IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 2 August 2023, accepted 18 August 2023, date of publication 16 October 2023, date of current version 8 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3323701

RESEARCH ARTICLE

High-Resolution Aerial Photo Categorization Model by Cross-Resolution Perceptual Experiences Transfer

SIDA LI[1] AND YE LIU[2]
[1]Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321007, China
[2]College of Computer Sciences, Zhejiang University, Hangzhou 310058, China

Corresponding author: Sida Li (20131044@jhc.edu.cn)

This work was supported by the Special Project of Center for Scientific Research and Development in Higher Education Institutes of Ministry of Education under Grant ZJXF2022174.

ABSTRACT There are thousands of observation satellites orbiting the earth, each of which captures massive-scale photographs covering millions of square kilometers everyday. In practice, these aerial photos are with high-resolution and usually contain tens to hundreds of ground objects (e.g., vehicles and rooftops). Understanding the categories of a rich variety of high-resolution aerial photos is an indispensable technique for many applications, such as intelligent transportation, natural disaster prediction, and smart agriculture. In this work, we propose a cross-resolution perceptual experiences transfer framework for categorizing high-resolution aerial photos, focusing on leveraging the perceptual features from low-resolution aerial photos to enhance the feature selection of high-resolution ones. More specifically, we first construct gaze shifting path to mimic human visual perception to both low-resolution and high-resolution aerial photos, wherein the corresponding deep gaze shifting path features are engineered. Afterward, a kernel-induced feature selection algorithm is formulated to obtain a succinct set of deep gaze shifting path features discriminative across low- and high-resolution aerial photos. Based on the selected features, low- and high-resolution aerial photos' labels are collaboratively utilized to train a linear classifier for categorizing high-resolution ones. Extensive comparative studies have validated the superiority of our method.

INDEX TERMS High-resolution, human visual perception, perception experiences, feature selection.

I. INTRODUCTION

Due to the development of delivering plenty of satellites during a single rocket, there are many earth observation satellites launched since 1980. As we know, high-resolution aerial photos (typical resolutions over $5K \times 5K$) containing ground objects with sophisticated spatial interactions are well captured by these satellites. Semantically understanding these ground objects as well as the inherent spatial topologies is an important technology in lots of state-of-the-art AI systems. As an example, we can spatially parse the distribution of different animals and forests. Then we can intelligent understand the trends of wildlife. Such application is informative for keeping habitats in the sanctuaries, especially for the endangered animals.

In geoscience and remote sensing, searchers have designed many visual annotation or classification models to characterize aerial images with normal resolutions (typically $800 \times 800 \sim 2K \times 2PK$). Plenty of experiments and modern AI systems have demonstrated their superior performance and convenience. Nevertheless, in practice, the previous models cannot effectively encode high-resolution aerial photos because of the following reasons:

1) Typically, there exists a rich set of multi-scale foreground objects inside an high-resolution aerial photo, as shown in Fig. 1. To calculate the semantics of an high-resolution aerial photo, we expect a bionic model that simulates the process of human perceiving the foreground salient regions. Actually, building a deep model that can simultaneously extract the visually/semantically salient regions and engineer the deep features for these extracted regions is non-trivial.

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

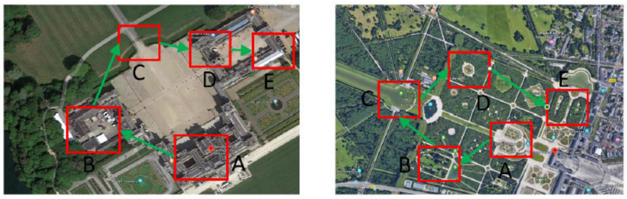

FIGURE 1. Pairwise high-resolution aerial photos with their gaze shifting paths.

2) Toward an efficient and interpretable image model for semantic understanding, we want high quality features shared between high- and low-resolution aerial images. 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 toward the high-order feature space computationally intractable.

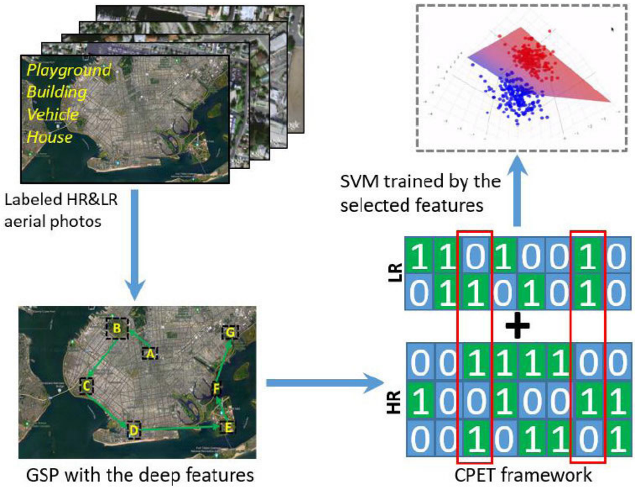

FIGURE 2. Categorizing aerial photos with high-resolution by leveraging cross-resolution perceptual experiences transfer.

