Aerial Photograph Categorization by Cross-Resolution Deep Human Gaze Behavior Learning

Luming Zhang, Ming Chen, Guifeng Wang, and Zhiming Wang

***Abstract*—Accurately recognizing aerial photographs is a useful technique in many domains like autonomous driving and environmental evaluation. In practice, both low-resolution and high-resolution aerial photos are captured asynchronistically for each region, as there are hundreds of Earth observation satellites orbitting the Earth. Realizing such multiresolution-based region semantic understanding is a difficult task due to three challenges: 1) mimicking human visual perception when they actively viewing the semantic objects inside each aerial photo; 2) deeply modeling the visually/semantically salient objects sequentially perceived by human visual system; and 3) developing a cross-resolution knowledge transferal module to enhance the feature representation for an area. To solve these challenges, we propose a cross-domain aerial photograph categorization system by leveraging the low-resolution spatial composition to enhance the deep encoding of human gaze shifting path (GSP) with a high-resolution. More specifically, we first use an active learning algorithm to discover multiple visually/semantically salient object patches for constructing GSP from a high-resolution aerial photo. Then, an aggregation-based deep model is formulated to sequentially link the deep features learned from the object patches inside each GSP. Subsequently, a novel knowledge transferal algorithm leverages the global spatial composition from low-resolution counterparts to upgrade the deeply-learned GSP feature of the high-resolution aerial photo. Using the upgraded deep GSP feature, a multilabel SVM classifier is trained for categorizing aerial photographs. Comparative studies on our million-scale aerial photograph set have demonstrated the competitiveness of our approach.**

***Index Terms*—Active learning, aerial photo, cross domain, deep feature, gaze behavior, machine learning.**

I. Introduction

RECOGNIZING aerial photo's categories is a key technique in multiple modern remote sensing systems [26]–[28]. For example, in autonomous navigation, it is necessary to ensure that the smart vehicle can recognize the shortest path between two cities intelligently. This requires that a set of aerial-photograph-related cues, e.g., mountain terrain, traffic network topology, and road gradient, can be rapidly incorporated. Besides, calculating the regional categories of many low/high-resolution aerial photos can assist evaluating the village/city/state environment, e.g., the forest/crop coverage ratio and the impact of flood disaster. Moreover, in many video-based pedestrian tracking systems, it is standard to exploit local contexts (encoded by aerial photos) like road direction and intersection topology, to enhance the tracking accuracy.

In the past few years, hundreds of deep recognition models were designed for recognizing scene/object categories, such as the well-known AlexNet-CNN [1] and ResNet [42]. Experimental evaluations have demonstrated their advantages toward the shallow recognition models. Nevertheless, previous deep scene/object recognition models cannot fulfill aerial photograph categorization satisfactorily due to the following three shortcomings.

1) There are tens and hundreds of visual salient objects (such as rooftops and vehicles) within a high-resolution aerial photo. They are indicative to the process of how humans perceive an aerial photo, which is informative to recognize aerial photo's categories. However, it is difficult to propose a mathematical model extracting these salient objects and calculate the sequence of human gaze allocation simultaneously. Specifically, how to discover the GSP that mimics human visual perception? as exemplified on the top of Fig. 1.
2) Due to the impressive performance of deep representations in scene/object description, we believe that deep GSP features can well represent aerial photos both visually and semantically. Notably, the movements of human gaze and the GSP's geometry jointly capture how humans perceive each aerial photograph. As far as we know, only the entire image or its internal regions can be represented by the existing deep models, while the GSP and the path's geometry are not discovered. Integrating these two attributes into a unified deep and solvable recognition framework is challenging.
3) There are hundreds of satellites observing the Earth (e.g., GF1 and ZY302), where both low-resolution and

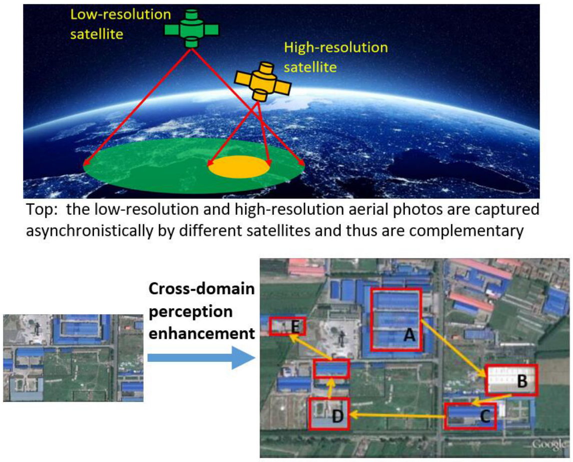
Manuscript received February 20, 2020; revised June 2, 2020 and July 9, 2020; accepted July 16, 2020. Date of publication June 2, 2022; date of current version June 24, 2022. This work was supported by the Basic Public Welfare Research Project of Zhejiang Province under Grant LGG19E050010. *(Corresponding author: Ming Chen.)*

Luming Zhang is with the Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Hangzhou 321016, China, and also with Hithink RoyalFlush Information Network Company, Ltd., Hangzhou 310023, China (e-mail: zglumg@gmail.com).

Ming Chen is with the Hithink RoyalFlush Information Network Company, Ltd., Hangzhou 310023, China (e-mail: chm_@zju.edu.cn).

Guifeng Wang and Zhiming Wang are with the Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Hangzhou 321016, China (e-mail: guifengwang@zju.edu.cn; zhimingwang@jin12.cn).

Digital Object Identifier 10.1109/JSTARS.2022.3179663


Low-resolution satellite
High-resolution satellite
Cross-domain perception enhancement
E
A
B
C
D
Google

Top: the low-resolution and high-resolution aerial photos are captured asynchronistically by different satellites and thus are complementary

Bottom: enhancing the high-resolution perceptual feature (GSP: $A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$) learning by leveraging the low-resolution global composition

Fig. 1. Top: The GSP ($A \rightarrow B \rightarrow C \rightarrow D \rightarrow E$) can simulate humans visual perception of an aerial photograph (the red italic text denotes the aerial photo's semantic categories). In detail, human first fix on region "A," and then shift his/her gaze onto region "B," and so on. Bottom: each landmark is captured by multiple low/high-resolution aerial photos in a multisatellite earth observation system.

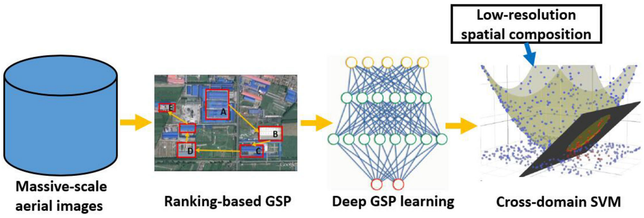

Low-resolution spatial composition
Massive-scale aerial images
Ranking-based GSP
Deep GSP learning
Cross-domain SVM
E
A
B
D
C

Fig. 2. Pipeline of our designed cross-domain deep perception-aware aerial photograph categorization model.

high-resolution aerial photographs are taken for an area, as exemplified at the bottom of Fig. 1. As the low-resolution and high-resolution aerial photos are captured asynchronistically, they characterize an area complementarily. For example, a playground is occluded in a high-resolution aerial photo but clear in a low-resolution one. Thus we have to combine the deep features from both low-resolution and high-resolution aerial photos for visual recognition. But building a knowledge sharing mechanism that discriminatively fuzes deep features from aerial photos with multiple resolutions is unsolved.

To address the aforementioned problems, we propose a novel multisatellite aerial photo categorization framework, focusing on deeply learning high-resolution GSP features by leveraging the global spatial attributes from low-resolution aerial photos. An overview of our approach is shown in Fig. 2. Given a collection of high-resolution aerial photos, each corresponds to multiple low-resolution counterparts,[1] we adopt the BING (binarized normed gradients) operator [60] to produce a succinct set of object patches. To mimic human visual perception, an active learning algorithm is proposed to discover the visually/semantically representative object patches within each aerial photo. Thereby, the so-called GSP is constructed by sequentially linking these discovered objected patches. Subsequently, we design an aggregation-based deep neural network that statistically fuses the deep features of the GSP's internal object patches. To refine the learned deep GSP feature, a cross-domain knowledge transferal algorithm is proposed to utilize the low-resolution counterparts to enhance its descriptiveness. Finally, by leveraging the refined deep GSP representation, a multilabel SVM is learned to categorize each high-resolution aerial photo into multiple categories. Extensive experimental evaluations have demonstrated the superiority of our categorization model, as well as the descriptiveness of the actively learned GSP. In our work, features from the low-resolution aerial image and high-resolution aerial image represent two domains, respectively. Previously, only single low-resolution aerial image or high-resolution aerial image is employed for aerial image categorization. Apparently, this strategy is suboptimal if the low-resolution or high-resolution aerial image is blurred or noisy. When using two domains, the low-resolution and high-resolution aerial images are complementary to each other in describing different regions. In this way, better categorization performance can be achieved.

The key novelties of this work are three-fold as follows.

1) An aggregation-based deep network for GSP representation, which encodes multiple visually/semantically salient regions sequentially perceived by humans.
2) A knowledge transferal algorithm that leverages the spatial composition of low-resolution counterparts to enhance the deep learning of GSP feature from a high-resolution aerial photo.
3) A massive-scale multiresolution aerial photo set, based on which empirical evaluation is conducted to validate our approach.

