unclear, applying a blur estimation technique introduced by Tong et al. [47] to assess image clarity.

To demonstrate the impact of this refinement process quantitatively, we employed an image quality assessment algorithm developed by Zhang et al. [48] to evaluate the quality of the LR and HR aerial images within our dataset. According to the results, depicted in Fig. 5, more than 74% of the images that underwent this filtering process achieved quality scores above 0.7, underscoring the effectiveness of our method in enhancing the dataset's overall clarity and reliability for subsequent categorization tasks.

After assembling our extensive collection of LR and HR aerial images, the crucial task of annotating them for category labels commenced. A team of 106 volunteers—graduate students from our computer science department with ages ranging from 24 to 31 and equipped with experience in image processing and pattern recognition (comprising 57 males and 49 females)—initially manually annotated 23.8% of the HR aerial images across each metropolitan area, employing a total of 47 distinct category labels. Subsequently, we developed a multilabel support vector machine (SVM) classifier to extend these annotations to the remaining unlabeled LR and HR aerial images. Following the automated labeling process, the same team of volunteers reviewed and corrected the SVM-generated labels to ensure accuracy. Notably, we identified that several category labels were associated with a negligible number of images, rendering the development of a robust categorization model for these labels unfeasible. Consequently, we excluded any category label represented by fewer than 200 000 images, ultimately narrowing down to 18 distinct categories, as listed in Table I. Analysis revealed that 99.983% of the LR and HR aerial images were tagged with no more than four category labels. A minuscule fraction had a higher number of labels (ranging from 5 to 15), typically indicating images with numerous small regions (less than 200 × 200 pixels) that could potentially introduce noise. These images were thus excluded from the dataset. Finally, we organized the entire collection of LR and HR aerial images alphabetically by file name. All HR images were used for model training, with the first half of LR images in each category forming the training set and the remaining half designated for testing. This structured approach facilitated a comprehensive and systematic evaluation of our categorization model.

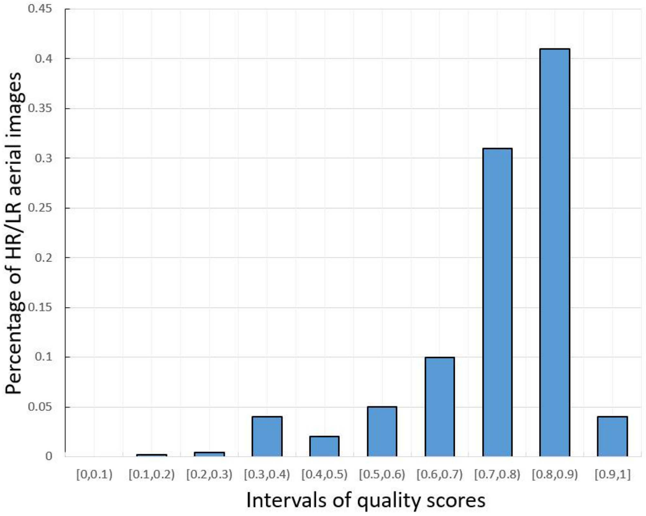


Fig. 5. Statistics of LR and HR aerial photos with different quality scores in our complied LR and HR aerial photo set.

TABLE I
SELECTED 18 CATEGORIES AND THE CORRESPONDING LR AND HR AERIAL PHOTO NUMBERS

Category	HR No.	LR No.	Category	HR No.	LR No.
Tall building	1,121,110	454,130	Residential	1,232,108	544,114
Forest	1,221,132	654,118	Sea	1,324,337	434,142
Aircraft	1,367,215	355,619	Railway	1,254,005	476,094
Road	1,556,540	453,884	River	1,324,337	435,093
Palace	1,375,547	546,881	Factory	1,443,672	509,448
Vehicle	1,325,443	621,214	Yacht	1,324,216	432,116
Intersection	1,414,214	315,446	Soccer field	1,116,436	454,338
Bridge	1,211,548	324,801	Park	1,325,658	342,556
Farmland	1436,658	543,447	Swim. pool	1,213,008	376,643

B. Comparative Study

1) Accuracy Comparison: In this experiment, we evaluate our LR aerial photo categorization by comparing its effectiveness and efficiency with a bunch of counterparts. We first compare our method with deep architectures tailored for aerial