We design a new cross-resolution perceptual experiences transfer framework that adopts the deeply-learned perceptual experiences of low-resolution aerial images to facilitate categorizing high-resolution one. An overview of our low-resolution aerial photo categorization is presented in Fig. 2. By utilizing a considerable quantity of high-resolution and low-resolution aerial photos. A machine learning algorithm is used to detect those salient regions, based on which the gaze shifting paths are generated and the deep features are calculated. Aiming at a concise set of discriminative features shared between high- and low-resolution aerial images, we explicitly map the deep gaze shifting path features onto a high-order and kernel-induced feature space. To inherit the perceptual knowledge of low-resolution aerial photos, a feature selection algorithm is developed to jointly 1) minimize the marginal/conditional distribution discrepancy between high-resolution and low-resolution aerial photos, and 2) maximize the linear classification accuracy. Based on the selected features, both labeled high-resolution and low-resolution aerial photos are employed to train the classifier. This can mitigate the sample insufficiency problem, which may cause the classifier overfitting during high-resolution aerial photo categorization. Comparative study with 17 image recognition models have demonstrated the advantage of our method.

II. RELATED WORK

Dozens of image recognition models were developed to analyze aerial photos. For image-level modeling, Chalavadi et al. [34] constructed a novel topological feature to model the inter-region connection inside each aerial photo. And a kernel-induced vector is calculated as the image representation for categorization. The authors [35] presented a weak model that semantically labels high-resolution aerial photos at image-level. The authors [36] proposed to combine the so-called random forest and semantics-aware feature extractor to classify each aerial photo into multiple categories. Akar et al. [37] developed a hierarchical CNN architecture for annotating the multiple labels of high-resolution aerial photos describing many downtown areas. Cai and Wei [5] proposed a cross-attention mechanism to learn the weights of aerial image features both horizontally and vertically. Costea et al. [39] formulated a vision transformer for aerial image classification, wherein the long-term contextual dependencies among regions can be intrinsically encoded.

For region-level modeling, Pan et al. [4] formulated a novel deep neural network for discovering multi-scale salient objects within each aerial photo. In [1], a focal loss deep architecture is proposed that optimally discovers vehicles from aerial images. Sameen et al. [38] developed a geo-localization model toward aerial photos by intelligently extracting intersections and streets. Wang et al. [8] integrated feature enhancement and soft label assignment into an anchor-independent object detector toward aerial images. Yu et al. [9] proposed a deep rotation-invariant detector that effectively estimates the angles of multi-scale objects inside aerial images. The authors [31] proposed a parallel deep model called mSODANet that hierarchically learns contextual features from multi-scale and multi-FoV (field-of-views) ground objects. Notably, different from the above methods, our approach is bionic-inspired and accurately mimics human gaze behavior.

III. OUR PROPOSED METHOD

A. DEEP GAZE SHIFTING PATH LEARNING

There are hundreds of objects and their parts in each high-resolution aerial photo. According to the recent biological and psychological studies [2], humans typically attend a succinct set of visually prominent objects in their visual perception process. When human perceiving a high-resolution image, human vision system will perceive the foreground salient objects beforehand, such as an aircraft and its components. Meanwhile, the remaining backgrounds are typically kept unhandled in practice. We have to incorporate such human visual perceptual experience in a high-resolution aerial photo categorization task. Herein, a rapid object parts extraction coupled with a novel active learning paradigm is deployed to detect the foreground salient objects.

The well-known BING [7] operator is leveraged as the object descriptor. By applying the BING operator, we receive a rich set of object patches inside a high-resolution aerial photo. Actually, humans usually attend to very few objects within each scenery. To mimic this, we use an effective active learning [6] to sequentially find K representative object patches from each high-resolution aerial photo. It encodes the following attributes: 1) high-resolution aerial photo's spatial features and 2) object patches' semantic labels.

