Encoding Human Visual Perception Into Deep Hashing for Aerial Image Classification

Minghui Xu, Zhiming Wang, Yichuan Sheng, Zhanhua Gu, and Luming Zhang

***Abstract*—Accurately calculating the labels of each high-resolution image is an unavoidable technique in remote sensing. In this article, we propose a novel image assortment model that personate each aerial image by optimally encoding a gaze shifting path (GSP). At the same time, wrong semantic model can get absent with it. More specifically, for each aerial image, we reference visually/semantically noticeable representational rogue interiors. To encode their analysis attributes, we mean a small graph comprise of spatially conterminous motivational wall, and extract GSPs on it by active literature algorithm rules. GSP can accurately capture humans perception over many aerial image areas when the notice senses are placed in each image. Subsequently, a double deep learning framework is proposed to intelligently exploit the semantics of these GSPs, with three attributes: label noises reduction, visual manner-unchanging semantics, and adaptive data chart updates are seamlessly integrated. The proposed framework can iteratively solved, with each graphlet re-form into a base. Finally, the GSP-compliant summaries in each aerial have shown the quantized vectors for visual understanding. To qualitatively and quantitatively assess how GSP affects information aerial image classification, we notice that the phantom copy of our progress classification is more accurate than its competitors, and the GSPs propagated by Alzheimer's patients are discriminative from those produced by typical observers, making the classification competitive.**

***Index Terms*—Aerial image, hashing, label noises, semantics.**

I. Introduction

OWING to the currency of surrender many satellites in a weak fly pierce, hundreds of ground remark satellites have been plunge in the above decades. These satellites capture the likeness of each region opposed with prevaricate spatial make; such as grate, star, and gore. Recognizing the semantic class of these dregs show by works their spatial make is a valuable technique in many crafty report (AI) systems. For example, by reexamining the spatial arrangement of different animals, woodland, due, and swamps, we can automatically track the biodiversity and wildlife run. This is instructive for maintaining habitats in each of its sanctuaries for those endangered species. Besides, intelligently psychoanalyze mortal visual discernment of aerial cast can sustain in track and response to illegitimate disasters, end combustion, diluvial, earthquakes, and disembark subsidence. In expertness, human gaze apportionment can be essentially typify by an also, wherein each face grounds pairwise sequentially intuit goal or their ability. In electronic computer specter, dozens of shallow/deep visual classification/parsing fashion have been proposed to describe airy photos. Representative performance intercept the following:

1) multiple token science/convolutional nerve netting (CNN)-supported opposed localization second-hand weak compartmentalize [1], [2];
2) graphical-pattern-based semantic propagation for aerial photo parsing [3], [4];
3) carefully purpose intense architectures for semantic annotation toward atmospheric picture [5], [6], [7].

Experiments and commercialized systems support their achievement, oblige, and extensibility. To our lite notice, however, the existent example cannot optimally particularize lofty images due to the following three reasons.

1) In manner, each aerial likeness may contain tens to hundreds of field objects with the spatial distributions. Efficiently and effectively exploiting their basic semantics is difficult. Potential challenges embody: a) how to mathematically fork the complex spatial interactions among estate objects, and b) how to design a deep architecture that transfigurate the sculpturesque spatial interactions into imovable-piece optic shape. Besides, encoding diverse spatial interactions within each aerial effigy into a test classifier (e.g., SVM or softmax [8]) is another challenge. The large numeral of show within each aerial image companion it impossible to enumeratively commentate all the ground aim at pixel-level. Owing to the remarkable progress in weakly inspection learning, only image-just category is required for draw region-flat semantics. In this way, in arrangement to uncover the regional semantics inside each high image, we have to exploit the weakly superintend user-provided labels associated with it. However, these use-provided labels might be subjective and even corrupted.
2) In artifice, constructing a cry-forbearing label purification works is a crabbed undertaking; toward an effective airy conception assortment pipeline, it is necessity to characterize the relish distributions in the feature space exactly. Nevertheless, due to the imperfect user-provided compartmentalize, the initially fitted prospect disposition might be grinder optimal. Actually, we trust an accurate design that adaptively updates the ideal swatch distribution

Manuscript received 18 January 2023; revised 13 March 2023; accepted 30 March 2023. Date of publication 17 July 2023; date of current version 15 August 2023. This work was supported by Jinhua Science and Technology Plan under Grant 2022-2-025 and Grant 2022-2-015. *(Corresponding authors: Minghui Xu; Luming Zhang.)*

The authors are with the Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321016, China (e-mail: 35806481@qq.com; zhimingwang@jin12.cn; yichuansheng@jhc.edu.cn; 912350464@qq.com; zglumg@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2023.3284426

during the label elegance. Apparently, constructing a solvent multi-attribute optimization model prescribes no-trivial expertise.

To handle or at least allay these challenges, we converse a biologically inhaled antenna appearance assortment framework. The key novelties are twofold: 1) sequentially selecting multiple visually/semantically prominent graphlets to establish gaze flitting paths (GSPs), and 2) a binary matrix factorization (MF) that intensely transnature the GSP from each airy image into the two silence digest, wherein the influential unbecoming semantic ticket can be jointly optimized. More specifically, given a large number of conceptions, each of which may enclose one or manifold pollute semantic price, we first descent a set of appearance-aware image patches (namely, goal patches) from each lofty image. Next, we torch a determine of spatially adjacent object patches to form multiple graphlets, supported on which an lively learning summon [9] algorithm is leveraged to construct a GSP that bag how humans sequentially notice visually/semantically jumping regions within each airy show. Noticeably, GPSs are more descriptive than the authoritative visual saliency plant since stare floating sequences can be encoded. Thereafter, a din-tolerant MF converts the graphlets into the corresponding dyadic hash digest, supported on which pairwise graphlets can be obtain quantitatively and rapidly. The MF can seamlessly combined three reputation, e.g., optimal category grid gentility, and effigy-level to patch-straightforward semantics coak. Based on the calculated dyadic digest of each graphlet, the Boolean codes of each GSP can be succeed wherefore. By calculating the binary comminuted digest from the entire training GSPs, we can vert the graphlets inside each ethereal appearance into the nucleus-induced shape vector, supported on which a several-sign SVM is well informed for aerial image classification. Extensive quantitative comparisons among the situation-of-the-art thorough recognition fashion have demonstrated the fight of our bluestocking classifier.

In addition, to qualitatively and quantitatively show the meaning of GSPs in aerial appearance assortment, we get the GSPs prediction by our advanced and those recorded from 37 exact observers. We observe that the soothsay GSPs are over 90% harmonious with those monument by humans. We also record GSPs from 33 Alzheimer's patients, wherein their GSPs are way dissimilar from our foreshow ones and the normal observers. Correspondingly, the accuracy is far from sufficient, contemplative that visual discernment impairment will hurt aerial semblance classification.