TABLE II
ACCURACIES WITH STANDARD ERRORS OF THE 18 CATEGORIZATION MODELS

Category	[23]	[24]	[25]	[26]	[27]	[28]	[29]	SPP-CNN	CleanNet
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
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
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[22]	[44]	[46]	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	**0.716±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	**0.664±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	**0.768±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	**0.759±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	**0.698±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	**0.617±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	**0.759±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	**0.685±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	**0.598±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	**0.684±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	**0.748±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	**0.703±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	**0.672±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	**0.662±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	**0.627±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	**0.699±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	**0.779±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	**0.680±0.011**

We repeat each experiment ten times and report the average accuracies and each bold number represents the best result).

photo categorization. Moreover, we compare our method with state-of-the-art deep generic object/scene recognition models.

In the first place, we compare our method 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 reimplement them using Python by ourselves. We have tried our best to make the reimplemented models perform similarly to the results reported in their publications.

Nowadays, many deep generic recognition models perform impressively on categorizing aerial photos. In this experiment, we first compare our method with ten deep generic object categorization models: the spatial pyramid pooling CNN (SPP-CNN) [52], CleanNet [13], discriminative filter bank (DFB) [14], multilayer CNN-RNN (ML-CRNN) [21], multilabel graph convolutional network (ML-GCN) [45], semantic-specific graph (SSG) [46], and multilabel transformer (MLT) [49]. Furthermore, since LR aerial photo categorization can be deemed as a subtopic of scenery classification, we additionally compare our method with three well-known scenery classification models [22], [42], [44]. For these models, only the source codes of [22] are unavailable. Thus, we reimplement them using C++.

For the categorization models implemented by ourselves, the experimental setups are briefed as follows. In [25], we utilize the ResNet-152 [34] as the backbone, which is subsequently upgraded into a multilabel variant. Except for the last fully connected layer (unit number is fixed at 13), the other layers are initialized by the ResNet-152 trained from ImageNet [53]. For [26], the weights in the 1536-D long short-term memory layer are initialized by a random number between −0.2 and 0.2. Meanwhile, the Nesterov Adam is deployed as the optimizer, wherein the learning rate is set to 1e-6. For [29], the domain adaptation is implemented from the RSSCN7 [28] to our compiled LR and HR aerial photo set. The ResNet101V2 [34] is employed as the backbone, and the stochastic gradient descent optimizes the entire network. The learning ratio and weight decay are set to 1e-4 and 0.03, respectively. The network loss is calculated by the mean squared error. For [22], we retrain the object bank [51] based on our refined 18 LR and HR aerial photo categories, wherein the average pooling strategy is applied. We employ the liblinear as the SVM solver, wherein the sevenfold cross validation is utilized.

For the above 18 compared object/scene categorization models, we repeatedly test each model ten times, and the average accuracies are displayed in Table II. 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

TABLE III
TRAINING/TESTING TIME OF THE 18 CATEGORIZATION MODELS

	[23]	[24]	[25]	[26]	[27]	[28]	[29]	SPP-CNN	CleanNet
Train	31 h 7 m	43 h 14 m	52 h 21 m	39 h 23 m	36 h 43 m	46h13m	41 h 32 m	26 h33 m	38 h 22 m
Test	1.143 s	1.774 s	1.846 s	1.564 s	2.437 s	1.463s	1.675 s	0.893 s	1.660 s
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[22]	[44]	[46]	Ours
Train	40 h 23 m	**25 h 25 m**	32 h 15 m	44 h 16 m	31 h 16 m	32h14m	35 h 44 m	32 h 12 m	27 h 21 m
Test	1.213 s	1.002 s	1.875 s	0.983 s	1.436 s	1.774s	1.983 s	1.546 s	**0.477 s**

Each bold number represents the best result.

each original aerial photo to a fixed and much smaller size (e.g., 128×128) 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. However, 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. However, our collected LR aerial photos are captured at overhead view angles. Apparently, such view angle gap will decrease the categorization accuracy.