Based on the sequentially selected K object patches, each path is constructed by connecting the K object patches (as the path $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E \rightarrow F \rightarrow G$ exemplified in Fig. 2). The constituent object patches and their spatial interactions simultaneously contribute to the gaze shifting path's appearance. Herein, given a K-sized gaze shifting path, we represent it by matrix $G = [G_1, G_2]$, where G_1 is a $K \times T$-sized matrix. The T dimensions describe the CNN feature from each image patch within a gaze shifting path. G_2 represents the $K \times K$ matrix indicating node linkage. Toward a simple yet effective feature, matrix G is row-wise concatenated into a long feature vector $\mathbf{u}$.

B. CROSS-RESOLUTION PERCEPTUAL EXPERIENCES TRANSFER

Theoretically, the extracted deep gaze shifting path features are usually distributed in the high-dimensional high-order feature space. Comparatively, the number of labeled high-resolution aerial photos is relatively small. This inevitably causes the dimensionality curse and will in turn hurt high-resolution aerial photo categorization. To handle this problem, a cross-resolution perceptual experiences transfer framework is formulated to select a succinct set of highly discriminative features shared between high-resolution and low-resolution aerial photos. Thereby, the selected features from high-resolution and low-resolution aerial photos can be collaboratively utilized to train the categorization model. In a word, cross-resolution perceptual experiences transfer can simultaneously reduce the feature dimensionality and increase the training sample number, based on which the dimensionality curse can be mitigated substantially.

1) FEATURE MAPPING BY APPROXIMATING POLYNOMIAL KERNEL

The polynomial kernel can be mathematically represented as:

$$\varphi\left(\mathbf{u}, \mathbf{v}\right) = \left(\tau \mu^{T} + \kappa\right)^{Q}, \tag{1}$$

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

$$\varphi_{q,e}\left(\mathbf{u}\right) = \sqrt{C_{Q}^{q} \cdot \kappa^{Q-q} \prod\nolimits_{j=1}^{q} \mathbf{u}_{e_j}}, i = 1, \cdots, Q, \tag{2}$$

where $e \in \{1, \cdots, K\left(K+T\right)\}^{q}$ enumerates the entire selections of q-dimensional coordinates in $\mathbf{u}$, and $K(K+T)$ is the dimensionality of deep gaze shifting path feature. By leveraging the multinomial theorem, (2) can be reorganized into:

$$\varphi\left(\mathbf{u}\right) = \cup_{q=1}^{Q}\{\varphi_{q,e\in\{1,\cdots,K(K+T)\}^{q}}(\mathbf{u})\}, \tag{3}$$

For degree Q, there are a total of $S = C_{K(K+T)+Q}^{Q}$ candidate features for feature selection, where operator C_{i}^{j} counts the combinations of selecting j features from i features.

2) OBJECTIVE FUNCTION OF FEATURE SELECTION

By leveraging the above explicit feature map, deep gaze shifting path feature engineered from high-resolution and low-resolution aerial photos can be represented by $\{\left(\varphi\left(\mathbf{u}_i\right) \in \mathbb{R}^{\mathbf{S}}\right), r_i^{H}\}_{i=1}^{M^H}$ and $\{\left(\varphi\left(\mathbf{u}_i\right) \in \mathbb{R}^{\mathbf{S}}\right), r_i^{L}\}_{i=1}^{M^L}$ respectively, where M^H and M^L count the high-resolution and low-resolution aerial photos respectively. r^H and r^L denote the category labels of the high-resolution and low-resolution aerial photos respectively. Herein, a novel feature selection algorithm is proposed to select features discriminative to both high-resolution and low-resolution aerial photos.

We denote the high-resolution aerial photos as $\left\{\mathbf{u}_i^{H}, r_i^{H} \in \{1, \cdots, B\}\right\}_{i=1}^{M^H}$, where $\mathbf{u}_i^{H}$ denotes the $K(K+T)$-dimensional deep gaze shifting path feature and r_i^{H} the corresponding category label. We denote $\mathbf{U}^{H} = \{\mathbf{u}_i, r_i^{L}\}_{i=1}^{M^H}$ as deep gaze shifting path feature from the entire high-resolution aerial photos and the labels. Let $p^{H}\left(\mathbf{U}^{H}\right)$ and $p^{L}\left(\mathbf{U}^{L}\right)$ be the marginal distributions of $\mathbf{U}^{H}$ and $\mathbf{U}^{L}$. The objective of our feature selection is to select an optimal feature set that predicts labels $\{r_i^{H}\}_{i=1}^{M^H}$ using the input high-resolution aerial photos $\{\mathbf{u}_i^{H}\}_{i=1}^{M^H}$ under assumptions $p^{H}\left(\mathbf{u}^{H}\right) \neq p^{L}\left(\mathbf{u}^{L}\right)$ and $q^{H}\left(\mathbf{u}^{H}\right) \neq q^{L}\left(\mathbf{u}^{L}\right)$.