The rest of this article is organized as follows. Section II briefly reviews the previous work related to ours. Section III introduces the three key modules in our framework: 1) active learning visually/semantically salient object patches, 2) an aggregation-based deep model that hierarchically derives the deep GSP features, and 3) a cross-resolution knowledge sharing algorithm to enhance deep GSP encoding. Comprehensive evaluations in Section IV demonstrated the effectiveness of our method. Finally, Section V concludes this article.

II. Related Work

Our method is closely related to two research fields in computer vision and remote sensing: 1) deep-learning-based object/scene categorization and 2) visual semantic models for aerial photos.

A. *Deep-Learning-Based Object/Scene Recognition*

Recently, a rich variety of deep architectures has been proposed for object/scene recognition. A few representative works are briefly introduced in the following. Multilayer CNNs with tailored architectures make building visual recognition models for million-level image set such as ImageNet [48] feasible.

[1] In our work, each high-resolution aerial photograph and its corresponding low-resolution counterparts capture the same region.

Krizhevsky *et al.* [1] proposed to learn large-scale CNNs using a subset of ImageNet [48], where impressive visual categorization performance has been achieved. Although the trained ImageNet-CNN focuses on generic object recognition, the engineered deep features can enhance many computer vision tasks, e.g., semantic annotation and human reidentification. Recently, the conventional ImageNet-CNN has been upgraded in two directions. The first direction handles the problem of generating high quality region samples from a rich set of images. Selective search [49] integrates the exhaustive search and semantic segmentation into a unified framework. A concise set of data-dependent and class-aware image regions are generated. For the other direction, Girshick *et al.* [50] proposed the well-known regions with CNN features (R-CNN). The core technique of R-CNN is a high quality image regions sampling strategy. Moreover, Zhou *et al.* [51] enhanced the CNN-based scene recognition by collecting qualified training samples. They collected a scene-centric image set consisting of seven million labeled scene images. In practice, it is suboptimal to train a CNN by leveraging each entire scene image or random image patches. Thereby, Wu *et al.* [57] developed a preprocessing strategy to enhance deep models for scene classification. It leverages a pretrained deep CNN by generating local and discriminative meta objects. He *et al.* [42] developed the ResNet, a residual learning algorithm to facilitate the training of very deep neural network compared to the standard ones. The network layers are formulated as learning a series of residual functions. Empirical results have demonstrated that highly competitive recognition performance has been observed on the ImageNet [48]. Further in [58], Wu *et al.* [42] introduced BlockDrop to enhance ResNet. It dynamically activates each deep network layer during inference, and, thereby the total computational cost is drastically reduced without performance loss.

It is worth emphasizing that, our method fundamentally differs from the deep recognition models in two aspects: 1) it naturally encodes human visual perception of high-resolution aerial photos by deeply learning the actively discovered GSPs and 2) it refines the deep GSP features of high-resolution aerial photos by exploring the knowledge of the low-resolution counterparts.

B. *Semantic Aerial Photograph Modeling*

A number of visual semantic models have been proposed to annotate aerial photos, either at image-level or at region-level. For semantic annotation at image-level, Zhang *et al.* [46] proposed a visual descriptor called graphlets to explicitly characterize aerial photo's geometry. And, thereby a discriminative model is trained for categorizing aerial photos into multiple classes. Xia *et al.* [47] formulated an aerial photo recognition framework by weakly supervised encoding region-level semantics. A multichannel hashing algorithm is unitized to fast calculate the image kernel for categorizing aerial photos. Akar *et al.* [2] employed the rotation forest and exploited object-level information for categorizing aerial photos. Experimental evaluations demonstrated its superiority toward competitors like Adaboost. Sameen *et al.* [16] proposed a deep CNN for recognizing aerial photo from high-resolution urban regions. It seamlessly encodes optical bands, digital surface, and ground-truth maps into the deep architecture. Cheng *et al.* [20] designed a pretuned deep CNN for high-resolution aerial photo categorization. The model is fine-tuned by a domain-specific scenery set. Moreover, Yao *et al.* [45] used CNN for semantic classification of aerial images. They proposed to semantically label pixels of urban regions by designing a multiresolution CNN to learn spatial–spectral features.

Many deep/shallow models have been proposed for detecting different types of region-level semantics. Fu. *et al.* proposed a fine-grained aircraft localization algorithm based on high-resolution aerial photos. The method employs a multiclass activation framework to discover the multiple parts within each aircraft. Wang *et al.* [44] formulated a multiresolution and end-to-end deep network for visual attention learning, associated with a classification and regression branch, for object detection in aerial photographs. Yang *et al.* [21] developed a double focal loss deep CNN for aerial-photo-based vehicle detection, where the skip connection is employed. Wang *et al.* [32] complied a waste plastic bottle set with 25407 aerial images. Correspondingly, they proposed to utilize unmanned aerial vehicles (UAV) to localize these plastic bottles. Costea *et al.* [3] proposed automatic geo-localization of aerial photos by identifying and matching of roads and intersection. Experiments using aerial photos from two European cities and OpenStreetMap-based roads annotations have shown its advantages.

Compared to the aforementioned techniques, our approach supports multilabel aerial photo annotation and the semantic categories can be flexibly defined. Thereby, our method is highly compatible with different circumstances. Moreover, as far as we know, only our method can explore cross-domain knowledge for multilabel semantic understanding, whereas our competitors only leverage one single aerial photo or multiview aerial photos from the same domain.

Cheng *et al.* [52] proposed a comprehensive review of the recent progress. They compiled a large-scale and public dataset, NWPU-RESISC45, for Remote Sensing Image Scene Classification. And several representative methods are evaluated using the proposed dataset. In [53], the authors proposed a simple but effective method to learn discriminative CNNs to enhance the performance of remote sensing image scene classification. The models are trained by optimizing a new discriminative objective function. Further Yao *et al.* [54] proposed a unified annotation framework for high-resolution aerial image modeling, They seamlessly combined the discriminative high-level feature learning and weakly supervised feature transferring.

III. Our Proposed Method

A. *Encoding Active Visual Perception*

BING [60] Object Patches: There are a rich number of fine-grained objects and their components inside each high-resolution aerial photo. Both biological and psychological studies have uncovered that humans are prone to attend a small proportion of visually/semantically salient objects during visual perception. In practice, before understanding each aerial photo, humans will first perceive objects, e.g., localizing an aircraft and its parts. Subsequently, they will attend to only a few prominent regions in detail, while the rest are kept almost unprocessed. Apparently, it is worthwhile to incorporate human visual perception during aerial photo categorization. In our categorization pipeline, a fast object proposals extraction associated with a geometry-preserved active learning algorithm is employed to select representative object patches for characterizing human gaze behavior during aerial photo perception.

Given a high-resolution detailed aerial photograph, humans usually attend to the semantically meaningful objects, e.g., rooftops and vehicles. These objects combined with their spatial compositions collaboratively determine the process of human perceiving each aerial photography. To localize objects that potentially draw human attention, we employ a state-of-the-art objectness measure to generate a concise set of object patches. In our work, we employ the BING [60] operator as the objectness measure due to its inherent competitiveness: 1) receiving a satisfactory object detection performance while keeping an extremely low calculation time; 2) generating a set of highly representative and low redundant object patches to enhance the deep encoding of human gaze behavior; and 3) having a high generalization capability to unknown aerial photo categories, and, thereby the categorization model is adaptable across different image sets.

Geometry-preserved Active Learning: There are still lots of object patches ($10^2 \sim 10^4$) extracted by the BING [60]. In practice, however, humans actively attend to fewer than 15 objects within each scenery. To characterize such active visual perception, a novel active learning is presented to discover K object patches from each aerial photo for GSP construction. It seamlessly integrates two representative attributes: 1) aerial photo's spatial configurations and 2) descriptiveness of the object patches at semantic-level.

Generally, a well-designed aerial photo categorization model should capture aerial photo's spatial compositions, i.e., the relative position between the foreground and background objects. To quantify such attributes, we believe that each object patch can be approximated by a linear combination of its neighboring object patches. During reconstruction, the contribution of each object patch is determined by the objective function as follows:

$$\arg\min_{\mathbf{E}} \sum_{i=1}^{N} ||z_i - \sum_{j=1}^{N} \mathbf{E}_{ij} z_j||$$
$$\text{s.t.} \sum_{j=1}^{N} \mathbf{E}_{ij} = 1, \mathbf{E}_{ij} = 0, \quad \text{if } z_i \notin \mathcal{N}(z_j) \tag{1}$$

where $\{z_1, \ldots, z_N\} \in \mathbb{R}^{N \times F}$ is the deep representations derived from the N BING [60] object patches in each aerial photo, F is the the deep feature dimensionality to each object patch, matrix $\mathbf{E}_{ij}$ quantifies the contribution of the ith object patch to reconstruct the jth one, and $\mathcal{N}(z_i)$ comprises of the spatial neighbors of object patch z_i.