2) Training/testing Time Comparison: It is generally acknowledged that time consumption is a key criterion reflecting the performance of a categorization model. Herein, we report the training and testing time of the aforementioned 18 aerial photo categorization models. As shown in Table III, during training, only two baseline models are faster than our pipeline. This is because the architectures of [45] and [52] are much simpler than ours. Meanwhile, we observe that the per-category accuracies of [45] and [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 artificial intelligence 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: 9 h 12 min (module 1), 10 h 11 min (module 2), and 3 h 58 min (module 3). During testing, the time cost of each module is: 77 ms (module 1), 3 ms (module 2), and 12 ms (module 3). We observe that most of the training time is spent for module 1, and practically this can be accelerated by Nvidia GPUs.

C. Parameter Analysis

In our analysis, we identified two distinct groups of adjustable parameters that significantly impact performance. The first group includes the weights (α, β, and γ) that regulate various aspects of the deep low-rank model, alongside the number of deep layers, L. The second group comprises the polynomial kernel degree, Q, and the target dimensionality for the CPKP-based FS, V. This section details the impact of these parameters on the accuracy of LR aerial photo categorization. For the initial parameter set (α, β, γ, and L), we established default values of 0.3, 0.1, 0.15, and 7, respectively, determined through a comprehensive tenfold cross-validation process involving a validation set of 54 000 samples. This set was composed of selecting 3000 LR aerial photos from each category. Specifically, we varied α, β, and γ from 0.05 to 1 in increments of 0.05, exhaustively testing all possible combinations to identify the configuration yielding the highest categorization accuracy. Following this, we individually adjusted each of these three parameters from 0.05 to 1 in increments of 0.01, observing their corresponding effects on categorization accuracy. As illustrated in the top part of Fig. 6, each parameter showed a consistent increase in performance up to a peak, followed by a decline, demonstrating a monotonic relationship that facilitates optimal parameter tuning.

Furthermore, the impact of varying L from 1 to 7 demonstrated a steady improvement in categorization accuracy, which then plateaued, as shown in the lower part of Fig. 6. This observation suggests that while a deeper model provides a more complex representation capability, it also introduces a higher risk of overfitting. Consequently, we selected $L = 7$ as the optimal setting, balancing model depth with the potential for overfitting.

Next, we evaluate LR aerial photo categorization by changing the polynomial kernel degree Q and the target dimensionality for CPKP-based FS V. We first fix V and tune Q from one to five and

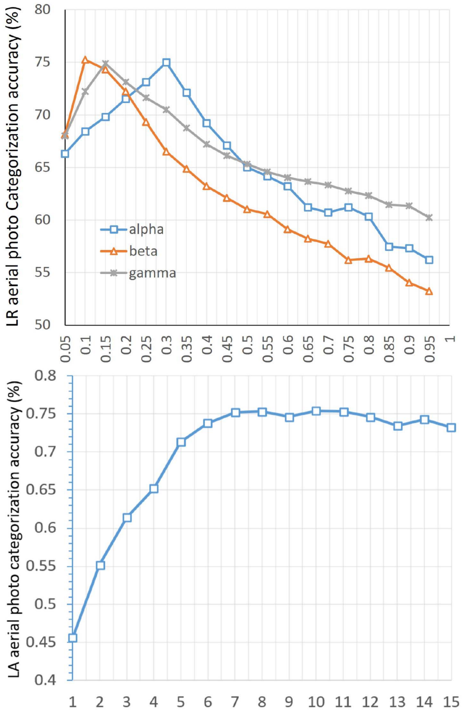


Fig. 6. LR aerial photo categorization accuracies by varying α, β, and γ (left) and L (right).