It is reasonable to assume that there exists a binary indicator $\mathbf{s} \in \{0, 1\}^{\mathbf{S}}$, such that $p\left(\varphi\left(\mathbf{u}^{H}\right) \odot \mathbf{s}\right) \approx p\left(\varphi\left(\mathbf{u}^{L}\right) \odot \mathbf{s}\right)$ and $p\left(\varphi\left(\mathbf{r}^{H}\right) \odot \mathbf{s}\right) \approx p\left(\varphi\left(\mathbf{r}^{L}\right) \odot \mathbf{s}\right)$, where $\odot$ denotes the inner product of pairwise matrices. Our target is to learn the indicator $\mathbf{s}$. Since we practically have insufficient high-resolution aerial photos, $\mathbf{s}$ cannot be effectively learned due to the overfitting problem. In this way, we propose to learn binary indicator s and a linear classifier H jointly, in order to satisfy the following three criteria: 1) the distance between the marginal distribution $p(\varphi\left(\mathbf{u}^{H}\right) \odot s)$ and $p(\varphi\left(\mathbf{u}^{L}\right) \odot s)$ is sufficiently small, 2) $\varphi\left(\mathbf{u}^{H}\right) \odot \mathbf{s}$ and $\varphi\left(\mathbf{u}^{L}\right) \odot \mathbf{s}$ preserve the discriminative dimensions of deep gaze shifting path features $\varphi(\mathbf{U}^{H})$ and $\varphi(\mathbf{U}^{L})$, based on which $p\left(\mathbf{r}^{H} \mid \varphi\left(\mathbf{u}^{H} \odot \mathbf{s}\right)\right) \approx p(\mathbf{r}^{L} | \varphi(\mathbf{u}^{L} \odot \mathbf{s}))$, and 3) the learned classifier $C\left(\mathbf{u}^{H}\right) \mathcal{D}\left(\varphi\left(\mathbf{u}^{H}\right) \odot \mathbf{s}\right)\mathbf{H}$ can optimally categorize the training low-resolution aerial photos $\varphi(\mathbf{u}^{L})$. These criteria can be mathematically represented as follows:

1) Marginal distribution discrepancy minimization: Given the polynomial-kernel-based feature mapping $\varphi(\mathbf{u})$ induced by (3), we aim to minimize the marginal

distribution discrepancy by feature selection. This can be formulated as:

$$\min_{\mathbf{s}\epsilon\mathbf{S}} \eta_1(\mathbf{s}) = \left\| \frac{\mathbf{1}}{M^H} \sum\nolimits_{\mathbf{u}^H\in\mathbf{U}^H} \varphi\left(\mathbf{u}^H\right)\odot\mathbf{s} - \frac{\mathbf{1}}{M^L} \sum\nolimits_{\mathbf{u}^L\in\mathbf{U}^L} \varphi\left(\mathbf{u}^L\right)\odot\mathbf{s} \right\|_F^{\mathbf{2}}, \quad (4)$$

where $\|\cdot\|_F^2$ denotes the squared Frobenius norm, the binary indicator's domain is represented by $\mathbf{S} = \{\mathbf{s}|\mathbf{s} \in \{\mathbf{0},\mathbf{1}\}^{\mathbf{S}}, \|\mathbf{s}\|_{\mathbf{0}} \leq \boldsymbol{A}$, and $\boldsymbol{A}$ is the maximum number of selected features.

Conditional distribution discrepancy minimization: Practically, the posterior probabilities $q^H(r^H|u^H)$ and $q^L(r^L|u^L)$ have complicated forms. Instead, we utilize the classconditional distributions $q^H(r^H|u^H = b)$ and $q^L(r^L|u^L = b)$. More specifically, we first calculate the conditional distribution distance between high-resolution and low-resolution aerial photos labeled by $b \in \{1, \cdots, B$. Thereafter, we attempt to minimize the conditional distribution discrepancy:

$$\min_{\mathbf{s}\epsilon\mathbf{S}} \eta_2(\mathbf{s}) = \left\| \frac{\mathbf{1}}{M_b^H} \sum\nolimits_{\mathbf{u}^H\in\mathbf{U}_b^H} \varphi\left(\mathbf{u}^H\right)\odot\mathbf{s} - \frac{\mathbf{1}}{M_b^L} \sum\nolimits_{\mathbf{u}_b^L\in\mathbf{U}_b^L} \varphi\left(\mathbf{u}^L\right)\odot\mathbf{s} \right\|_F^{\mathbf{2}}, \quad (5)$$

where $\mathbf{U}_b^H$ and $\mathbf{U}_b^L$ denote the high-resolution and low-resolution aerial photos with category label $\boldsymbol{b}$. $\boldsymbol{M}_b^H$ and $\boldsymbol{M}_b^L$ count their number respectively.

Empirical error minimization: As we mentioned, we expect that the selected features not only minimize the distribution difference, but also be succinctly 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 (ECOC) [4] is employed. The empirical error of the high-resolution and low-resolution aerial photos will be minimized, i.e.,

$$\min_{\mathbf{s}\epsilon\mathbf{S}} \min_{\mathbf{H}} \eta_3(\mathbf{s},\mathbf{H}) = \sum\nolimits_{\mathbf{u}_i\in\mathbf{U}^H} \frac{\mathbf{1}}{\mathbf{2}} \|\boldsymbol{\epsilon}_i\|_F^2 + \frac{\boldsymbol{\psi}}{\mathbf{2}} \|\mathbf{H}\|_F^{\mathbf{2}}$$
$$\text{s.t. } \boldsymbol{\epsilon}_i \in (\boldsymbol{\varphi}(\mathbf{u}_i)\odot\mathbf{s})\,\boldsymbol{H} - \mathbf{r}_i, \boldsymbol{i} = 1, \cdots \boldsymbol{M}^{\mathbf{H}} + \boldsymbol{M}^L, \quad (6)$$

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

$$\min_{\mathbf{s}\epsilon\mathbf{S}} \min_{\mathbf{H}} \eta(\boldsymbol{s},\boldsymbol{H}) = \eta_1(\mathbf{s}) + \eta_2(\mathbf{s}) + \eta_3(\boldsymbol{s},\boldsymbol{H}),$$
$$\boldsymbol{s.t.}, \boldsymbol{\epsilon}_i \in (\boldsymbol{\varphi}(\mathbf{u}_i)\odot\mathbf{s})\,\boldsymbol{H} - \mathbf{r}_i, \boldsymbol{i} = 1, \cdots \boldsymbol{M}^{\mathbf{H}} + \boldsymbol{M}^L, \quad (7)$$

This objective function is NP-hard due to the combinatorial integral constraints on s. Herein, we adopt an efficient solution as detailed in the document [40].

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. COMPARATIVE STUDY

In this experiment, we evaluate our high-resolution 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 photo categorization. Then, our method is compared with multiple state-of-the-art deep generic object/scene recognition models. The experimental data set is from [35].

In the first place, we compare our method with seven deep categorization models [14], [15], [16], [17], [18], [19], [20] that intrinsically encode some prior knowledge of different aerial photo categories. We notice that the source codes of [14], [15], [18], and [19] are publicly available. Thereby, we conduct comparative study wherein the parameter settings are set as default. For [16], [17], and [20], the source codes are unavailable to our knowledge. In this way, we re-implement 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) [33], CleanNet [11], discriminative filter bank (DFB) [8], multi-layer CNN-RNN (MLCRNN) [12], multi-label graph convolutional network (MLGCN) [29], semantic-specific graph (SSG) [30] and multilabel transformer (MLT) [31]. Furthermore, since low-resolution aerial photo categorization can be deemed as a sub-topic of scenery classification, we additionally compare our method with three well-known scenery classification models [3-41], [26], [28]. For these models, only the source codes of [13] are unavailable. Thus we re-implement them using C++.