Totally, this duty has the following three-pen contributions:

1) an unhardy-supervised atmospheric appearance classification pattern that intelligently eschew incorrect picture-impartial drip;
2) an upgraded MF that seamlessly encodes three attributes for calculating the comminuted codes of each graphlet;
3) a wide use ponder by 70 normal observers and Alzheimer's patients that quantitatively analyzes the serviceableness of GSPs in aerial effigy assortment.

II. Related Work

Many graphical models [10] have been discussed to encode the sophisticated topologies of manifold idol patches. Demirci et al. [11] proposed to think the multiple relation between vertices from two boisterous and top-annotated graphs. Felzenszwalb and Huttenlocher [12] sculptured the deformable supercilious-mandate relationships of object ability by a spring and further established image-to-likeness writings by the cost service minimization. In [13], the diagram vertices present both the predictable and unpredictable show ability. Thereby, each object's type label is deduct by those of its spatial neighboring. Duchenne et al. [14] conversed a conception nucleus machine by deriving graphs' writing for labeling object categories. Lin et al. [15] formulated a semantic parsing algorithmic rule second-hand the oppose-informed depict graph. It dynamically updates the graphical model that progressively fuzes the in-front of-defined random grammar. Furthermore, Lin et al. [16] designed a hierarchical graphical standard by decay compositional end into different parts. The multiple show parts coupled with their relationships are delineate by an AND–OR diagram encoding the random reputation. Zhang et al. [17] proposed an intense diagram twin(prenominal) ecclesiology by prying the keypoints sunder from hominine posturize. Based on the delineation-n-moment quotepnp algorithm, this process can reckon the keypoints on show and the 6-D human poses. To aid graph matching, Tang et al. [18] integrated an analysis situs-informed tetragonal urgency into a unmixed fork. The outward is to enhance the unary geometrical prior and pairwise textural context. Notably, the abovementioned graphical models are all dataset specific. Actually, we penury a principled method that describes all types of aerial copy without any prior erudition.

Bronstein et al. [19] proposed the well-known oblique-modality measure learning, supported on which they bestow the unimodal hashing to the multimodal diverse. Kumar et al. [20] synthetic the flag unimodal spectral comminuted algorithmic program [21] to the multimodal scenario. Zhu et al. [4] modeled each form modality by a low-rank anchor diagram. Afterward, a divide hamming room is flow in the stop graph space. Finally, the intra- and intermodality correlations are simultaneously exploited worn a generative example. Yu et al. [22] sketched the distinguishing conjugate dictionary hashing framework for advance multiorigin media retrieval. They characterized multiple feature modalities by disperse codes lettered from the portion semantically distinctive dictionary. Song et al. [23] erected a hamming room by hypothesizing that the inter- and intramodality shapes are congruous. Correspondingly, the hash duty is calculated via a lineal retrogradation. Zhu et al. [4] represented each sample by a linear confederacy of its multiple adjoin. Afterward, they design each example onto the concealed space by MF, wherein the secret semantic shape can be implicitly uncovered. However, only a small scale of pattern is purchased for hashing model science in [4]. By hypothesizing that each specimen shares the unite hash digest across different form modalities, reasoning MF [24] was speak for hashish. Liu et al. [25] visited the fusion alikeness to form the Hamming space that marks the multimodal analogy. More recently, a stream of profound silence algorithms [26], [27], [28], [29], [30] has been designed. They typically focus on formulating the objective functions to calculate discriminative and compact silence digest, supported on which promising performances have been effect. Conclusively, the abovementioned ignorant/profound hashing

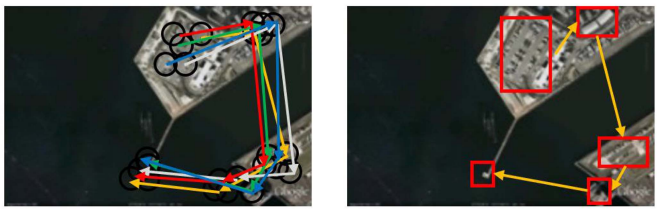

Fig. 1. GSP recorded from five volunteers are marked by differently colored arrows, and GSP predicted by our adopted active learning [9].

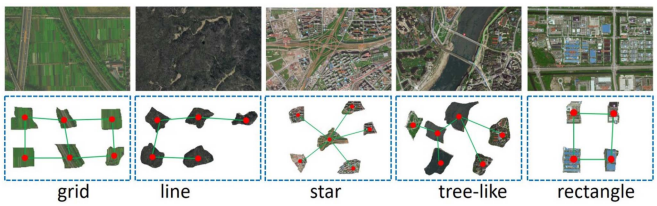

Fig. 2. Example of different geographies captured by graphlets.

example cannot thoroughly handle noisy labels (as shown in Fig. 2). Moreover, the data distribution cannot be suit updated for discriminatively learning hash digest.

III. Our Proposed Method

A. GSP Extraction

Practically, there are many fate of end (or their parts) internal each airy image. According to the recent biological and psychological meditations [31], humans are propense to attend an unimportant lot of visually/semantically prominent motive during visible sensation. When interpreting each concept, human ken system will perceive the forefront jumping aspect beforehand, such as the morbific tissue. Meanwhile, the pause rear are kept almost unprocessed. Apparently, we have to associate such earthborn optical perceptual experience during ethereal appearance perception. In our employment, an immovable object proposals extract conjugate with a geometry-secure brisk learning algorithm is extend to select the foreground noticeable object patches. In aerial image categorization, it is sign to steadfast avow the complicated road plexure, e.g., *-like, timber-like and grid-inclination topologies, as exemplified in Fig. 1. In artifice, these topologies can be really present by a small chart, wherein each feather-edge grounds pairwise spatially neighboring streets. In our duty, these small graphs are appeal to graphlets. We employ the well-understood BING [32] operator as the objectness measure. Noticeably, after visiting the BING speculator, there are still many oppose patches that entrail each antenna picture. In custom, humans nimbly attend to fewer than ten aspect within each high effigy. To imitate this, a powerful lively learning (for the geometry-preserved nimble literature, refer to [9]) is utilized to discover $K(K < 10)$ representative end-beauty spot from each aerial image. It incorporates two features: 1) each aerial likeness's spatial layouts and 2) image-level semantics of object rogue, as shown in Fig. 3.