Besides spatially encoding each aerial photo, the semantic descriptiveness of the selected object patches that constructing GSP is another important attribute. Based on the reconstruction error in (1), we denote $\{g_1, \ldots, g_N\}$ as the reconstructed object patches. Afterward, we identify the K selected object patches by optimizing the following objective function:

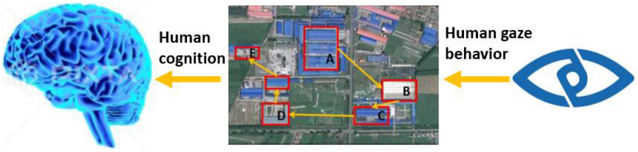


Fig. 3. Humans Sequentially perceive five semantic objects within a high-resolution aerial photo (indicated by GSP: $A \to B \to C \to D \to E$).

$$\eta(g_1, \ldots, g_N)$$
$$= \sum_{i=1}^{K} ||g_{q_i} - g_{q_i}||^2 + \tau \sum_{i=1}^{N} ||g_i - \sum_{j=1}^{N} \mathbf{E}_{ij} g_j||^2 \tag{2}$$

where τ is the weight of the regularizer, and $\{g_{q_1}, \ldots, g_{q_K}\}$ denotes the K actively selected object patches. Specifically, the first term optimizes the cost of fixing the coordinates of the K selected object patches. The second term enforces that the reconstructed object patches are maximally similar to the original ones semantically. In general, optimizing (2) can produce a set of semantically descriptive object patches, which can reflect human visual/semantic perception toward different aerial photos.

Let matrix $\mathbf{Z} = [z_1, \ldots, z_N]$ and $\mathbf{G} = [g_1, \ldots, g_N]$, and denoting $\mathbf{\Delta}$ as an $N \times N$ diagonal matrix indicating the selected object patches, i.e., diagonal element $\mathbf{\Delta}_{ii} = 1$ if $i \in \{q_1, \ldots, q_K\}$ and 0 otherwise. Based on these, the cost function (2) can be upgraded into the matrix form

$$\eta(\mathbf{Q}) = \text{tr}((\mathbf{G} - \mathbf{Z})^T \Delta(\mathbf{G} - \mathbf{Z})) + \tau\text{tr}(\mathbf{G}^T \mathbf{T} \mathbf{G}) \tag{3}$$

where matrix $\mathbf{T} = (\mathbf{I} - \mathbf{E})^T (\mathbf{I} - \mathbf{E})$. To minimize (3), the gradient of $\eta(\mathbf{G})$ is set to zero and we obtain

$$\mathbf{\Delta}(\mathbf{G} - \mathbf{Z}) + \tau \mathbf{T} \mathbf{G} = 0. \tag{4}$$

Then, the reconstructed object patches are calculated by

$$\mathbf{G} = (\tau \mathbf{T} + \mathbf{\Delta})^{-1} \mathbf{\Delta} \mathbf{Z}. \tag{5}$$

Based on the reconstructed object patches, the new reconstruction error is updated into

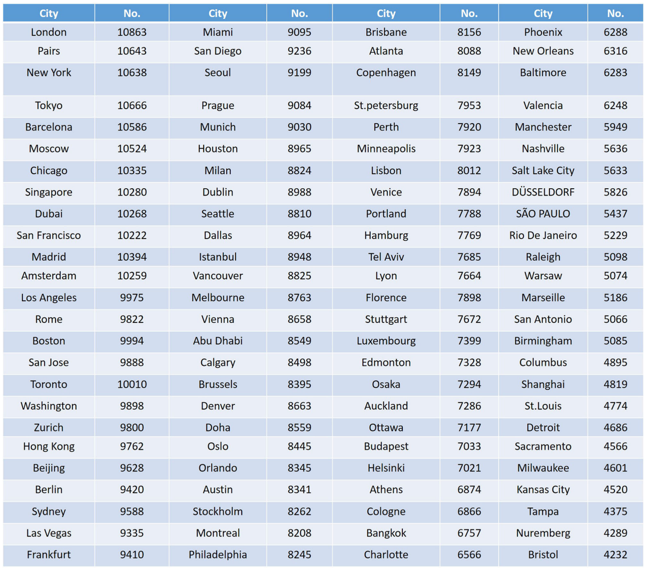
$$\eta(z_{q_1}, \ldots, z_{q_K}) = ||\mathbf{Z} - \mathbf{G}||_F^2 = ||\mathbf{Z} - (\tau \mathbf{T} + \mathbf{\Delta})^{-1} \mathbf{\Delta} \mathbf{Z}||_F^2$$
$$= ||(\tau \mathbf{T} + \mathbf{\Delta})^{-1} \tau \mathbf{T} \mathbf{Z}||_F^2 \tag{6}$$

where $|| \cdot ||_F^2$ is the matrix Frobenius norm.

Following (6), each GSP can be constructed by sequentially linking the actively discovered K object patches inside each aerial photograph, as exemplified in Fig. 3. More specifically, suppose k object patches have been determined, the $(k+1)$th object patch is selected by

$$q_{k+1} = \arg \min_{i \notin \{q_1, \ldots, q_k\}} ||(\tau \mathbf{T} + \mathbf{\Delta}_k + \mathbf{\Upsilon}_i)^{-1} \tau \mathbf{T} \mathbf{Z}||_F^2 \tag{7}$$

where the jth diagonal element of matrix $\mathbf{\Delta}_k \in \mathbb{R}^{N \times N}$ is one if z_j has been selected in the kth iteration and zero otherwise, and $\mathbf{\Upsilon}_i \in \mathbb{R}^{N \times N}$ is a diagonal matrix whose ith diagonal entity is one, while the rest are zeros.



City	No.	City	No.	City	No.	City	No.
London	10863	Miami	9095	Brisbane	8156	Phoenix	6288
Pairs	10643	San Diego	9236	Atlanta	8088	New Orleans	6316
New York	10638	Seoul	9199	Copenhagen	8149	Baltimore	6283
Tokyo	10666	Prague	9084	St.petersburg	7953	Valencia	6248
Barcelona	10586	Munich	9030	Perth	7920	Manchester	5949
Moscow	10524	Houston	8965	Minneapolis	7923	Nashville	5636
Chicago	10335	Milan	8824	Lisbon	8012	Salt Lake City	5633
Singapore	10280	Dublin	8988	Venice	7894	DÜSSELDORF	5826
Dubai	10268	Seattle	8810	Portland	7788	SÃO PAULO	5437
San Francisco	10222	Dallas	8964	Hamburg	7769	Rio De Janeiro	5229
Madrid	10394	Istanbul	8948	Tel Aviv	7685	Raleigh	5098
Amsterdam	10259	Vancouver	8825	Lyon	7664	Warsaw	5074
Los Angeles	9975	Melbourne	8763	Florence	7898	Marseille	5186
Rome	9822	Vienna	8658	Stuttgart	7672	San Antonio	5066
Boston	9994	Abu Dhabi	8549	Luxembourg	7399	Birmingham	5085
San Jose	9888	Calgary	8498	Edmonton	7328	Columbus	4895
Toronto	10010	Brussels	8395	Osaka	7294	Shanghai	4819
Washington	9898	Denver	8663	Auckland	7286	St.Louis	4774
Zurich	9800	Doha	8559	Ottawa	7177	Detroit	4686
Hong Kong	9762	Oslo	8445	Budapest	7033	Sacramento	4566
Beijing	9628	Orlando	8345	Helsinki	7021	Milwaukee	4601
Berlin	9420	Austin	8341	Athens	6874	Kansas City	4520
Sydney	9588	Stockholm	8262	Cologne	6866	Tampa	4375
Las Vegas	9335	Montreal	8208	Bangkok	6757	Nuremberg	4289
Frankfurt	9410	Philadelphia	8245	Charlotte	6566	Bristol	4232

Fig. 4. Our proposed deep model for representing a GSP sequentially connecting five object patches.

B. Deeply Learned GSP Representation

By actively discovering the GSP from a high-resolution aerial photo, we propose a scale-invariant network to calculate the deep GSP representation hierarchically. The pipeline of our designed deep model is shown in Fig. 4. In brief, the proposed deep architecture is comprised of two modules. First, a so-called adaptive spatial pooling (ASP) is deployed to characterize object patches with multiple scales. Afterward, the patch-level deep representations are concatenated into the image-level one through a statistical combination operation.

Module 1: Generally speaking, maintaining the original image resolution and aspect ratio is significant to represent aerial photo's spatial configurations [63]. Practically, distinguished from the fixed-sized image patches, supporting object patches with varied scales are more descriptive to semantic objects/components [62]. Aiming at this, the standard five-layer CNN [61] is upgraded to support differently sized input patches.

The our proposed deep network (Fig. 4) is detailed as follows. Given a set of salient object patches sequentially linked by each GSP, we flip/rotate each randomly to increase the sample number. The entire deep architecture involves four operations: convolution, ASP, and local response normalization, associated with a fully connected layer with 256 latent units. Afterward, the deep network branches out a fully connected layer with R units. Each unit corresponds to the R latent aerial-photos-relevant topics. Noticeably, R is a parameter depending on a particular dataset. Herein, we set R to the number of aerial photo categories.

Module 2: Given a high-resolution GSP that is constituted by a set of sequentially linked object patches with multiple scales, we extract the A-dimensional deep feature for each object patch using the aforementioned patch-level deep CNN. Then, we statistically concatenate these patch-level deep features into the deep representation characterizing each GSP.

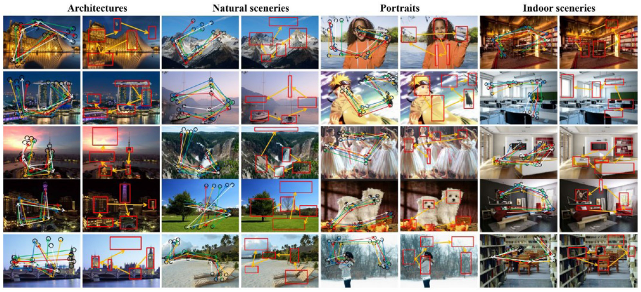


Fig. 5. Using the global composition from low-resolution aerial photo to enhance the high-resolution deep GSP feature learning, where the occluded regions can be largely recovered.