For the above 18 compared object/scene categorization models, we repeatedly test each model ten times and the average accuracies are displayed in Table 1. We method performs the best as expected. To quantify the stability of these categorization models, we report their standard errors simultaneously. 1) Our method outperforms the other aerial photo categorization models remarkably due to three reasons. First, to facilitate deep model training, our competitors typically resize each original aerial photo to a fixed and much smaller size (e.g. 128 × 128) for the subsequent hierarchical feature engineering. This hurts the learning of an low-resolution 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 high-resolution aerial photos. Third, only our method generates gaze shifting paths sequentially capturing the semantics of low-resolution aerial photos perceived by humans. They are further incorporated into a CPKP-based feature selection for calculating category labels. Comparatively, the seven counterparts only globally/locally characterize each low-resolution 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

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

Category	[14]	[15]	[16]	[17]	[18]	[19]	[20]	SPP-CNN	CleanNet
Tall building	0.642± 0.012	0.589±0.009	0.646±0.013	0.606±0.014	0.620±0.009	0.594 ±0.016	0.633±0.015	0.691±0.014	0.681±0.013
Residential	0.587±0.012	0.594±0.011	0.612±0.017	0.588±0.013	0.601±0.016	0.607±0.009	0.589±0.018	0.615±0.011	0.615±0.014
Intersection	0.703±0.014	0.715±0.012	0.694±0.016	0.685±0.014	0.716±0.019	0.684±0.017	0.721±0.010	0.684±0.013	0.695±0.012
Forest	0.684±0.013	0.673±0.014	0.698±0.013	0.664±0.014	0.682±0.012	0.658±0.014	0.685±0.012	0.713±0.011	0.705±0.014
Sea	0.674±0.013	0.647±0.015	0.684±0.015	0.633±0.013	0.665±0.017	0.646±0.018	0.673±0.013	0.662±0.013	0.686±0.010
Soccer field	0.546±0.014	0.565±0.016	0.587±0.013	0.577±0.016	0.583±0.009	0.562±0.014	0.584±0.012	0.570±0.021	0.583±0.018
Aircraft	0.732±0.012	0.704±0.014	0.721±0.013	0.695±0.015	0.705±0.013	0.718±0.017	0.685±0.015	0.716±0.014	0.705±0.013
Railway	0.621±0.014	0.613±0.016	0.635±0.013	0.643±0.015	0.607±0.015	0.596±0.016	0.605±0.014	0.614±0.017	0.616±0.015
Bridge	0.547±0.016	0.564±0.015	0.584±0.014	0.578±0.017	0.557±0.016	0.584±0.014	0.573±0.017	0.562±0.015	0.583±0.011
Road	0.613±0.013	0.624±0.012	0.635±0.014	0.615±0.016	0.625±0.013	0.621±0.014	0.605±0.016	0.616±0.013	0.627±0.014
River	0.721±0.015	0.708±0.017	0.716±0.010	0.716±0.014	0.726±0.013	0.699±0.013	0.702±0.015	0.709±0.016	0.715±0.019
Park	0.654±0.014	0.665±0.012	0.674±0.015	0.682±0.016	0.673±0.013	0.669±0.015	0.673±0.014	0.691±0.018	0.688±0.014
Palace	0.665±0.013	0.643±0.015	0.673±0.017	0.631±0.015	0.626±0.014	0.647±0.014	0.651±0.011	0.637±0.013	0.619±0.012
Factory	0.624±0.014	0.621±0.013	0.616±0.012	0.610±0.015	0.627±0.013	0.612±0.012	0.608±0.014	0.608±0.016	0.618±0.017
Farmland	0.604±0.013	0.602±0.016	0.608±0.012	0.598±0.016	0.584±0.014	0.614±0.013	0.592±0.015	0.609±0.018	0.611±0.16
Vehicle	0.685±0.009	0.674±0.013	0.658±0.011	0.694±0.015	0.653±0.012	0.668±0.014	0.670±0.016	0.684±0.014	0.671±0.014
Yacht	0.703±0.016	0.724±0.013	0.706±0.015	0.721±0.017	0.716±0.014	0.708±0.013	0.714±0.018	0.716±0.016	0.713±0.014
Swim. pool	0.654±0.014	0.636±0.012	0.641±0.015	0.652±0.013	0.633±0.016	0.665±0.011	0.673±0.015	0.631±0.013	0.636±0.018
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[13]	[26]	[28]	Ours
Tall building	0.625±0.013	0.664±0.014	0.659±0.012	0.682±0.016	0.673±0.014	0.625±0.014	0.642±0.016	0.647±0.014	**0.706±0.011**
Residential	0.594±0.014	0.614±0.013	0.618±0.012	0.624±0.015	0.613±0.014	0.576±0.015	0.597±0.016	0.588±0.014	**0.663±0.009**
Intersection	0.715±0.011	0.695±0.013	0.722±0.016	0.734±0.014	0.736±0.017	0.684±0.014	0.673±0.013	0.664±0.011	**0.778±0.007**
Forest	0.694±0.014	0.723±0.013	0.707±0.012	0.726±0.016	0.714±0.020	0.654±0.016	0.668±0.017	0.673±0.015	**0.758±0.009**
Sea	0.674±0.015	0.645±0.013	0.658±0.016	0.673±0.013	0.657±0.012	0.671±0.016	0.663±0.013	0.675±0.014	**0.697±0.010**
Soccer field	0.584±0.016	0.567±0.015	0.594±0.014	0.585±0.016	0.583±0.014	0.557±0.013	0.563±0.018	0.559±0.014	**0.615±0.012**
Aircraft	0.685±0.013	0.684±0.021	0.705±0.023	0.722±0.015	0.728±0.017	0.675±0.013	0.687±0.017	0.693±0.018	**0.761±0.007**
Railway	0.624±0.014	0.632±0.015	0.617±0.013	0.606±0.017	0.625±0.015	0.607±0.014	0.611±0.016	0.603±0.013	**0.683±0.006**
Bridge	0.564±0.015	0.547±0.017	0.568±0.013	0.574±0.016	0.536±0.017	0.530±0.014	0.543±0.013	0.532±0.016	**0.592±0.009**
Road	0.612±0.018	0.615±0.015	0.604±0.016	0.642±0.014	0.633±0.020	0.610±0.017	0.606±0.012	0.615±0.017	**0.682±0.008**
River	0.724±0.015	0.714±0.014	0.721±0.016	0.718±0.017	0.715±0.013	0.675±0.015	0.663±0.016	0.684±0.018	**0.761±0.008**
Park	0.674±0.015	0.663±0.017	0.690±0.018	0.684±0.014	0.684±0.016	0.694±0.017	0.682±0.015	0.683±0.014	**0.709±0.007**
Palace	0.624±0.011	0.631±0.023	0.614±0.018	0.621±0.019	0.635±0.017	0.596±0.016	0.604±0.015	0.609±0.014	**0.685±0.010**
Factory	0.614±0.015	0.608±0.016	0.612±0.012	0.607±0.017	0.614±0.015	0.603±0.017	0.615±0.013	0.613±0.012	**0.665±0.008**
Farmland	0.594±0.014	0.592±0.016	0.587±0.013	0.612±0.016	0.617±0.014	0.585±0.013	0.597±0.012	0.603±0.015	**0.6222±0.008**
Vehicle	0.654±0.016	0.685±0.016	0.675±0.017	0.646±0.014	0.686±0.016	0.639±0.014	0.654±0.017	0.673±0.016	**0.709±0.008**
Yacht	0.724±0.015	0.720±0.021	0.716±0.018	0.714±0.016	0.718±0.017	0.709±0.014	0.706±0.018	0.724±0.013	**0.781±0.007**
Swim. pool	0.621±0.016	0.654±0.014	0.643±0.017	0.657±0.015	0.626±0.014	0.607±0.013	0.614±0.013	0.628±0.016	**0.681±0.006**