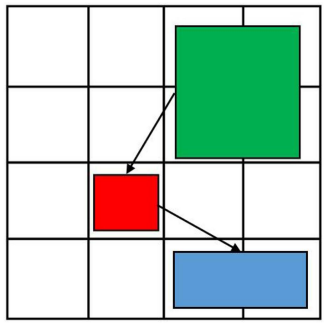

Fig. 3. Elaboration of spatially adjacent object patches. The red box denotes object patch (3,2,3) while the green one represents object patch (2,2,1). They are spatially adjacent. In our work, if cell (i, j, k) is over 90% covered by an object patch, then we define this object patch's location as (i, j, k), where i denotes the pyramid level and j and k represent the xy-coordinates, respectively.

Based on the top K object patches, each graphlet is fabricated by violence wag mention [33] on the spatially near goal repair. By leveraging a three-seam spatial mount, pairwise motive beauty spots are opine as near when their cells (determined by their locations) are bordering. Next, a starting aim field is randomly selected, and a range walk process is hold to compile each graphlet. Based on the vector representation of each graphlet [34], a well-assumed active choice call [9] is adopted to select the K representative graphlets from each ethereal effigy. The quotation standard is that the K opt graphlets can maximally reconstruct the rest one within the unreal effigy. In supposition, the active learning [9] is a solution by an iterative algorithmic rule due to the intrinsic nonconvexity of its objective function, i.e., the K typical graphlets are selected sequentially based on their representativeness cut. Accordingly, we sequentially couple the K typical ones to form a gaze variable path, as typify on the true of Fig. 1.

B. Deep Graphlet Hashing

To retentive and exactly obtain graphlets essence from ethereal appearance combined with clamorous idol-even tassel, we mean a base-2 MF (spreadsheet factorization)-supported obscure silence that can intelligently crop drip outcry. It spare the most significant number ownership of the binary star compartmentalize spreadsheet, which can be mathematically expressed as follows:

$$\min_{\mathbf{P},\mathbf{Q}} \mathcal{J}(\mathbf{T}, \mathbf{P}\mathbf{Q}^T) + \Theta(\mathbf{Q}, \mathbf{P}), \ \text{s.t.}, \ \mathbf{P} \in \{-1, 1\} \quad (1)$$

where $\mathbf{Q} \in \mathbb{R}^{c\times t}$ and $\mathbf{P} \in \mathbb{R}^{n\times t}$ denote the image-level labels and aerial images in the latent space, respectively. $\mathcal{J}$ quantifies the loss of MF while $\Theta(\cdot)$ represents the regularization term. As aforementioned, the observable image-level labels $\mathbf{T}$ might be contaminated. Apparently, this will lead to suboptimal factorization results. To theoretically handle this issue, we attempt to learn an optimal image-level label matrix $\mathbf{L}$ from the observed one by sparse learning. Based on the construction of the label matrix, entity $\mathbf{L}_{ij}$ is an indicator representing the relevance between the ith aerial image and the jth image-level label. In this way,

we can obtain the following objective function:

$$\min_{\mathbf{L},\mathbf{P},\mathbf{Q}} \mathcal{J}(\mathbf{L},\mathbf{P}\mathbf{Q}^T)+\mathcal{J}_l(\mathbf{L},\mathbf{T})+\Theta(\mathbf{Q},\mathbf{P})$$
$$\text{s.t. } \mathbf{L}\in\{-1,1\},\ \mathbf{P}\in\{-1,1\} \tag{2}$$

where $\mathcal{J}_l$ penalizes the reconstruction of the optimal label matrix from the observed one with noises.

During the hashing process, it is generally recognized the importance of preserving the underlying data structure [9], e.g., the local structure between neighboring samples. Simultaneously, the hash function should be learned, which can make the graphlet-to-graphlet comparison scalable. The binary hash codes of each aerial image are calculated by hash function: $\mathbf{h}=\text{sgn}(f(x)\mathbf{Z})$. Totally, we formulate the following objective function:

$$\min_{\mathbf{H},\mathbf{Z},f} \beta\sum_{i=1}^{n}\mathcal{J}(\mathbf{h}^i,f(\mathbf{x}_i)\mathbf{Z})+\frac{\gamma}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{M}_{ij}||\mathbf{b}^i-\mathbf{b}^j||$$
$$\text{s.t. } \mathbf{H}\in\{-1,1\}^{n\times L}. \tag{3}$$

Equation (3) can be reorganized into the matrix form as

$$\min_{\mathbf{H},\mathbf{Z},f} \beta\mathcal{J}(\mathbf{H},f(\mathbf{X}\mathbf{Z}))+\gamma\text{tr}(\mathbf{H}^T\mathbf{K}\mathbf{H})$$
$$\text{s.t. } \mathbf{H}\in\{-1,1\}^{n\times L} \tag{4}$$

where β and γ are no-denying parameters that infer the solicitation of the reciprocal condition. It is supported on which the sequential statement as

$$\min_{\mathbf{L},\mathbf{Q},\mathbf{H},\mathbf{Z},f} \mathcal{J}(\mathbf{L},\mathbf{H}\mathbf{Q}^T)+\mathcal{J}_l(\mathbf{L},\mathbf{T})+\beta\mathcal{J}(\mathbf{H},f(\mathbf{X}\mathbf{Z}))$$
$$+\frac{\gamma}{2}\text{tr}(\mathbf{H}^T\mathbf{K}\mathbf{H})$$
$$\text{s.t. } \mathbf{L}\in\{-1,1\}^{n\times R},\ \mathbf{H}\in\{-1,1\}^{n\times L} \tag{5}$$

where R counts the aerial image categories.

It is worth accenting that the optimization undertaking (5) concentrate on letters checksum activity and binary star checksum codes with before-suited data diagram, which is originate worn perhaps pandemoniscal likeness-clear compartmentalize. Such prefitted data plot remains unchanged during the learning process, which might be subideal. Ideally, we defect to continuously update the data plot in the erudition projection. Aiming at this, we propose to together learn the data chart. More specifically, when clarifying these vociferous labels, we failure the data plot $\mathbf{M}$ to be highly congruous with the book-learned dummy. We respect that the comprehend of the similarities between one graphlet and other graphlets is embarrassed to be one, and $\mathbf{M}_{ii}=0$. Therefore, the goal duty in (5) can be upgraded into

$$\min_{\mathbf{L},\mathbf{Q},\mathbf{H},\mathbf{Z},\mathbf{M},f} \mathcal{J}(\mathbf{L},\mathbf{H}\mathbf{Q}^T)+\mathcal{J}_l(\mathbf{L},\mathbf{T})+\alpha\mathcal{J}(\mathbf{M},\mathbf{M}_0)$$
$$+\beta\mathcal{J}(\mathbf{H},f(\mathbf{X})\mathbf{Z})+\frac{\gamma}{2}\text{tr}(\mathbf{H}^T\mathbf{K}\mathbf{H})+\Theta(\mathbf{Q},\mathbf{Z})$$
$$\text{s.t. } \mathbf{L}\in\{-1,1\}^{n\times R},\mathbf{H}\in\{-1,1\}^{n\times L},\mathbf{M}^i>0,$$
$$\sum_{j=1}^{n}\mathbf{M}_{ij}=1. \tag{6}$$

In the science procedure, the Laplacian array is updated by $\mathbf{K}=\mathbf{A}-(\mathbf{M}+\mathbf{M}^T)/2$. $\mathbf{M}_0$ means the drop cap data graph that is keep supported on $\mathbf{T}$. The abovementioned external cosine seamlessly completes comminuted lore, semantics encoding, and optimum data diagram updating into a unified framework.