We denote $\Psi = \{\psi_i\}_{i\in[1,K]}$, where $\psi_i \in \mathbb{R}^F$ is the deep feature learned from each of the K object patches inside a GSP. Subsequently, we represent $\mathcal{T}_m$ as the set of values of the mth component of all $\psi_i \in \Psi$, i.e., $\mathcal{T}_m = \{\phi_{mj}\}_{j\in[1,K]}$. The statistic aggregation layer contains a set of statistic functions: $\mathcal{F} = \{\kappa_u\}_{u\in[1,4]}$. Each κ is a particular statistic operation for a set of patch-level deep features learned from the K CNNs (as shown in Fig. 4). For our proposed SA-Net, $\mathcal{F} = \{\min, \max, \text{mean}, \text{median}\}$. The outputs of $\mathcal{F}$ are concatenated and then aggregated using a fully connected layer to produce a S-dimensional vector as the deep representation for a GSP. Formally, these operations can be formulated as

$$f(\Psi) = \mathbf{W} \times (\oplus_{u=1}^{4} \oplus_{m=1}^{F} \kappa_u(\mathcal{T}_m)). \qquad (8)$$

where $\mathbf{W} \in \mathbb{R}^{S\times 4F}$ denotes the matrix parameterizing the fully connected aggregation layer, and $\oplus$ concatenates the $4F$ short vector $\kappa_u(\mathcal{T}_m)$ into a long one.

In our method, we discriminatively pretrain the CNN model toward different object patches. The CNN is pretrained by leveraging the large-scale auxiliary dataset (ILSVRC2012 [56]) with image-level annotations only. The pretraining was performed using the open source Caffe CNN library [55]. In our implementation, we carefully tuned the inherent parameters of the pretrained CNN (i.e., the number of layers is adjusted from 4 to 8, the kernel size is tuned from 64 × 64 to 512 × 512, and the output dimensionality of each layer is adjusted from 128 to 1024) until the best performance is reached.

C. Cross-Resolution Deep GSP Feature Enhancement

As aforementioned, the deep high-resolution GSP representation describes how humans perceive visually/semantically salient objects, which locally characterize each aerial photo. Meanwhile, the global spatial composition from one or multiple low-resolution counterparts are also contributive to human semantic perception, especially when the high-resolution aerial photo is blurred/occluded. Taking Fig. 5 as an example, many areas in the high-resolution aerial photograph are occluded by clouds, making the semantic categorization task difficult. But these areas are clear and nonoccluded in the low-resolution counterparts. Based on this observation, it is necessary to transfer the knowledge from low-resolution aerial photos to the high-resolution one to enhance visual semantic categorization. In our work, a domain-transfer paradigm is proposed to leverage the global spatial compositions from low-resolution aerial photos to improve the high-resolution deep GSP feature.

Domain-Transfer-Based SVM: To complementarily optimize the low-resolution global compositional feature and high-resolution deep GSP feature, a common subspace that bridges the high-resolution feature and the low-resolution one is constructed.[2] Specifically, a D_c-dimensional common subspace is built, where the low-resolution feature f^l and high-resolution one f^h can be projected onto it by transformation matrices $\mathbf{U} \in \mathbb{R}^{D_c \times D_l}$ and $\mathbf{V} \in \mathbb{R}^{D_c \times D_h}$, respectively. Inspired by the impressive performance of feature augmentation technique [19] in fusing heterogeneous features, we incorporate the original high/low-resolution features (f^l and f^h) and subsequently augment them by two augmented feature mapping functions, i.e.,

$$\theta(f^l) = [\mathbf{U}f^l, f^l, \mathbf{0}_{D_l}] \tag{9}$$

$$\theta(f^h) = [\mathbf{V}f^h, f^h, \mathbf{0}_{D_h}]. \tag{10}$$

Based on the above augmented features, the low-resolution global compositional feature and the high-resolution deep GSP feature are readily comparable. Afterward, we incorporate the augmented feature into a multiclass SVM framework for aerial photo semantic categorization, where the standard SVM formulation with the hinge loss is adopted. Mathematically, a weight vector $\mathbf{g} = [\mathbf{g}_c^T, \mathbf{g}_l^T, \mathbf{g}_h^T]^T$ is defined for the augmented feature, where $\mathbf{g}_c$, $\mathbf{g}_l$, and $\mathbf{g}_h$ denote the weight vectors for the common subspace, the low-resolution global composition feature, and the high-resolution deep GSP feature, respectively. Based on these, the optimal transformation matrices $\mathbf{U}$ and $\mathbf{V}$ as well as the weight vector $\mathbf{g}$ are calculated by minimizing the SVM structural risk. Formally, the above formulation can be represented as

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \min_{\mathbf{g},b,\rho_i^l,\rho_k^h} \frac{1}{2}||\mathbf{g}||^2 + H\left(\sum_{i=1}^{M_l}\rho_i^l + \sum_{i=1}^{M_h}\rho_i^h\right) \\ &\text{s.t. } l_i^l(\mathbf{g}^T\pi_l(f_i^l) + b) \geq 1 - \rho_i^l, l_i^h(\mathbf{g}^T\pi_h(f_i^h) + b) \geq 1 - \rho_i^h \\ &\rho_i^h \geq 0, ||\mathbf{U}||_F^2 \leq o_l, ||\mathbf{V}||_F^2 \leq o_h \end{aligned} \tag{11}$$

where $H > 0$ is a parameter balancing the tradeoff between the model complexity and the empirical losses of training low/high-resolution aerial photographs; o_l and o_h are two prespecified positive numbers that control the complexity of matrices $\mathbf{U}$ and $\mathbf{V}$, respectively.

To optimize (11), the dual form of its inner optimization is derived as follows. Mathematically, the dual variables ω_i^l and ω_i^h are introduced. Then, we set the derivatives of the Lagrangian of (11) with respect to $\mathbf{g}, b, \rho_i^l, \rho_k^h$ to zeros. Accordingly, we obtain the KKT conditions as: $\mathbf{g} = \sum_{i=1}^{M_l} \omega_i^l l_i^l \pi_l(f_i^l) + \sum_{i=1}^{M_h} \omega_i^h l_i^h \pi_h(f_i^h)$, $\sum_{i=1}^{M_l} \omega_i^l l_i^l + \sum_{i=1}^{M_h} \omega_i^h l_i^h = 0$, and $0 \leq \omega_i^l \leq \omega_i^h \leq H$. Then, the dual problem is given as

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \max_{\boldsymbol{\omega}} \mathbf{1}^T\boldsymbol{\omega} - \frac{1}{2}(\boldsymbol{\omega} \circ \mathbf{y})^T \mathbf{K}_{\mathbf{U},\mathbf{V}}(\boldsymbol{\omega} \circ \mathbf{y}) \\ &\text{s.t. } \mathbf{1}^T\boldsymbol{\omega} = 0, \mathbf{0} \leq \boldsymbol{\omega} \leq H\mathbf{1}, ||\mathbf{U}||_F^2 \leq o_l, ||\mathbf{V}||_F^2 \leq o_h \end{aligned} \tag{12}$$

where vector $\boldsymbol{\omega} = [\omega_1^l, \ldots, \omega_{M_l}^l, \omega_1^h, \ldots, \omega_{M_h}^l]^T \in \mathbb{R}^{M_l + M_h}$ contains the dual variables; $\mathbf{y} = [\mathbf{y}_l^T, \mathbf{y}_h^T]^T \in \{1, \ldots, L\}^{M_l + M_h}$ is the training samples' category labels (L denotes the number of aerial photo's categories); $\mathbf{y}_l$ and $\mathbf{y}_h$ denote the category labels from the low-resolution and high-resolution aerial photos, respectively; $\mathbf{K}_{\mathbf{U},\mathbf{V}} = \begin{bmatrix} \mathbf{F}_l^T(\mathbf{I}_{D_l} + \mathbf{U}^T\mathbf{U}) & \mathbf{F}_l^T\mathbf{U}^T\mathbf{V}\mathbf{F}_h \\ \mathbf{F}_h^T\mathbf{V}^T\mathbf{U}\mathbf{F}_l & \mathbf{F}_h^T(\mathbf{I}_{D_h} + \mathbf{V}^T\mathbf{V}) \end{bmatrix}$ is the $(M_l + M_h) \times (M_l + M_h)$ kernel matrix characterizing both the low- and high-resolution aerial photos; and $\mathbf{F}_l = [f_l^1, \ldots, f_{M_l}^1] \in \mathbb{R}^{D_l \times M_l}$ and $\mathbf{F}_h = [f_h^1, \ldots, f_{M_h}^h] \in \mathbb{R}^{D_h \times M_h}$ are feature matrices for the low- and high-resolution aerial photos, respectively.

Optimizing objective function (12) needs the dimensionality of the common space D_c to be specified, which might be infeasible in practice. It is observable that in the kernel matrix $\mathbf{K}_{\mathbf{U},\mathbf{V}}$, the transformation matrices $\mathbf{U}$ and $\mathbf{V}$ typically appears in the form of $\mathbf{U}^T\mathbf{U}$, $\mathbf{U}^T\mathbf{V}$, $\mathbf{V}^T\mathbf{U}$, and $\mathbf{V}^T\mathbf{U}$. We then define a semidefinite intermediate matrix $\mathbf{R} = [\mathbf{U}, \mathbf{V}]^T[\mathbf{U}, \mathbf{V}]$. Thereby, the common subspace becomes latent since we do not have to solve matrices $\mathbf{U}$ and $\mathbf{V}$ explicitly.