discover the tiny but discriminative regions inside each low-resolution aerial photo. Second, our method can flexibly incorporate the prior knowledge of high-resolution aerial photos. Contrastively, the seven generic object recognition models cannot encode such information. Third, by leveraging our CPKP-based feature selection, our method can dynamically abandon those indiscriminative regions. But the seven generic object recognition models do not have this function. 3) The three scene categorization models perform unsatisfactorily on low-resolution 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 low-resolution aerial photo set. Moreover, the three categorization methods can successfully handle sceneries captured at horizontal view angles. But our collected low-resolution aerial photos are captured at overhead view angles. Apparently, such view angle gap will decrease the categorization accuracy. To quantitively analyze the importance of cross-resolution perceptual experiences transfer (CPET), we set the number of low-resolution aerial photo to zero. This means that no perceptual information from low-resolution aerial photos is utilized. We notice that the average categorization accuracy is reduced by 5.443%, which clearly shows the importance of cross-resolution perceptual experiences transfer.

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 categorization models. As shown in Table 2, during training, only two baseline categorization models outperform our pipeline. This is because the architectures of [29] and [33] are much simpler than ours. Simultaneously, we observe that the per-category accuracies of [29] and [33] are both about 5% lower than ours. For the testing time comparison, our method can be conducted at a much faster speed than all the baseline methods.

TABLE 2. Trainimg/testing time of the 18 categorizamon monels (Exch bold nimben represevts the best restut).