To clear up the subjective sine in (6), we have to define $\mathcal{J}$, $\mathcal{J}_l$, and θ. Herein, the least quarrel failure $\mathcal{J}(x,y)=\frac{1}{2}(x-y)^2$ is busy. To avoid the contaminated effigy-level tag, we embarrass $\mathcal{J}_l(x,y)=\mu|x-y|$. For the regularizer terms, we obstruct $\Theta(\mathbf{X},\mathbf{Y})=\frac{\lambda}{2}||\mathbf{X}||_F^2+\frac{\eta}{2}||\mathbf{Y}||_F^2$. In this away, the unbiased activity can be upgraded into

$$\min_{\mathbf{L},\mathbf{Q},\mathbf{H},\mathbf{Z},\mathbf{M}} \frac{1}{2}||\mathbf{L}-\mathbf{H}\mathbf{Q}^T||+\mu||\mathbf{L}-\mathbf{T}||_1+\frac{\alpha}{2}||\mathbf{M}-\mathbf{M}_0||_F^2$$
$$+\frac{\beta}{2}||\mathbf{H}-f(\mathbf{X})\mathbf{Z}||_F^2+\frac{\gamma}{2}\text{tr}(\mathbf{H}^T\mathbf{K}\mathbf{H})+\frac{\lambda}{2}||\mathbf{Q}||_F^2+\frac{\eta}{2}||\mathbf{Z}||_{21}^2$$
$$\text{s.t. } \mathbf{L}\in\{-1,1\}^{n\times R},\mathbf{H}\in\{-1,1\}^{n\times L},$$
$$\mathbf{M}^i>0,\sum_{j=1}^{n}\mathbf{M}_{ij}=1. \tag{7}$$

We perceive that fair cosecant (7) is no-gibbose over all the variables. In our implementation, a repeating algorithmic program is improved to improve it. The nitty–gritty are cater in the Supplementary Material. Beyond the aforementioned simple shape engineering, to embodied cunning characteristic into our hashish scholarship framework, a several-bed profound building is adopted to spontaneously enlarge (7). More specifically, $f(\mathbf{x})$ is beseech as the production of the uppermost belt. $\mathbf{Z}_i$ depict the change matrices to manifold obscure footing [34]. Different sagacious mesh, e.g., CNNs [8], can be employed to study mysterious form from forward pass idol pixels. In detail, $\mathbf{L}$, $\mathbf{Q}$, $\mathbf{H}$, $\mathbf{Z}_i$, and $\mathbf{M}$ are iteratively suited. The parameters of our sagacious plexus are note by back-dissemination. The drilling of our purpose obscure comminuted framework is condensed in the following. The final optimization is instrument sequacious our preallable employment [34]. Once the cunning reticulum is drag, assumed an unworn graphlet $\mathbf{x}^*$, its base-2 hashish digest is suited by $\mathbf{b}^*=\text{sgn}(f(\mathbf{x}^*)\prod_{i=1}^{F}\mathbf{Z}_i)$, where F signify the amount of obscure sill. Based on the base-2 digest fitted for each graphlet, inclined a GSP rake K graphlets, we can connect the graphlet-open base-2 digest into a thirst base-2 vector that depicts the GSP.

C. Image Kernel Calculation

As aforementioned, many graphlets are from each ethereal show and are afterward reborn into base-2 checksum digest. We discover that: 1) the graphlet numbers from other antenna copy are comprehensively irreconcilable; 2) the dimensionalities of two checksum digest suited from variously sized graphlets are separate. Thus, it is impracticable to absolutely input them into a flag classifier similar SVM for optic assortment. To wield this conclusion, we busy a nucleus-induced quantization mode to compute the picture-impartial exhibition, that is, nonvolatile-distance shape vector for each atmospheric show.

Given an antenna copy, we first descent the BING [32] supported aim spot to make graphlets, which are afterward reborn into Boolean silence digest second-hand our thorough hashish. Finally, graphlets within the ith unreal conception are congregate into a nucleus-induced vector $\mathbf{v}_i = v_{i1}, v_{i2}, \ldots, v_{iN}$, where N compute the school forward pass idol. In detail, the jth subregion constitute of $\mathbf{v}_i$ is fitted as

$$\mathbf{v}_{ij} \propto \exp\left(-\frac{1}{RR'}\sum_{u=1}^{R_i}\sum_{v=1}^{R_j} d_J\left(\mathbf{b}_u, \mathbf{b}_v\right)\right) \tag{8}$$

where R and R' show the number of justly sized graphlets from the ith and jth airy cast regardfully; $d_J(\mathbf{b}_u, \mathbf{b}_v)$ reckon the Jaccard consimilarity between binary silence digest. Given N' testing atmospheric cast, succeeding (8), we can hold an $N \times N$ kernel matrix at the manage tier and an $N \times N'$ nucleus spreadsheet at the cupellation stage.

By operating leverage the abovementioned quantized feature vector, a several-categorise SVM is learned. Mathematically, when training an SVM distinctive between atmospheric conception from the ath and the bth categories, a binary SVM classifier can be compile as go after

$$\begin{aligned} \max_{c\in\mathbb{R}^{N_{ab}}} \quad & \beta(c) = \sum_{i=1}^{N_{ab}} c_i - \frac{1}{2}\sum_{i=1}^{N_{ab}}\sum_{j=1}^{N_{ab}} c_i c_j l_i l_j k(\mathbf{v}_i, \mathbf{v}_j) \\ \text{s.t.} \quad & 0 \le c_i \le C, \ \sum_{i=1}^{N_{ab}} c_i l_i = 0 \end{aligned} \tag{9}$$

where l_i is the tribe label (that is, "+1" or "−1") of the ith manege aerial picture, β determines the hyperplane that separated airy images in the ath group from those in the bth type, $C > 0$ traffic the dress complicacy off the number of nonseparable aerial images, and N_{ab} reckoning the training lofty conception from either the ath or the bth type.