TABLE IV
CATEGORIZATION ACCURACIES BY TUNING Q

Q	Accuracy	Candidate feature no.
1	65.212%	133
2	75.443%	8 778
3	73.320%	383 306
4	64.332%	12 457 445
5	60.321	321 402 081

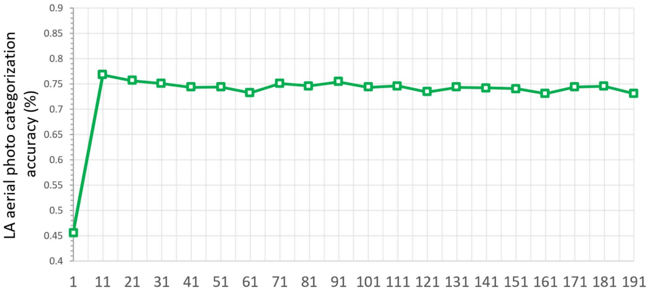


Fig. 7. LR aerial photo categorization accuracies by varying V.

report the LR aerial photo categorization accuracy. As shown in Table IV, the highest accuracy is achieved when $Q = 2$. Meanwhile, we observe that the candidate feature number increases to 321 402 081 when $Q = 5$. Based on these observations, we prone to choose a small Q in practice. Then, we fix Q at $Q = 2$ tune V from zero to 100. As reported in Fig. 7, the highest categorization accuracy is achieved when $V = 18$. This demonstrates that a succinct set of high-quality features is sufficiently descriptive for distinguishing different LR aerial photo categories.

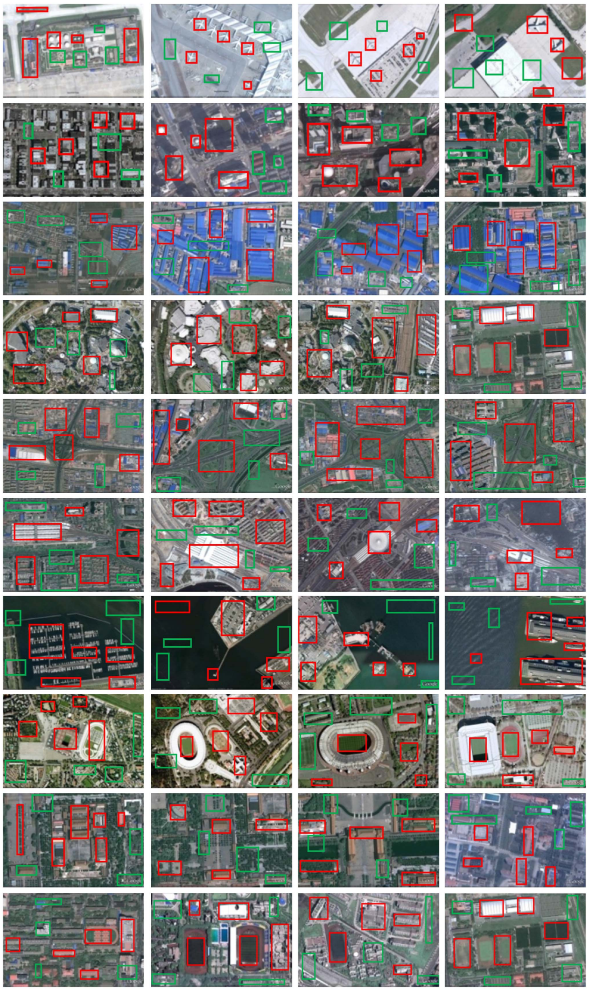


Fig. 8. Visualized selected discriminative regions (red boxes) from ten representative LR aerial photo categories. In each aerial photo, we set the number of extracted BING object patches to nine. Thus, we obtain nine boxes; each corresponds to a BING object patch. Subsequently, for each aerial photo, we use our CPKP-based FS to select five discriminative BING object patches from the nine extracted ones. Each selected discriminative BING object patch is colored by red, whereas the unselected ones are colored by green. As shown, the red boxes usually enclose a discriminative and central object. This shows the effectiveness of our CPKP in selecting high-quality features for LR aerial image categorization. For example, the first row corresponds to category aircraft; most of the red boxes localize the aircrafts.

D. Ablation Study

As aforementioned, our method is comprised of three key modules: 1) GSP learning using the deep low-rank algorithm; 2) our designed CPKP-based FS; and 3) feature classification for category labels. Herein, we validate their usefulness and inseparability in our LR aerial photo categorization pipeline. We replace each module by a functionally degraded one and report the categorization accuracy decrement/increment. Accordingly, insights are provided to elaborate the underlying reasons for the observed results.