	[14]	[15]	[16]	[17]	[18]	[19]	[20]	SPP-CNN	CleanNet
Train	31h7m	43h14m	52h21m	39h23m	36h43m	46h13m	41h32m	26h33m	38h22m
Test	1.143s	1.774s	1.846s	1.564s	2.437s	1.463s	1.675s	0.893s	1.660s
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[13]	[26]	[28]	Ours
Train	40h23m	**25h25m**	26h14m	44h16m	31h16m	32h14m	35h44m	32h12m	27h21m
Test	1.213s	1.002s	1.875s	0.983s	1.436s	1.774s	1.983s	1.546s	**0.477s**

B. PARAMETER ANALYSIS

We evaluate high-resolution aerial photo categorization by changing the polynomial kernel degree Q and the target dimensionality V for cross-resolution perceptual experiences transfer-based feature selection. We first fix V and tune Q from one to five and report the high-resolution aerial photo categorization accuracy. We observe that the highest accuracy is achieved when $Q = 2$. Meanwhile, we observe that the candidate feature number increases to 321402081 when $Q = 5$. Based on these observations, we prone to choose a small Q in practice. Subsequently, we fix Q at $Q = 2$ tune V from one to 100. Noticeably, the highest categorization accuracy is achieved when $V = 15$. This demonstrates that a succinct set of high quality features is sufficiently descriptive for distinguishing different high-resolution aerial photo categories.

V. CONCLUSION

Recognizing aerial images is an indispensable application in remote sensing [21], [22], [23], [24], [25]. We proposed a novel cross-resolution-enhanced high-resolution aerial photo categorization pipeline, wherein deep perceptual features are extracted and refined by propagating the prior knowledge of low-resolution aerial photos into high-resolution ones. Sufficient experiments have shown the competitiveness of our proposed method.

REFERENCES

[1] M. Y. Yang, W. Liao, X. Li, and B. Rosenhahn, "Deep learning for vehicle detection in aerial images," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 3079–3083.

[2] F. van Ede, S. R. Chekroud, and A. C. Nobre, "Human gaze tracks the focusing of attention within the internal space of visual working memory," *J. Vis.*, vol. 19, no. 10, p. 133b, Sep. 2019.

[3] 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.

[4] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[5] 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.

[6] 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.

[7] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P. L. Rosin, and P. H. S. Torr, "Bing: Binarized normed gradients for objectness estimation at 300fps," *Comput. Vis. Media*, vol. 5, no. 1, pp. 3–20, Mar. 2019.

[8] 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.

[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. 5624216.

[10] 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.

[11] 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.

[12] A. Caglayan and A. B. Can, "Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition," in *Proc. ECCV Workshops*, 2018, pp. 1–8.

[13] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorizations," in *Proc. PRAM*, 2015, pp. 22–41.

[14] 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.

[15] C. Kyrkou and T. Theocharides, "Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 517–525.

[16] Y. Hua, S. Lobry, L. Mou, D. Tuia, and X. X. Zhu, "Learning multi-label aerial image classification under label noise: A regularization approach using word embeddings," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep. 2020, pp. 525–528.

[17] 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.

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

[19] H. Sun, Y. Lin, Q. Zou, S. Song, J. Fang, and H. Yu, "Convolutional neural networks based remote sensing scene classification under clear and cloudy environments," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 713–720.

[20] 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.

[21] W. Mu and B. Liu, "Voice activity detection optimized by adaptive attention span transformer," *IEEE Access*, vol. 11, pp. 31238–31243, 2023.

[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] 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.

[27] 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.

[28] Y. Li, M. Dixit, and N. Vasconcelos, "Deep scene image classification with the MFAFVNet," in *Proc. ICCV*, 2017, pp. 5746–5754.

[29] 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.

[30] 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.

[31] 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.

[32] R. Diestel, *Graph Theory*. Springer-Velag, 2005.

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

[35] L. Zhang, Z. Pan, and L. Shao, "Semi-supervised perception augmentation for aerial photo topologies understanding," *IEEE Trans. Image Process.*, vol. 30, pp. 7803–7814, 2021.

[36] 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.

[37] Ö. Akar, "Mapping land use with using rotation forest algorithm from UAV images," *Eur. J. Remote Sens.*, vol. 50, no. 1, pp. 269–279, Jan. 2017.

[38] 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.

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

[40] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," in *Proc. UAI*, 2011.

SIDA LI is currently a Faculty Member with Jinhua Polytechnic. His research interests include machine learning, computer vision, and image processing.

YE LIU is currently a Professor with the College of Computer Sciences, Zhejiang University, Hangzhou, China. His research interests include visual modeling and image processing.

• • •