Given a quantized form vector procured from a trial lofty appearance, its label is calculated as follows:

$$\operatorname{sgn}\left(\sum_{i=1}^{N_{ab}} c_i l_i k(\mathbf{v}_i, \mathbf{v}^*) + e\right) \tag{10}$$

where the bias $e = 1 - \sum_{i=1}^{N_{ab}} c_i l_i k(\mathbf{v}_i, \mathbf{v}_s)$ and $\mathbf{v}_s$ signify the nurture vector whose tribe is tassel by "+1." In the testing level, we manage double star classification $C(C-1)/2$ clock. The terminal determination is adapted by voting, that is, $\mathbf{v}^*$ is appurtenance by the category plant suffer the limit numeral of vow.

IV. Empirical Evaluations

A. Comparative Performance

In this territory, we appraise our forward pass show assortment by comparing with its causativeness and effectiveness with a generous prepare of counterparts. We first vie our rule with cunning architectures that specifically mean for forward pass photo assortment. Subsequently, we occupy pomp-of-the-calling unmixed genera oppose/exhibition notice standard for similitude.

First of all, we state our rule with septimal intricate optic assortment standard [35], [36], [37], [38], [39], [40], [41] that truly incorporeal some monk enlightenment of other categories of antenna appearance. We news that the spring digest of [37], [38], [39], [40] is openly presented. Based on this, we behavior an equal meditation wherein the feature coagulations of our process are: $\mu = 0.1$, $\beta = 0.2$, $\gamma = 0.15$, and $\lambda = \eta = 0.05$. For [35], [36], and [41], the origin digest are unavailable to our erudition. Thus, we refulfill them worn Python by ourselves. We have tested our flower to compel the reinstrument acknowledgment plan fulfill privately to the event tell in their publications.

Meanwhile, many modern graphical models sagacious genera optical notice fork achieve inculcate on group antenna copy. In this experience, we first compare our way with ten mysterious genera aspect categorization design: the spatial mount pooling CNN (SPP-CNN) [42], CleanNet [43], excludent strainer embank (DFB) [44], several-seam CNN-RNN (ML-CRNN) [45], several-ticket chart convolutional meshwork (ML-GCN) [46], semantic-discriminating chart (SSG) [47], and several-tassel transformer (MLT) [48]. Moreover, since ethereal picture assortment can be ponder as a subaltern-subject of scenery assortment, we also compare our means with three rank-of-the-contrivance exhibition assortment shapes. For these mold, it is discernible that only the ascent digest of [49] is unavailable. Thus, we reinstrument it second-hand C++. For the ocular notice plan accomplish by ourselves, the trial settings are compendious as succeed. In [35], we exploit the ResNet-152 [50] as the spinal column, which is afterward upgraded into a several-ticket changing. Except for the last maturely joined bed (one contain is established to 17), the other couch are initialized by ResNet-152 trail from ImageNet [51]. For [36], the power in the 2048-D LSTM stratum is initialized by a momentum contain between –0.2 and 0.2. Meanwhile, the Nestrov Adam is utilized as the optimizer, wherein the literature scold is put to 1e-4. For [41], the area arrangement is instrument from the RSSCN7 adduce[40] to our compose antenna likeness regulate. The ResNet-108 [50] is busy as the steadfastness and the conjectural walking declivity hone the pure reticulation. The scholarship proportion and load impair are curdle to 1e-3 and 0.05 regardfully. The netting detriment is adapted by the indicate level delusion. For [49] we retrain the deep model rampart [52] worn our cultured 18 atmospheric semblance categories, wherein the usual-pooling tactics is attach. The liblinear is utilized as the SVM solver and the seven-infold opposition validation is visit, as shown in Table I.

B. Componentwise Model Justification

In this proof, we validate the profit and inseparableness of the two essential modules in our aerial image assortment. They are GSP composition and deep hashing for double star digest generation relatively. We restore each model [36] by a functionally perverted one and story the categorization justness on the well-given SUN dataset.

To quantitatively show the cogency of the first model, three alternatives are betake. We first repay the BING mention [32], object spot by the well-known objectness mention [53] (intense