According to the definition of matrix $\mathbf{R}$, the objective function (12) can be upgraded into

$$\begin{aligned} &\min_{\mathbf{R} \succeq 0} \max_{\boldsymbol{\omega}} \mathbf{1}^T\boldsymbol{\omega} - \frac{1}{2}(\boldsymbol{\omega} \circ \mathbf{y})^T \mathbf{K}_{\mathbf{R}}(\boldsymbol{\omega} \circ \mathbf{y}) \\ &\text{s.t. } \mathbf{1}^T\boldsymbol{\omega} = 0, \mathbf{0} \leq \boldsymbol{\omega} \leq H\mathbf{1}, \text{trace}(\mathbf{R}) \leq o \end{aligned} \tag{13}$$

where $\mathbf{K}_{\mathbf{R}} = \mathbf{F}^T(\mathbf{R} + \mathbf{I})\mathbf{F}$, matrix $\mathbf{F} = \begin{bmatrix} \mathbf{F}_l & \mathbf{O}_{D_l \times M_h} \\ \mathbf{O}_{D_h \times M_l} & \mathbf{F}_h \end{bmatrix} \in \mathbb{R}^{(D_s + D_t) \times (M_l + M_h)}$, and $o = o_l + o_h$. To solve the objective function (13), we employ the alternating optimization [41]. That is, an SVM problem is solved with respect to ω, followed by a semidefinite programming (SDP) problem with respect to $\mathbf{R}$.

Based on the learned SVM hyperplane $\boldsymbol{\omega}$ and the intermediate matrix $\mathbf{R}$, we can categorize each aerial photo into the corresponding multiple semantic categories. By summarizing this section, the pipeline of our cross-resolution deep aerial photo categorization is shown in Algorithm 1.

IV. Experimental Evaluation

In this section, we validate the performance of our designed perception-aware cross-resolution deep aerial photograph categorization using three experiments. We first detail the massive-scale cross-resolution aerial photo set we compiled, based on which we compare our method with well-known deep/shallow scene/object categorization models. Subsequently, we evaluate three key modules in our cross-resolution categorization

[2] Without loss of generality, we assume that each high-resolution aerial photo is associated with one low-resolution counterpart. When more than one low-resolution counterparts are employed, we can combine them using a standard multiview learning algorithm [43] and further use the combined feature for domain transferal.

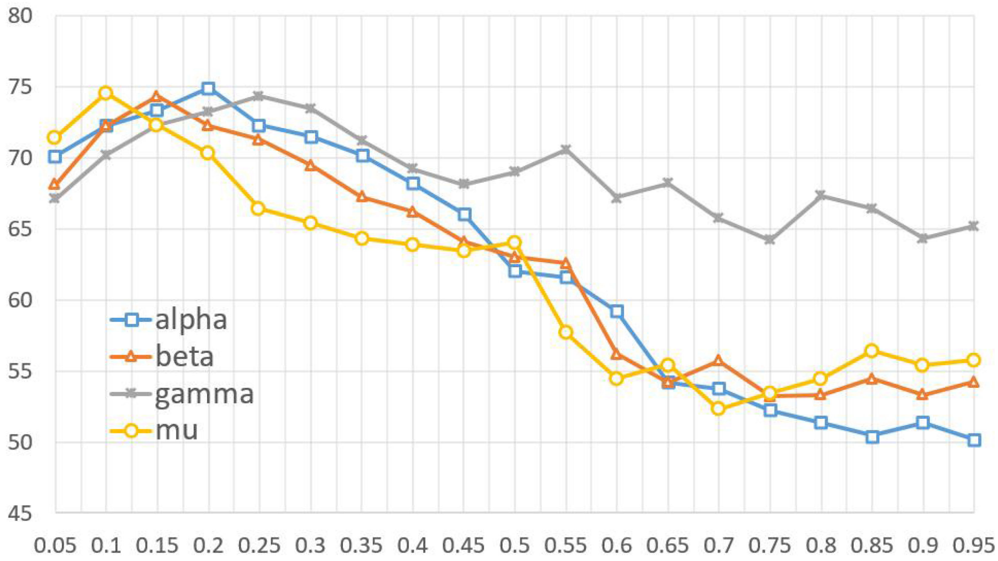


Fig. 6. Number of high-resolution aerial photos crawled from each of the top 100 metropolises.

Algorithm 1: Perception-Aware Cross-Resolution Deep Aerial Photo Categorization .

```
input: Massive-scale high-resolution aerial photos, each
  associated with one/multiple low-resolution
  counterparts; Training aerial photos' semantic labels;
output: Predicted semantic label of a test aerial photo;
1) Use BING [60] to extract many object patches from
  each high-resolution aerial photo, and then adopt
  geometry-based active learning to construct the GSP;
2) Adopt the aggregation-based deep network to learn the
  deep GSP representation;
3) Use the domain-transfer SVM to fuse low-resolution
  global composition feature and high-resolution deep
  GSP feature for aerial photo categorization.
```

pipeline: 1) active learning for GSP extraction, 2) aggregation-based deep GSP learning, and 3) domain-transfer-based deep GSP feature enhancement. Finally, we evaluate the influences of different parameters.

The experiments were carried out on a PC workstation equipped with dual Intel E5-2630 CPUs, 128 GB RAM, and a Tesla K40 GPU. All the compared categorization models were realized using C++.

A. Dataset

As far as we know, there is no public million-scale cross-resolution aerial photo set currently. In our work, we spent lots of human resources to collect about 2 M aerial photos from the Google Earth.[3] As Google Earth photos from metropolises are generally more clearer and detailed, a crawler software is designed to download and crop million-scale aerial photos from the 100 famous cities throughout the world, as reported in Fig. 6. Typical resolutions of these aerial photos are between 100×100 and 1200×1200. Then, we consider an aerial photo whose resolution is smaller than 400×400 as low resolution, while that larger than 800×800 as high resolution. In this way, we collect 0.83 M high-resolution aerial photos and 1.35 M low-resolution ones.

TABLE I
15 REPRESENTATIVE SEMANTIC CATEGORIES AND THE CORRESPONDING NUMBER OF HIGH-RESOLUTION AERIAL PHOTOS

Tall building	105 342	Residential	121 768	Intersection	87 843
Forest	85 874	Sea	131 436	Soccer field	107 435
Aircraft	97 685	Railway	118 453	Bridge	98 562
Road	86 894	River	132 692	Park	103 321
Palace	98 676	Factory	121 009	Farmland	112 324

Distinguished from the crawler that is conducted automatically, each high-resolution aerial photos and its low-resolution counterparts are paired semiautomatically. In detail, we record the XY-coordinate of each crawled aerial photo. For each pair of low-/high-resolution aerial photos, if their XY-coordinates are sufficiently small, then we consider they are potentially pairable. Subsequently, we recruit 28 master/Phd students from the Computer Sciences Department of Zhejiang University for dataset compilation. They worked about ten hours every working day to double-check the potentially pairable low/high-resolution aerial photos. This double-checking took two weeks. Finally, we obtained 0.69-M high-resolution aerial photos, associated with 1.12-M low-resolution counterparts.

To semantically annotate each aerial photograph at image-level, we inspected the top 100 cities throughout the world and summarized the 15 most representative semantic categories, as shown in Table I. For our aerial photo set, each aerial photo is typically associated with one to four semantic categories. In our implementation, the semantic categories of each aerial photo are annotated semiautomatically. First of all, object detectors corresponding to the 15 semantic categories are learned, which are

[3] [Online]. Available: https://www.google.com/earth/

utilized to label the semantic categories inside each aerial photo. To further refine these semantic labels, the same 28 master/Phd students carefully double-checked each aerial photo's labels.

B. Comparative Study

1) Categorization Accuracy: In the first place, we compare our proposed perceptual cross-resolution deep recognition model with several well-known shallow categorization models: 1) fixed-length walk and tree kernels, abbreviated as FLWK and FLTK [33], respectively, 2) multiresolution histogram (MRH) [38], 3) standard spatial pyramid matching (SPM)-based image kernel associated with its three variants: LLC-SPM [34], SC-SPM [59], and OB-SPM [35], and 4) the super vector (SV) [36]-based image encoding and supervised image encoding algorithm (SSC) [37]. In our implementation, the setups of the aforementioned baseline algorithms are briefed as follows: the FLWK's and FLTK's lengths are tuned from two to ten. For the MRH, the adopted scene images are smoothed by the Gaussian kernel ($\sigma = 2$) calculated with 15 gray levels. For SPM and its variants, the entire training scene images were decomposed into 1.5 million SIFT points. Thereafter, an 800-sized codebook is learned via k-means.

Owing to the impressive performance of deep categorization models, our method is also compared with a set of deep scene/object categorization models: the ImageNet CNN [1], the R-CNN [50], the meta object CNN (M-CNN) [57], the deep mining CNN (DM-CNN) [40], spatial pyramid pooling CNN (SPP-CNN) [39], ClearnNet [4], deep transferable architecture (DTA) [5], discriminative filter bank (DFB) [6], and multilayer CNN-RNN (ML-CNN-RNN) [7]. Moreover, we compare our cross-resolution categorization model with three semantic scene categorization models; they are proposed by Mesnil *et al.* [8], Xiao *et al.* [9], and Cong *et al.* [10], respectively.