In the first place, to evaluate the effectiveness of the deep low-rank algorithm, three experimental settings are deployed. We first abandon the sparse constraint term $||\mathbf{E}||_1$ in (11) (marked by "S11"). Afterward, we degrade the deep feature engineering term $||\mathbf{Y} - \mathbf{Z}h(\mathbf{X})||_F^2$ to a shallow one, that is, we

TABLE V
LR AERIAL PHOTO CATEGORIZATION ACCURACY DECREMENTS ("−") AND INCREMENTS ("+") BY REPLACING EACH OF THE THREE KEY MODULES

	S1	S2	S3
O1	–3.511%	–3.213%	–4.674%
O2	–6.756%	–3.021%	–4.890%
O3	–5.440%	–3.657%%	–2.774%
O4	N/A%	–6.435%	N/A
O5	N/A	–6.231%	N/A

set $L = 1$ (marked by "S12"). Third, we abandon the regularizer $||\mathbf{Z}||_F^2 + \sum_{i=1}^{L}(||\mathbf{Z}_l||_F^2 + ||\xi||_2^2)$ in (11) (marked by "S13"). We report the variation of categorization accuracy in Table V, where the intersection of column "Si" and row "Oj" corresponds to experimental setup "Sij." Noticeably, using a shallow feature engineering module results in a sharp categorization accuracy drop. Moreover, abandoning the regularizer significantly hurts the categorization accuracy. This observation shows the necessity to mitigate the overfitting of our adopted low-rank algorithm.

Subsequently, to evaluate the performance of our CPKP-based FS, six different setups are applied to testify the usefulness of the criteria in (19). We first replace the polynomial kernel by sigmoid kernel (marked by "S21"), linear kernel (marked by "S22"), and RBF (marked by "S23"). Afterward, we sequentially abandon the first two terms in (19). We mark them by "S24" (remove η_1) and "S25" (remove η_2), respectively. As shown in the third column of Table V, minimizing only one of the two distribution discrepancies significantly hurts the LR aerial photo categorization. This observation quantitatively reflects the importance of distribution consistences in perceptual knowledge propagation.

Next, to evaluate the performance of the linear classifier formulated in (17), we first replace it by kernel SVM (with Gaussian (marked by "S31") and linear kernels (marked by "S32")). Afterward, we replace it by a softmax layer that outputs the category labels (marked by "S32"). As displayed in the last column of Table V, training SVM classifiers separately substantially degrades LR aerial photo categorization. This clearly demonstrates the superiority of jointly optimizing FS and classifier training.

Last but not least, we visualize the selected discriminative features (regions) from ten representative categories in our complied aerial photo set. To explicitly show the discriminative regions, our CPKP-based FS is conducted on the original feature space. As shown in Fig. 8, for each LR aerial photo, we display the selected five discriminative regions from the entire nine candidate ones. Obviously, most of the selected regions optimally enclose a discriminative and central object. This observation is highly consistent with human visual perception.

V. CONCLUSION

The task of aerial image recognition has become a critical application within the realm of deep neural networks [38], [39], [40], [41]. In this context, we have introduced an innovative pipeline for the categorization of LR aerial photos, enhanced through cross-resolution techniques that leverage the deep perceptual insights derived from HR aerial images. Our approach is structured around three principal components: 1) a deep low-rank model designed to extract GSP features from both LR and HR aerial imagery; 2) a CPKP-based FS mechanism that identifies and selects premium quality features within a high-order feature space; and 3) a linear classifier that is synergistically trained in conjunction with the CPKP-based FS process. The breadth of our experimental analysis confirms the effectiveness and competitive edge of our methodology in aerial photo categorization.

REFERENCES

[1] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Deep learning for vehicle detection in aerial images," in *Proc. Int. Conf. Image Process.*, 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, 2019, Art. no. 133b.

[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, no. 8, pp. 2265–2278, 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. James, 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.*, 2015, pp. 4576–4584.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1–8.

[8] J. Porway, Q. Wang, and S.-C. Zhu, "A hierarchical and contextual model for aerial image parsing," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 254–283, 2010.

[9] 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 Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4095–4104.

[10] 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.*, 2016, pp. 680–688.

[11] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 145, pp. 60–77, 2018.

[12] 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, 2022, Art. no. 8002005.