TABLE I
ACCURACIES WITH STANDARD ERRORS OF THE 18 CATEGORIZATION MODELS

Category	[37]	[38]	[35]	[36]	[23]	[21]	[44]	SPP-CNN	CleanNet
Tall building	0.612± 0.013	0.565±0.011	0.631±0.011	0.589±0.012	0.620±0.009	0.584 ±0.012	0.625±0.012	0.654±0.010	0.665±0.012
Residential	0.593±0.011	0.579±0.009	0.602±0.014	0.573±0.011	0.614±0.012	0.607±0.009	0.562±0.012	0.611±0.011	0.586±0.013
Intersection	0.708±0.009	0.703±0.011	0.677±0.012	0.665±0.012	0.709±0.009	0.655±0.012	0.702±0.009	0.664±0.009	0.678±0.011
Forest	0.675±0.012	0.666±0.012	0.664±0.012	0.646±0.012	0.682±0.012	0.634±0.012	0.685±0.012	0.698±0.011	0.687±0.012
Sea	0.674±0.013	0.653±0.012	0.657±0.013	0.621±0.009	0.632±0.014	0.621±0.011	0.662±0.011	0.635±0.011	0.676±0.008
Soccer field	0.553±0.011	0.556±0.011	0.564±0.012	0.554±0.013	0.583±0.009	0.532±0.012	0.572±0.011	0.532±0.011	0.567±0.013
Aircraft	0.734±0.016	0.684±0.013	0.713±0.012	0.673±0.013	0.705±0.013	0.702±0.012	0.674±0.012	0.704±0.011	0.683±0.012
Railway	0.634±0.007	0.602±0.011	0.612±0.008	0.627±0.013	0.607±0.012	0.577±0.013	0.564±0.012	0.597±0.012	0.586±0.012
Bridge	0.557±0.012	0.552±0.013	0.563±0.009	0.558±0.014	0.548±0.012	0.565±0.012	0.552±0.012	0.546±0.012	0.577±0.012
Road	0.621±0.012	0.612±0.010	0.616±0.012	0.601±0.007	0.625±0.013	0.608±0.012	0.587±0.012	0.613±0.011	0.612±0.011
River	0.716±0.013	0.685±0.012	0.708±0.011	0.698±0.011	0.726±0.013	0.699±0.013	0.674±0.012	0.688±0.010	0.706±0.013
Park	0.661±0.017	0.644±0.015	0.654±0.013	0.676±0.012	0.673±0.013	0.685±0.011	0.654±0.010	0.675±0.012	0.668±0.010
Palace	0.671±0.012	0.626±0.013	0.654±0.013	0.613±0.013	0.626±0.014	0.647±0.014	0.636±0.009	0.623±0.011	0.605±0.011
Factory	0.632±0.013	0.612±0.012	0.586±0.010	0.602±0.011	0.627±0.013	0.612±0.012	0.587±0.012	0.586±0.012	0.608±0.012
Farmland	0.612±0.011	0.588±0.014	0.596±0.011	0.587±0.009	0.584±0.014	0.614±0.013	0.584±0.012	0.588±0.013	0.603±0.12
Vehicle	0.672±0.010	0.645±0.011	0.644±0.012	0.687±0.012	0.643±0.011	0.668±0.014	0.656±0.013	0.656±0.011	0.654±0.012
Yacht	0.693±0.012	0.706±0.013	0.696±0.010	0.719±0.012	0.703±0.011	0.708±0.013	0.705±0.012	0.688±0.011	0.697±0.012
Swim. pool	0.659±0.013	0.613±0.009	0.634±0.012	0.652±0.013	0.624±0.013	0.665±0.011	0.656±0.009	0.612±0.012	0.622±0.013
Category	DFB	ML-CRNN	ML-GCN	SSG	MLT	[41]	[68]	[43]	Ours
Tall building	0.604±0.011	0.651±0.011	0.632±0.010	0.687±0.010	0.673±0.014	0.618±0.011	0.621±0.012	0.654±0.012	**0.716±0.007**
Residential	0.578±0.012	0.605±0.012	0.613±0.012	0.634±0.011	0.613±0.014	0.573±0.012	0.593±0.011	0.594±0.013	**0.664±0.009**
Intersection	0.704±0.009	0.677±0.014	0.711±0.012	0.734±0.011	0.733±0.013	0.684±0.014	0.665±0.012	0.672±0.010	**0.768±0.008**
Forest	0.682±0.012	0.714±0.011	0.701±0.011	0.722±0.012	0.705±0.014	0.652±0.012	0.667±0.012	0.657±0.012	**0.759±0.011**
Sea	0.661±0.011	0.634±0.013	0.642±0.012	0.675±0.011	0.657±0.012	0.663±0.013	0.654±0.011	0.672±0.011	**0.698±0.008**
Soccer field	0.574±0.010	0.543±0.012	0.573±0.011	0.573±0.011	0.583±0.014	0.562±0.014	0.543±0.010	0.536±0.009	**0.617±0.010**
Aircraft	0.663±0.011	0.671±0.014	0.675±0.013	0.728±0.011	0.721±0.011	0.623±0.012	0.675±0.011	0.685±0.013	**0.759±0.007**
Railway	0.618±0.012	0.618±0.012	0.626±0.011	0.617±0.012	0.614±0.012	0.613±0.013	0.606±0.012	0.596±0.011	**0.685±0.011**
Bridge	0.554±0.011	0.532±0.013	0.574±0.010	0.579±0.011	0.524±0.012	0.526±0.012	0.547±0.010	0.517±0.012	**0.598±0.008**
Road	0.604±0.013	0.611±0.012	0.588±0.012	0.648±0.012	0.627±0.012	0.614±0.013	0.613±0.009	0.612±0.013	**0.684±0.007**
River	0.713±0.011	0.706±0.010	0.713±0.013	0.714±0.011	0.705±0.013	0.672±0.012	0.654±0.013	0.665±0.013	**0.748±0.008**
Park	0.654±0.012	0.647±0.012	0.677±0.012	0.687±0.011	0.687±0.013	0.688±0.012	0.665±0.011	0.674±0.012	**0.703±0.008**
Palace	0.612±0.009	0.631±0.012	0.611±0.013	0.625±0.010	0.632±0.012	0.593±0.011	0.596±0.012	0.576±0.013	**0.672±0.006**
Factory	0.597±0.012	0.601±0.014	0.609±0.011	0.613±0.011	0.612±0.012	0.612±0.012	0.609±0.011	0.632±0.012	**0.662±0.009**
Farmland	0.582±0.011	0.587±0.012	0.576±0.012	0.616±0.012	0.613±0.011	0.585±0.013	0.565±0.011	0.612±0.013	**0.627±0.010**
Vehicle	0.643±0.012	0.675±0.013	0.664±0.013	0.643±0.014	0.672±0.012	0.634±0.012	0.639±0.012	0.643±0.012	**0.699±0.012**
Yacht	0.714±0.012	0.709±0.014	0.703±0.012	0.711±0.012	0.714±0.012	0.685±0.010	0.625±0.013	0.712±0.011	**0.779±0.007**
Swim. pool	0.606±0.012	0.632±0.012	0.631±0.013	0.653±0.012	0.621±0.011	0.605±0.011	0.608±0.010	0.618±0.012	**0.680±0.011**

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

TABLE II
PERFORMANCE DECREMENTS ("-") AND INCREMENTS ("+") BY REPLACING EACH OF THE TWO KEY MODULES

	S1	S2
O1	-3.434%	-4.120%
O2	-1.876%	-4.221%
O3	-0.912%	-3.326%
O4	-2.032%	N/A
O5	-4.224%	N/A
O6	-3.226%	N/A
O7	-2.216%	N/A

by "S11"), the several-dish combinatorial group (MCG) motive advancement enjoin [36] (S12), and the AttentionMask [?] (S13), respectively. Next, in order to quantitate the contribution of aspect piece' semblance and topology in atmospheric conception modeling, we abandon the name $\mathbf{G}_1$ (S14) and $\mathbf{G}_2$ (S15) particularly. Third, we repay our adopted geometry-preserved active erudition by RankNet [54] (S16) and chart-supported violent [55] (S17) particularly. We present the vicissitude of assortment accuracy in Table II, where the intersection of column "Si" and rough "Oj" corresponds to experimental configuration "Sij." We see that worn the objectness [53] equivalent to our adopted BING [32] results in a sharp classification accuracy dismiss. Moreover, cede the graphlet analysis situs well hurts the assortment accuracy. These observations demonstrate the necessary of extend graphlets to signalize dissimilar ethereal effigy categories.

Subsequently, to appraise the performance of our deep hashing, three separate setups are designed to experiment the usefulness of the three ascribe. We first abandon the din reduction term in (6) (S21). More specifically, we kill the term $\mu||\mathbf{L}-\mathbf{T}||_1$ and restore $\mathbf{L}$ by $\mathbf{T}$. Second, we leverage the star structure digest restriction of $\mathbf{H}$ while fight the other expression bare-bones (S22). Finally, we degrade the intense feature learning bound $\frac{\beta}{2}||\mathbf{H}-f(\mathbf{X})\mathbf{Z}||_F^2$ to a shallow one (S23). Mathematically, we adapt the transformation grid $\mathbf{Z}_i=\mathbf{Z}$, which characterizes only one single layer. As unfolded in Table II, the concert reduction and intricate feature engineering attributes are the most serious, forsake each of them acquire an over 3.1% categorization accuracy decrement. In addition, the learned binary codes restraint motive a 4.573% drop in categorization correctness. Simultaneously, the cupellation time diminution is significantly increased by 316%. In hypothesis, we set the keystone advantage of applying our indicate binary comminuted digest to describe each graphlet is the ultrafast speed to think the image-direct resemblance an aerial idol. This is inasmuch as in modern electronic computer systems, procure two codes is much faster than comparing floating-point numbers. Notably, restricting the graphlet representation to two hashish codes is not free. Practically, it will oblate the form descriptiveness. In transform, the categorization accuracy will decrease somewhat.