As can be seen from Tables II and III, a quantitative comparison is conducted among the above 22 deep/shallow categorization models. Each experiment is repeated 20 times and the standard derivations are reported accordingly. In total, we make the following observations.

1) Our cross-resolution deep categorization model outperforms the shallow ones significantly due to three reasons: a) SPM and its variants, and SV only depend on SIFT descriptors. And it is infeasible to integrate multichannel visual descriptors into these models, such as color moment [64]; b) the external object detectors in OB-SPM are custom-built by the generic object categories. They are less representative to the aerial photo's categories; 4) RHM encodes image rough structure approximately for each aerial photo, and thereby it performs the worst; and d) due to the inherent totter phenomenon for graphical-model-based descriptors, the walk/tree kernel is less descriptive for describing aerial photographs.
2) Our approach performs competitively to the compared deep models in aerial photo categories categorization. The reasons are twofold as follows: a) based on the domain-transfer technique, our categorization model can enhance the learned deep high-resolution GSP feature by leveraging the global feature of its low-resolution counterparts. The inherent correlations between low-/high-resolution aerial photos are modeled explicitly. Comparatively, the other deep categorization models simply engineer the deep feature from the combined low-/high-resolution aerial photos, where the correlations are undiscovered; and b) our method can precisely capture human gaze allocation, which is informative to mimic human perceiving each aerial photo. In contrast, the other methods cannot encode such informative feature. We notice that for the rest of the deep categorization models, the traditional ImageNet-CNN [48] performs the worst. This is because only global image layout is encoded, and those fine-grained region-level details are neglected.
3) Although the internal semantic objects are well discovered, Mesnil *et al.* [8]'s and Cong *et al.* [10]'s methods cannot challenge our method. Moreover, they cannot support feature refinement by incorporating other domains. For Xiao *et al.* [9]'s method, it underperforms our approach since only low-level visual descriptors are exploited during scene modeling, where high-level semantic features are ignored completely.

2) Time Consumption: The time cost is an important criteria that reflects the performance of each deep/shallow categorization model. In this experiment, we report the training and test time cost on our massive-scale aerial photo set. First, we report the training time consumed for each module as follows: 7 h 53 m (active GSP extraction), 5 h 12 m (deep aggregation network for GSP encoding), and 7 h 41 m (domain-transfer-based SVM). During testing, given a new aerial photo, the time consumed for each module is given as follows: 154 ms (active GSP extraction), 435 ms (deep aggregation network for GSP encoding), and 98 ms (domain-transfer-based SVM).

In addition, the time cost of our method and a set of deep/shallow categorization models are compared. We present the time cost (both training and test) of these approaches in Table IV. The following conclusions can be made. 1) The training time of the shallow categorization models is much shorter than those of deep models. But during testing, the deep models are carried out faster. For our method, each aerial photo needs only 0.871 s to predict the categories. 2) Compared to the other deep categorization models, our method is more efficient during both training and test. This is because deep models like CNN and R-CNN typically produce thousands of object patches during training/test stage. But our method produces a succinct set of linked object patches for deep model learning. 3) Although the cross-domain transfer mechanism consumes extra time beyond the deep model training/test, we believe that it is worthwhile since the categorization precision can be enhanced by 6.7% after incorporating global features of the low-resolution aerial photos.

C. Stepwise Model Justification

In our proposed deep cross-resolution categorization pipeline, there are three key modules: 1) BING [60]-based active GSP learning; 2) aggregation-based deep model for GSP encoding;

TABLE II
AVERAGE ACCURACIES OF THE ABOVE DEEP/SHALLOW CATEGORIZATION MODELS (EACH EXPERIMENT IS REPEATED 20 TIMES)

Category	FLWK	FLTK	MRH	SPM	LLC-SPM	SC-SPM	OB-SPM	SV
Tall building	0.537	0.513	0.571	0.614	0.637	0.593	0.673	0.682
Residential	0.635	0.539	0.593	0.624	0.631	0.601	0.657	0.661
Intersection	0.605	0.682	0.695	0.633	0.658	0.613	0.694	0.687
Forest	0.642	0.618	0.675	0.651	0.637	0.629	0.557	0.676
Sea	0.661	0.676	0.681	0.537	0.642	0.683	0.591	0.671
Soccer field	0.583	0.617	0.703	0.626	0.639	0.648	0.674	0.684
Aircraft	0.597	0.664	0.613	0.681	0.694	0.706	0.688	0.693
Railway	0.667	0.621	0.652	0.714	0.564	0.559	0.635	0.701
Bridge	0.538	0.597	0.601	0.615	0.634	0.616	0.651	0.695
Road	0.543	0.606	0.652	0.643	0.657	0.661	0.634	0.636
River	0.653	0.627	0.685	0.713	0.692	0.658	0.647	0.668
Park	0.632	0.659	0.664	0.652	0.708	0.692	0.673	0.691
Palace	0.681	0.584	0.592	0.568	0.613	0.657	0.682	0.672
Factory	0.673	0.698	0.639	0.734	0.716	0.671	0.594	0.683
Farmland	0.605	0.681	0.637	0.684	0.598	0.664	0.684	0.667
Average	0.617	0.625	0.644	0.646	0.648	0.643	0.649	0.678
Category	SSC	ImageNet-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	CleanNet	DTA
Tall building	0.692	0.679	0.681	0.705	0.716	0.681	0.673	0.684
Residential	0.638	0.725	0.673	0.659	0.681	0.657	0.667	0.713
Intersection	0.657	0.682	0.674	0.665	0.637	0.628	0.634	0.691
Forest	0.643	0.695	0.634	0.656	0.703	0.657	0.672	0.685
Sea	0.694	0.687	0.672	0.651	0.692	0.659	0.697	0.699
Soccer field	0.726	0.751	0.697	0.731	0.726	0.692	0.716	0.706
Aircraft	0.689	0.704	0.721	0.708	0.715	0.724	0.695	0.711
Railway	0.703	0.697	0.714	0.721	0.657	0.637	0.701	0.683
Bridge	0.693	0.656	0.708	0.694	0.724	0.713	0.645	0.694
Road	0.659	0.711	0.694	0.682	0.706	0.715	0.692	0.687
River	0.689	0.671	0.721	0.684	0.692	0.687	0.704	0.667
Park	0.693	0.677	0.726	0.638	0.697	0.712	0.677	0.715
Palace	0.692	0.683	0.669	0.694	0.709	0.724	0.731	0.707
Factory	0.655	0.687	0.674	0.661	0.692	0.716	0.722	0.716
Farmland	0.683	0.714	0.726	0.699	0.707	0.675	0.691	0.692
Average	0.680	0.695	0.692	0.683	0.697	0.685	0.688	0.697
Category	DFB	ML-CNN-RNN	Mesnil *et al.*	Xiao *et al.*	Cong *et al.*	Ours		
Tall building	0.711	0.713	0.698	0.721	0.744	**0.747**		
Residential	0.694	0.723	0.715	0.724	0.733	**0.736**		
Intersection	0.706	0.712	0.709	0.683	0.715	**0.739**		
Forest	0.691	0.678	0.703	0.699	0.716	**0.733**		
Sea	0.698	0.711	0.707	0.689	**0.726**	0.709		
Soccer field	0.688	0.701	0.705	0.692	0.725	**0.742**		
Aircraft	0.724	0.716	0.742	0.731	0.706	**0.735**		
Railway	0.705	0.727	0.724	0.713	0.722	**0.738**		
Bridge	0.742	**0.749**	0.717	0.721	0.703	0.714		
Road	0.668	0.702	0.725	0.733	0.712	**0.739**		
River	0.724	0.735	0.709	0.721	0.705	**0.737**		
Park	0.694	0.705	0.673	0.734	0.693	**0.744**		
Palace	0.722	0.743	0.705	0.711	0.706	**0.748**		
Factory	0.703	0.697	0.684	0.692	0.715	**0.736**		
Farmland	0.694	0.713	0.695	0.725	0.732	**0.741**		
Average	0.704	0.715	0.707	0.712	0.717	**0.736**		

The bold entities indicate the best result.

and 3) cross-domain-based SVM learning. We evaluate the effectiveness of each component to demonstrate their indispensability and inseparability. Specifically, we replace each of the three modules, while keeping the rest two unchanged. Based on this, we report the corresponding categorization performance decrement or increment.

Module 1: To evaluate the active GSP extraction, we first replace the BING-based object patches extraction by superpixels (S1) and randomly cropped patches (S2), respectively. We report the corresponding average categorization accuracy decrements in Table V. As shown, our adopted BING operator performs much better than superpixels and random patches because it is intrinsically descriptive to objects and their components. Besides, we compare our geometry-based active learning with three competitors i.e., unified active learning by He *et al.* [24](S3), local representation active learning by Hu *et al.* [25] (S4), and multilabel active learing by Wu *et al.* [29] (S5). As reported in Table V, Hu *et al.*'s active learning achieves the closest performance to our method, lagging behind only 1.231%. Meanwhile, He *et al.*'s and Wu *et al.*'s algorithms perform moderately worse than ours. Actually, the superiority of our active learning is because of: 1) the aerial photo's global geometry is optimally preserved and 2) the sequentially solution [as shown in (7)] can well reflect human gaze allocation.