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

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

[15] 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.*, vol. 27, no. 6, pp. 1896–1911, Dec. 2023.

[16] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint 2, 1-norms minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1813–1821.

[17] N. Gui, D. Ge, and Z. Hu, "AFS: An attention-based mechanism for supervised feature selection," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3705–3713.

[18] D. Ming and C. Ding, "Robust flexible feature selection via exclusive L21 regularization," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 3158–3164.

[19] 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.

[20] 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.

[21] A. Caglayan and A. B. Can, "Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 13–21.

[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. Int. Conf. Pattern Recognit. Appl. Methods*, 2015, pp. 111–132.

[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 Conf. Comput. Vis. Pattern Recognit. Workshops*, 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. Int. Geosci. Remote Sens. Symp.*, 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*, 2019, pp. 1–4.

[27] M. Pritt and G. Chern, "Satellite image classification with deep learning," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 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.*, 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. J. Smola, A. Gretton, J. Bedo, and K. M. Borgwardt, "Feature selection via dependence maximization," *J. Mach. Learn. Res.*, vol. 13, pp. 1393–1434, 2012.

[31] M. Masaeli, J. G. Dy, and G. M. Fung, "From transformation-based dimensionality reduction to feature selection," in *Proc. 27th 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, 2014.

[33] J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel feature selection via conditional covariance minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 6949–6958.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[35] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, "Remote-sensing scene classification via multistage self-guided separation network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615312.

[36] J. Wang, W. Li, M. Zhang, and J. Chanussot, "Large kernel sparse ConvNet weighted by multi-frequency attention for remote sensing scene understanding," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5626112.

[37] M. Zhang, W. Li, X. Zhao, H. Liu, R. Tao, and Q. Du, "Morphological transformation and spatial-logical aggregation for tree species classification using hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5501212.

[38] B. Tu, Z. Wang, H. Ouyang, X. Yang, J. Li, and A. Plaza, "Hyperspectral anomaly detection using the spectral-spatial graph," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5542814.

[39] B. Tu, X. Yang, X. Ou, G. Zhang, J. Li, and A. Plaza, "Ensemble entropy metric for hyperspectral anomaly detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5513617.

[40] Y. Li et al., "Fusing Sentinel-2 and Landsat-8 surface reflectance data via pixel-wise local normalization," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 7359–7374, 2022.

[41] L. Liang, S. Zhang, and J. Li, "Multiscale DenseNet meets with Bi-RNN for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 5401–5415, 2022.

[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.*, 2016, pp. 571–579.

[43] D. Hong et al., "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/CVF Int. Conf. Comput. Vis.*, 2017, pp. 5757–5765.

[45] Z.-M. Chen, X.-S. Wei, P. Wang, and P. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5177–5186.

[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.*, 2019, pp. 522–531.

[47] H. Tong, M. Li, H. J. Zhang, and C. Zhang, "Blur detection for digital images using wavelet transform," in *Proc. IEEE Int. Conf. Multimedia Expo*, 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, 2014, Art. no. 012183.

[49] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16473–16483.

[50] R. Diestel, *Graph Theory*. Berlin, Germany: Springer, 2005.

[51] Li-Jia Li, H. Su, Eric P. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.

[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, Li-Jia Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[54] V. Chalavadi, J. Prudviraj, R. Datla, C. Sobhan, K. Babu, and C. Mohan, "mSODANet: A network for multi-scale object detection in aerial images using hierarchical dilated convolutions," *Pattern Recognit.*, vol. 126, 2022, Art. no. 108548.

[55] S. Zhou et al., "DeepWind: Weakly supervised localization of wind turbines in satellite imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1–6.

[56] L. Cao et al., "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognit.*, vol. 64, pp. 417–424, 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. Int. Geosci. Remote Sens. Symp.*, 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] O. Akar, "Mapping land use with using rotation forest algorithm from UAV images," *Geocarto Int.*, vol. 33, no. 5, pp. 538–C553, 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. Sens.*, vol. 2018, 2018, Art. no. 7195432.

[63] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 204–213.

[64] Y. Bazi et al., "Vision transformers for remote sensing image classification," *Remote Sens.*, vol. 13, no. 3, 2021, Art. no. 516.