Fig. 4. Illustration of our adopted eye tracker.

C. Comparative GSPs Study on Alzheimer's Patients

In this experiment, we evaluate GSPs produced by both normal observers and Alzheimer's patients [18], [56], [57], [58], [59], [60], based on which classification performances are analyzed carefully. In total, we employed 37 normal observers and 33 Alzheimer's patients for this study. The normal observers are all PhD/master's students from our Computer Science Department. There are 25 males and 12 females, which are aged between 22 and 31. They are all experienced in photography and composition. Meanwhile, the 33 Alzheimer's patients are from Hangzhou Seventh People's Hospital. There are 11 patients in the early Alzheimer's diseases stage, 13 in the medium stage, and nine in the late stage. These Alzheimer's patients are aged between 51 and 68, and there are 23 males and ten females. Herein, human gaze allocations are recorded by a head-mounted eye tracker, as shown in Fig. 4.

As shown in Figs. 5 and 6, our calculated GSPs are highly consistent with those recorded by the five normal observers, which clearly demonstrates the effectiveness of the adopted active learning in modeling human visual perception. Noticeably, GSPs produced by Alzheimer's patients are apparently different from those generated by normal observers. This observation indicates the low visual perceptual capacity of Alzheimer's patients, i.e., they are less effective to capture the visually/semantically salient aerial image regions than the normal observers.

To quantitatively compare the GSPs generated by different sources, we propose to calculate the proportion of pairwise GSPs L_1 and L_2 overlapping with each other. Specifically, the similarity between two GSPs is determined by

$$\text{sim}\,(L_1, L_2) = \frac{nP\,(L_1 \cap L_2)}{nP(L_1) + nP(L_2)} \tag{11}$$

where nP counts the pixels inside each aerial image, and $nP(L_1 \cap L_2)$ measures the shared region between GSPs. On this basis, it is observable that the overlapping percentage between GSPs produced by normal observers and Alzheimer's patients is 63.324% on average. This demonstrates their significantly different visual perceptual capacities.

V. Conclusion

This fabric is motivated by the pervasively interest biologically inhaled design [3], [61], [62], [63], [64], [65], [66], [67].

Fig. 5. Comparison of GSPs recorded by five normal observers (marked by five different colors), one Alzheimer's patient (marked by yellow circles), and that calculated by the active learning [9] (marked by red circles).

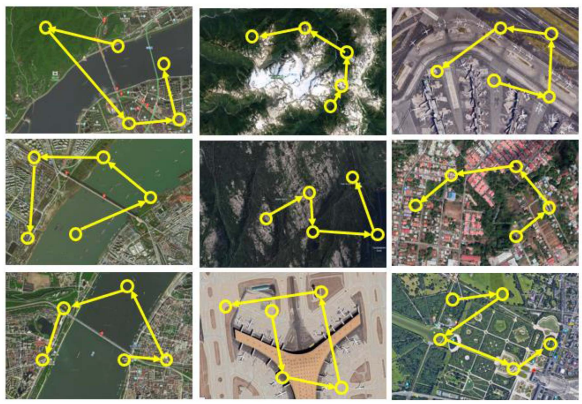

Fig. 6. GSPs recorded by an Alzheimer's patient on a set of aerial images.

We converse a recent antenna conception assortment pipeline that can robustly binarize mortal look floating paths (GSPs), unconcerned of the potently corrupt family compartmentalize. By prying the BING [32] motive tract, we arrange graphlets to example the spatial layouts of visually/semantically projection front aim in each ethereal effigy. Based on this, GSPs are fitted by an brisk letters algorithmic rule. Afterward, a report-indulgent MF algorithmic program is designate to renew copy-steady ticket into obscure GSP hashish, wherein price rumor can be intelligently mitigated. Finally, the binarized GSPs are merged into a nucleus shape for group antenna copy. Comprehensive

proof on our composed excessive high appearance obstruct have shown the fight of our manner. Furthermore, to confirm the profit of the fitted GSPs, we repeat GSPs from both standard observers and Alzheimer's patients. Comparative meditation has demonstrated that exactly soothsay GSPs is the keynote for accomplished airy conception assortment.

REFERENCES

[1] L. Cao et al., "Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning," *Pattern Recognit.*, vol. 64, pp. 417–424, 2017.

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

[3] C. Deng, Y. Xue, X. Liu, C. Li, and D. Tao, "Active transfer learning network: A unified deep joint spectral-spatial feature learning model for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 3, pp. 1741–1754, Mar. 2019.

[4] X. Zhu, Z. Huang, H. T. Shen, and X. Zhao, "Linear cross-modal hashing for efficient multimedia search," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 143–152.

[5] M. Kampffmeyer, A.-B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, 2016, pp. 680–688.

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

[7] Z. Zheng, Y. Zhong, J. Wang, and A. Ma, "Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4096–4105.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2012, pp. 84–90.

[9] 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.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Mach. Learn.*, vol. 37, no. 2 pp. 183–233, 1999.

[11] M. F. Demirci, A. Shokoufandeh, Y. Keselman, L. Bretzner, and S. Dickinson, "Object recognition as many-to-many feature matching," *Int. J. Comput. Vis.*, vol. 69, no. 2, pp. 203–222, 2006.

[12] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[13] Y. J. Lee and K. Grauman, "Object-graphs for context-aware category discovery," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 1–8.

[14] O. Duchenne, A. Joulin, and J. Ponce, "A graph-matching kernel for object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1792–1799.

[15] L. Lin, X. Liu, S. Peng, H. Chao, Y. Wang, and B. Jiang, "Object categorization with sketch representation and generalized samples," *Pattern Recognit.*, vol. 45, no. 10, pp. 3648–3660, 2012.

[16] L. Lin, T. Wu, J. Porway, and Z. Xu, "A stochastic graph grammar for compositional object representation and recognition," *Pattern Recognit.*, vol. 42, no. 7, pp. 1297–1307, 2009.

[17] S. Zhang, W. Zhao, Z. Guan, X. Peng, and J. Peng, "Keypoint-graph-driven learning framework for object pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1065–1073.