Module 2: We then testify the performance of our aggregation-based deep model. Two settings are adopted: 1) replacing the deep GSP feature by the ImageNet CNN [48] that captures each aerial photo globally (S1); and 2) abandoning the statistic operator "min" (S2), "max" (S3), "mean" (S4), and "median" (S5), respectively. As shown in Table V, our deep GSP feature performs significantly better than the ImageNet-CNN feature. This is because the actively selected GSP can well capture human gaze behavior and its constituent object patches are highly representative to aerial photo's discriminative parts. Contrastively, ImageNet-CNN feature only encodes the entire aerial photo without indicating the discriminative object patches. Besides,

TABLE III
STANDARD DERIVATIONS OF THE ABOVE DEEP/SHALLOW CATEGORIZATION MODELS (EACH EXPERIMENT IS REPEATED 20 TIMES)

Category	FLWK	FLTK	MRH	SPM	LLC-SPM	SC-SPM	OB-SPM	SV
Tall building	0.011	0.012	0.010	0.013	0.013	0.013	0.013	0.011
Residential	0.012	0.008	0.011	0.008	0.012	0.014	0.011	0.010
Intersection	0.011	0.014	0.016	0.011	0.011	0.011	0.009	0.009
Forest	0.012	0.008	0.015	0.012	0.012	0.011	0.009	0.012
Sea	0.013	0.011	0.012	0.009	0.011	0.012	0.011	0.015
Soccer field	0.012	0.011	0.015	0.013	0.013	0.012	0.011	0.013
Aircraft	0.011	0.012	0.013	0.013	0.013	0.013	0.008	0.008
Railway	0.014	0.010	0.011	0.010	0.012	0.012	0.009	0.009
Bridge	0.013	0.013	0.013	0.011	0.011	0.011	0.012	0.012
Road	0.014	0.008	0.011	0.014	0.013	0.010	0.009	0.011
River	0.011	0.012	0.011	0.009	0.008	0.010	0.012	0.010
Park	0.012	0.011	0.013	0.012	0.011	0.012	0.012	0.009
Palace	0.011	0.008	0.010	0.012	0.015	0.010	0.009	0.012
Factory	0.010	0.010	0.011	0.009	0.010	0.011	0.011	0.011
Farmland	0.013	0.008	0.013	0.012	0.013	0.010	0.012	0.009
Category	SSC	ImageNet-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	CleanNet	DTA
Tall building	0.013	0.012	0.011	0.013	0.015	0.011	0.008	0.013
Residential	0.010	0.011	0.014	0.010	0.013	0.014	0.009	0.010
Intersection	0.010	0.013	0.012	0.011	0.009	0.011	0.010	0.012
Forest	0.012	0.009	0.010	0.015	0.013	0.011	0.009	0.008
Sea	0.008	0.012	0.013	0.009	0.011	0.013	0.008	0.011
Soccer field	0.012	0.010	0.013	0.012	0.014	0.010	0.010	0.013
Aircraft	0.011	0.008	0.011	0.013	0.015	0.015	0.009	0.012
Railway	0.008	0.012	0.013	0.010	0.012	0.014	0.008	0.010
Bridge	0.013	0.013	0.012	0.011	0.014	0.012	0.010	0.012
Road	0.011	0.009	0.010	0.013	0.011	0.011	0.009	0.011
River	0.008	0.008	0.015	0.011	0.012	0.013	0.012	0.013
Park	0.015	0.010	0.011	0.015	0.013	0.009	0.008	0.012
Palace	0.009	0.008	0.012	0.009	0.011	0.012	0.009	0.009
Factory	0.014	0.012	0.014	0.011	0.016	0.014	0.012	0.008
Farmland	0.009	0.011	0.013	0.014	0.012	0.011	0.014	0.011
Category	DFB	ML-CNN-RNN	Mesnil *et al.*	Xiao *et al.*	Cong *et al.*	Ours		
Tall building	0.011	0.012	0.011	0.013	0.012	0.011		
Residential	0.008	0.008	0.012	0.011	0.014	0.009		
Intersection	0.012	0.013	0.010	0.009	0.013	0.010		
Forest	0.011	0.008	0.015	0.014	0.013	0.008		
Sea	0.011	0.012	0.011	0.008	0.009	0.010		
Soccer field	0.012	0.008	0.015	0.013	0.014	0.009		
Aircraft	0.011	0.012	0.011	0.013	0.012	0.010		
Railway	0.010	0.010	0.009	0.010	0.013	0.009		
Bridge	0.008	0.014	0.011	0.011	0.015	0.010		
Road	0.011	0.009	0.009	0.012	0.012	0.011		
River	0.008	0.012	0.011	0.009	0.010	0.008		
Park	0.012	0.011	0.013	0.012	0.009	0.010		
Palace	0.011	0.009	0.012	0.012	0.013	0.008		
Factory	0.008	0.012	0.011	0.009	0.009	0.009		
Farmland	0.012	0.011	0.015	0.013	0.014	0.010		

TABLE IV
TRAINING/TEST TIME CONSUMPTION OF THE ABOVE SHALLOW/DEEP CATEGORIZATION MODELS (EACH EXPERIMENT IS REPEATED 20 TIMES)

	FLWK	FLTK	MRH	SPM	LLC-SPM	SC-SPM	OB-SPM	SV
Training	6 h 32 m	8 h 15 m	3 h 35 m	6 h 34 m	8 h 47 m	7 h 4 m	12 h 34 m	5 h 42 m
Test	3.5 s	7.6 s	1.56 s	3.4 s	4.2 s	4.1 s	6.6 s	3.7 s
	SSC	ImageNet-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	CleanNet	DTA
Training	7 h 23 m	94 h 21 m	121 h 43 m	165 h 32 m	89 h 43 m	75 h 32 m	84 h 43 m	91 h 43 m
Test	2.5 s	8.7 s	7.4 s	9.2 s	7.5 s	4.5 s	6.1 s	5.5 s
	DFB	ML-CNN-RNN	Mesnil *et al.*	Xiao *et al.*	Cong *et al.*	Ours		
Training	115 h 21 m	67 h 45 m	4 h 32 m	2 h 21 m	5 h 32 m	35 h 21 m		
Test	13.3 s	7.7 s	1.3 s	0.8 s	1.4 s	3.1 s		

TABLE V
PERFORMANCE DECREMENTS ("-")/INCREMENTS ("+") BY REPLACING EACH OF THE THREE KEY MODULES

	Module 1	Module 2	Module 3
S1	-3.541%	-10.546%	-4.530%
S2	-6.435%	-1.434%	-3.023%
S3	-2.434%	-3.113%	-6.652%
S4	-1.231%	-2.158%	-3.324%
S5	-3.143%	-1.376%	-2.543%

we notice that abandoning each statistic operator will hurt the aerial photo categorization performance, especially the "max" operator.

Module 3: Finally, we evaluate the effectiveness of our cross-domain SVM learning. We first abandon the low-resolution global composition channel and directly train the multilabel SVM by the deep high-resolution GSP feature (S1). As reported in Table V, this operation makes the categorization accuracy decrease by 9.323%. This demonstrates the necessity

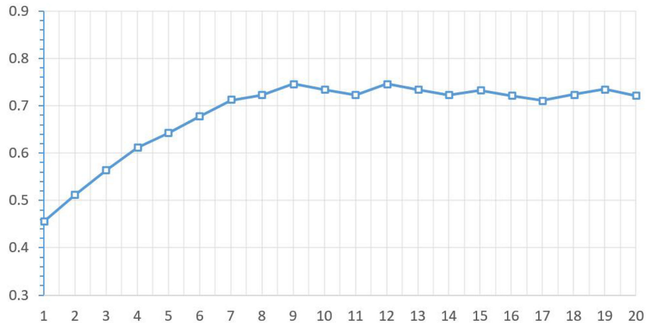


Fig. 7. Aerial photo's categorization accuracies by varying τ (left) and K (right).

TABLE VI
PERFORMANCE BY TUNING H

H	10	10^2	10^3
Acc	0.677	0.736	0.719%
H	10^4	10^5	10^6
Acc	0.675%	0.643	0.615%

of transferring the low-resolution feature to enhance deep GSP feature learning. Afterward, we compare our domain transfer algorithm with three competitors: 1) translated learning [11] (S2), 2) heterogenous transfer learning [12] (S3), and 3) heterogeneous domain adaption [13] (S4). As shown in Table V, all the competitors underperform our method by at least 2%.

D. Parameter Analysis

Totally, there are three key parameters to be adjusted for our cross-resolution deep categorization model: 1) τ, the regularizer weight encoding the aerial photo's geometry, 2) K, the number of object patches inside each GSP, and 3) H in the multilabel SVM formulation. Herein, we evaluate the influence of each parameter on the aerial photo categorization.

First, we tune τ from 0.05 to 0.5 with a step of 0.05 and report the average categorization accuracy. As shown on the left of Fig. 7, we notice that neither a too small τ nor a very large τ is an optimal choice. This observation reflects that we cannot neglect the global geometry or emphasize it aggressively. On our compiled aerial photo set, we notice that when $\tau = 0.25$, the best performance can be received.

Afterward, we adjust K from one to 19 with a step of two. Similarly, we report the average categorization accuracy. As displayed on the right of Fig. 7, the categorization accuracy increases significantly when K is tuned from one to nine, and subsequently remains stable. This observation indicates that $K = 9$ is sufficiently descriptive to our massive-scale aerial photo set. Since K determines the SubCNN number in our aggregation-based deep model, we set $K = 9$ in order to maintain the effectiveness and effectiveness of our categorization system. Each subCNN denotes the CNN trained toward the *i*th object patch inside the GSPs. For example, if there are five object patches inside each GSP, then there are five subCNNs. The *i*th subCNN is particularly trained using the *i*th object patches from all the training GSPs. Interestingly, we notice that, in our setting, the optimal K is larger than that on the nonaerial scenery set, such as the Scene-67 [14]. The reason lies in that, an aerial photo usually contains more foreground salient regions.

Last but not least, we evaluate the categorization performance by varying H. We adjust H from 10 to 10^6 with a factor of 10. As reported in Table VI, we notice that the best categorization accuracy is received when $H = 10^2$.

V. CONCLUSION

Categorizing aerial photographs is an important application in computer vision and remote sensing [28], [30], [31]. This article formulates a deep cross-resolution aerial photograph categorization framework. We leverage the low-resolution global compositional feature to enhance the deep learning of high-resolution GSP feature. By active learning GSP from many BING [60]-based object patches, an aggregation-based deep model is formulated to represent each GSP. Afterward, a cross-domain and multiclass SVM is derived by optimally combining low-level global compositional feature and high-resolution GSP feature. To evaluate our method, we compiled a million-scale aerial photo set. Comprehensive empirical results have demonstrated its effectiveness and efficiency.

The future work includes developing a semisupervised cross-domain SVM, wherein the low/high-resolution aerial photos are partially labeled. Moreover, we intend to release our compiled aerial photo set for public evaluation.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[2] O. Akar, "Mapping land use with using rotation forest algorithm from UAV images," *Geocarto Int.*, vol. 33, no. 5, pp. 538–553, 2017.

[3] N. Shervashidze, S. V. N. Vishwanathan, T. Petri, K. Mehlhorn, and K. M. Borgwardt, "Efficient graphlet kernels for large graph comparison," *AISTATS*, pp. 488–495, 2009.

[4] 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.*, 2018, pp. 5447–5456.

[5] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8697–8710.

[6] 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.*, 2018, pp. 4148–4157.

[7] 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. Vision. Workshops*, 2018, pp. 675–688.

[8] 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. Pattern Recognit. Appl. Methods*, 2015, pp. 209–224.

[9] Y. Xiao, J. Wu, and J. Yuan, "mCENTRIST: A multi-channel feature generation mechanism for scene categorization," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 823–836, Feb. 2014.

[10] Y. Cong, J. Liu, J. Yuan, and J. Luo, "Self-supervised online metric learning with low rank constraint for scene categorization," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3179–3191, Aug. 2013.

[11] W. Dai, Y. Chen, G.-R. Xue, Q. Yang, and Y. Yu, "Translated learning: Transfer learning across different feature spaces," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 678–694.

[12] Q. Yang, Y. Chen, G.-R. Xue, W. Dai, and Y. Yu, "Heterogeneous transfer learning for image clustering via the social web," in *Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Lang. Process.*, 2009, pp. 1–9.

[13] C. Wang and S. Mahadevan, “Heterogeneous domain adaptation using manifold alignment,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1–8.

[14] A. Quattoni and A. Torralba, “Recognizing indoor scenes,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 413–420.

[15] A. Krizhevsky, “Learning multiple layers of features from tiny images,” Univ. Toronto, Toronto, Canada, Tech. Rep., 2009.

[16] M. I. Sameen, B. Pradhan, and O. S. Aziz, “Classification of very high resolution aerial photos using spectral-spatial convolutional neural networks,” *J. Sensors*, 2009, Art. no. 7195432.

[17] H. Zhang, B. Li, J. Zhang, and F. Xu, “Aerial image series quality assessment,” *OP Conf. Ser.: Earth Environ. Sci.*, vol. 17, 2014, Art. no. 012183.

[18] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik, “Multiscale combinatorial grouping,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 328–335.

[19] H. Daume III, “Frustratingly easy domain adaptation,” in *Proc. Assoc. Comput. Linguistics*, 2007, pp. 256–263.

[20] 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.

[21] 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.*, 2018, pp. 3079–3083.

[22] 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.

[23] A. Vedaldi and K. Lenc, “MatConvNet-convolutional neural networks for MATLAB,” in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.

[24] X. He, M. Ji, and H. Bao, “A unified active and semi-supervised learning framework for image compression,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 65–72.

[25] Y. Hu, D. Zhang, Z. Jin, D. Cai, and X. He, “Active learning based on local representation,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1415–1421.

[26] H. You, S. Tian, and L. Yu, “Yalong lv: Pixel-level remote sensing image recognition based on bidirectional word vectors,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1281–1293, Feb. 2020.

[27] T. Gadhiya and A. K. Roy, “Superpixel-driven optimized wishart network for fast PolSAR image classification using global k-means algorithm,” *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 97–109, Jan. 2020.

[28] M. E. Paoletti, J. M. Haut, R. Fernndez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, “Deep pyramidal residual networks for spectral-spatial hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 740–754, Feb. 2019.

[29] K. Wu, D. Cai, and X. He, “Multi-label active learning based on submodular functions,” *Neurocomputing*, vol. 313, pp. 436–442, 2018.

[30] L. Pallotta, A. D. Maio, and D. Orlando, “A robust framework for covariance classification in heterogeneous polarimetric SAR images and its application to L-band data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 104–119, Jan. 2019.

[31] M. E. Paoletti, J. M. Haut, R. Fernndez-Beltran, J. Plaza, A. J. Plaza, and F. Pla, “Conditional random field and deep feature learning for hyperspectral image classification,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1612–1628, Mar. 2019.

[32] J. Wang, W. Guo, T. Pan, H. Yu, L. Duan, and W. Yang, “Bottle detection in the wild using low-altitude unmanned aerial vehicles,” in *Proc. IEEE 21st Int. Conf. Inf. Fusion*, 2018, pp. 439–444.

[33] Z. Harchaoui and F. R. Bach, “Image classification with segmentation graph kernels,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[34] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang, “Yihong gong, locality-constrained linear coding for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3360–3367.

[35] L.-J. Li, H. Su, E. P. Xing, and L. Fei-Fei, “Object bank: A. high-level image representation for scene classification and semantic feature sparsification,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1–9.

[36] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, “Image classification using super-vector coding of local image descriptors,” in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 141–154.

[37] J. Yang, K. Yu, and T. S. Huang, “Supervised translation-invariant sparse coding,” in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3517–3524.

[38] E. Hadjidemetriou, M. D. Grossberg, and S. K. Nayar, “Multiresolution histograms and their use for recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 7, pp. 831–847, Jul. 2004.

[39] 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.

[40] Y. Li, L. Liu, C. Shen, and A. van den Henge, “Mid-level deep pattern mining,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 971–980.

[41] L. Duan, D. Xu, and I. W. Tsang, “Learning with augmented features for heterogeneous domain adaptation,” in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 667–674.

[42] 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.

[43] S. Sun, “A survey of multi-view machine learning,” *Neural Comput. Appl.*, vol. 23, no. 7/8, pp. 2031–2038, 2016.

[44] 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.

[45] W. Yao, P. Poleswki, and P. Krzystek, “Classification of urban aerial data based on pixel labelling with deep convolutional neural networks and logistic regression,” *ISPRS - Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. XLI-B7, pp. 405–410, 2016.

[46] L. Zhang, Y. Han, Y. Yang, M. Song, S. Yan, and Q. Tian, “Discovering discriminative graphlets for aerial image categories recognition,” *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 5071–5084, Dec. 2013.

[47] 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.

[48] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[49] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, “Selective search for object recognition,” *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, 2013.

[50] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[51] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–9.

[52] G. Cheng, J. Han, and X. Lu, “Remote sensing image scene classification: Benchmark and state of the art,” *Proc. IEEE*, vol. 105, no. 10, pp. 1865–1883, 2017.

[53] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, “When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative CNNs,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2811–2821, May 2018.

[54] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, “Semantic annotation of high-resolution satellite images via weakly supervised learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[55] V. Lepetit, F. Moreno-Noguer, and P. Fua, “EPnP: An accurate on solution to the pnp problem,” *Int. J. Comput. Vis.*, vol. 81, no. 2, pp. 155–166, 2009.

[56] O. Russakovsky *et al.*, “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[57] R. Wu, B. Wang, W. Wang, and Y. Yu, “Harvesting discriminative meta objects with deep CNN features for scene classification,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1287–1295.

[58] Z. Wu *et al.*, “BlockDrop: Dynamic inference paths in residual networks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8817–8826.

[59] J. Yang, K. Yu, Y. Gong, and T. Huang, “Linear spatial pyramid matching using sparse coding for image classification,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1794–1801.

[60] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr1, “BING: Binarized normed gradients for objectness estimation at 300fps,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 3286–3293.

[61] N. Zhang, M. Paluri, M. A. Ranzato, T. Darrell, and L. D. Bourdev, “PANDA: Pose aligned networks for deep attribute modeling,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1637–1644.

[62] B. Cheng, B. Ni, S. Yan, and Q. Tian, “Learning to photograph,” in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 291–300.

[63] L. Mai, H. Jin, and F. Liu, “Composition-preserving deep photo aesthetics assessment,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 497–506.

[64] M. Stricker and M. Orengo, “Similarity of color images,” *Proc. SPIE*, vol. 2420, pp. 3179–3191, 1995.