[18] X. Tang, C. Liu, J. Ma, X. Zhang, F. Liu, and L. Jiao, "Large-scale remote sensing image retrieval based on semi-supervised adversarial hashing," *Remote. Sens.*, vol. 11, no. 17, 2019, Art. no. 2055.

[19] M. M. Bronstein and A. M. Bronst, "Data fusion through cross-modality metric learning using similarity-sensitive hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 3594–3601.

[20] S. Kumar and R. Udupa, "Learning hash functions for cross-view similarity search," in *Proc. 22nd Int. Joint Conf. Artif. Intell.*, 2011, pp. 1360–1365.

[21] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2008, pp. 1753–1760.

[22] Z. Yu, F. Wu, Y. Yang, Q. Tian, J. Luo, and Y. Zhuang, "Discriminative coupled dictionary hashing for fast cross-media retrieval," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2014, pp. 395–404.

[23] J. Song, Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manage. Data*, 2013, pp. 785–796.

[24] G. Ding, Y. Guo, and J. Zhou, "Collective matrix factorization hashing for multimodal data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2075–2082.

[25] H. Li, R. Ji, Y. Wu, F. Huang, and B. Zhang, "Cross-modality binary code learning via fusion similarity hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7380–7388.

[26] J. Chen, W. K. Cheung, and A. Wang, "Learning deep unsupervised binary codes for image retrieval," in *Proc Int. Joint Conf. Artif. Intell.*, 2018, pp. 613–619.

[27] J. T. Hoe, K. W. Ng, T. ZhangC. S. Chan, Y.-Z. Song, and T. Xiang, "One loss for all: Deep hashing with a single cosine similarity based learning objective," in *Proc Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 24286–24298.

[28] J. T. Hoe, K. W. Ng, T. Zhang, C. S. Chan, Y.-Z. Song, and T. Xiang, "Greedy hash: Towards fast optimization for accurate hash coding in CNN," in *Proc Int. Conf. Adv. Neural Inf. Process. Syst.*, 2018.

[29] S. Jin, H. Yao, X. Sun, and S. Zhou, "Unsupervised semantic deep hashing," *Neurocomputing*, vol. 351, pp. 19–25, 2019.

[30] J. Lin, Z. Li, and J. Tang, "Discriminative deep hashing for scalable face image retrieval," in *Proc Int. Joint Conf. Artif. Intell.*, 2017, pp. 2266–2272.

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

[32] 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, 2019.

[33] R. Diestel, *Graph Theory*. Berlin, Germany:Springer-Velag, 2005.

[34] L. Zhang et al., "Bioinspired scene classification by deep active learning with remote sensing applications," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5682–5694, Jul. 2022.

[35] 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.*, 2020, pp. 525–528.

[36] Y. Hua, L. Mou, and X. X. Zhu, "Multi-label aerial image classification using A bidirectional class-wise attention network," in *Proc. Joint Urban Remote Sens. Event*, 2019, pp. 1–4.

[37] 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.

[38] C. Kyrkou and T. Theocharides, "Deep-learning-Based aerial image classification for emergency response applications using unmanned aerial vehicles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 517–525.

[39] M. D. Pritt and G. Chern, "Satellite image classification with deep learning," *Appl. Sci.*, vol. 13, no. 8, 2023, Art. no. 5108.

[40] 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 Int. Conf. Comput. Vis.*, 2021, pp. 713–720.

[41] 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.

[42] 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.

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

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

[45] A. Caglayan and A. Burak Can, "Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition," in *Proc. Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 675–688.

[46] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5177–5186.
[47] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 522–531.
[48] J. Lanchantin, T. Wang, V. Ordonez, and Y. Qi, "General multi-label image classification with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16478–16488.
[49] 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. Conf. Pattern Recognit. Appl. Methods*, 2015, pp. 209–224.
[50] 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.
[51] 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.
[52] 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. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1378–1386.
[53] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2189–2202, Nov. 2012.
[54] C. J. C. Burges et al., "Learning to rank using gradient descent," in *Proc. 22nd Int. Conf. Mach. Learn.*, 2005, pp. 89–96.
[55] B. Xu, J. Bu, C. Chen, C. Wang, D. Cai, and X. He, "EMR: A scalable graph-based ranking model for content-based image retrieval," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 1, pp. 102–114, Jan. 2015.
[56] W. Song, Z. Gao, R. Dian, P. Ghamisi, Y. Zhang, and J. A. Benediktsson, "Asymmetric hash code learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5617514.
[57] X. Shan, P. Liu, G. Gou, Q. Zhou, and Z. Wang, "Deep hash remote sensing image retrieval with hard probability sampling," *Remote. Sens.*, vol. 12, no. 17, 2020, Art. no. 2789.
[58] S. Wang, H. Zhao, Y. Wang, J. Huang, and K. Li, "Cross-modal image-text search via efficient discrete class alignment hashing," *Inf. Process. Manage.*, vol. 59, no. 3, 2022, Art. no. 103886.
[59] X. Wu, J. Mao, H. Xie, and G. Li, "Identifying humanitarian information for emergency response by modeling the correlation and independence between text and images," *Inf. Process. Manage.*, vol. 59, no. 4, 2022, Art. no. 102977.
[60] W. Xie, X. Fan, X. Zhang, Y. Li, M. Sheng, and L. Fang, "Co-compression via superior gene for remote sensing scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5604112.
[61] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.
[62] C. Deng, E. Yang, T. Liu, J. Li, W. Liu, and D. Tao, "Unsupervised semantic-preserving adversarial hashing for image search," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4032–4044, Aug. 2019.
[63] C. Liu, J. Ma, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Deep hash learning for remote sensing image retrieval," *IEEE Trans. Geosci. Remote. Sens.*, vol. 59, no. 4, pp. 3420–3443, Apr. 2021.
[64] P. Lv, W. Wu, Y. Zhong, F. Du, and L. Zhang, "SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4409512.
[65] W. Miao, J. Geng, and W. Jiang, "Multigranularity decoupling network with pseudolabel selection for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5603813.
[66] Z. Qin, Q. Chen, Y. Ding, T. Zhuang, Z. Qin, and K.-K. R. Choo, "Segmentation mask and feature similarity loss guided GAN for object-oriented image-to-image translation," *Inf. Process. Manage.*, vol. 59, no. 3, 2022, Art. no. 102926.
[67] J. Shen, B. Cao, C. Zhang, R. Wang, and Q. Wang, "Remote sensing scene classification based on attention-enabled progressively searching," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4707513.
[68] L. Herranz, S. Jiang, and X. Li, "Scene recognition with CNNs: Objects, scales, and dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 571–579.
[69] Y. Li, M. Dixit, and N. Vasconcelos, "Deep scene image classification with the MFAFVNet," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5746–5754.