scientific reports

Check for updates

OPEN

A novel binary hashing for agricultural scenery classification

Han Wang[1], Jun-He Liu[1]✉ & Yi Yang[2]

In this research, we present PerceptHashing, a technique designed to categorize million-scale agricultural scenic images by incorporating human gaze shifting paths (GSPs) into a hashing framework. For each agricultural image, we identify visually and semantically significant object patches, such as fields, crops, and water bodies. These patches are linked to form a graphlet, establishing a network of spatially adjacent patches, and a GSP is then extracted using an active learning algorithm. The GSP reflects the distribution of human gaze across different regions of each agricultural scene, typically involving fewer than 12 regions, as validated by cross-validation. We then design a binary hashing framework that effectively leverages the semantics encoded in these GSPs. This framework integrates three key elements: (i) refinement of noisy labels, (ii) incorporation of deep image-level agricultural semantics, and (iii) updates to the adaptive data graph. The resulting hash codes from each GSP are converted into a kernelized visual descriptor for classification. To evaluate the influence of GSPs on agricultural image classification both qualitatively and quantitatively, we conducted an extensive user study comparing GSPs from typical observers with those from Alzheimer's patients. The results demonstrate that: (1) the classification accuracy of 1.22 million agricultural aerial images using our approach is significantly higher than those processed by other methods, and (2) GSPs from 33 Alzheimer's patients differ markedly from those of 37 normal observers, leading to notable differences in classification accuracy ($acc_{normal} = 0.694$ versus $acc_{patient} = 0.322$).

With the growing trend of launching multiple satellites on a single rocket, there has been a significant increase in the deployment of earth observation satellites dedicated to monitoring agricultural landscapes in recent years. These satellites capture agricultural scenic images that reveal intricate spatial patterns, such as crop grids, irrigation systems, and field boundaries. This study aims to categorize each agricultural image into different semantic groups, such as crop types, water sources, and farm infrastructure. Leveraging the spatial configurations of these agricultural features to determine their semantic categories is an effective strategy in many artificial intelligence-driven agricultural systems. For instance, by mapping the spatial layouts of different crops, irrigation patterns, dry lands, and wetlands, it is possible to proactively monitor crop growth, detect potential water shortages, and predict pest migrations, all of which are crucial for sustainable farming practices. Furthermore, smart analysis of how humans visually interpret agricultural images can aid in the surveillance and management of environmental challenges such as droughts, soil degradation, and flooding. In practice, the way humans direct their gaze when viewing agricultural landscapes can be represented as a pathway, with each link connecting sequentially noticed features, such as fields, irrigation lines, or patches of vegetation.

For research in AI applied to agriculture, various shallow and deep recognition and labeling models have been developed to analyze agricultural scenic imagery. These approaches can be categorized into the following types: (1) Multi-Instance Learning and CNN-based methods for localizing agricultural features using weak labels[2,3]; (2) the use of semantic diffusion through graphical models for interpreting agricultural scenes[4,5]; and (3) advanced deep learning frameworks for the semantic labeling of agricultural landscapes[6,7]. Testing and state-of-the-art technologies have validated their capabilities. However, to the best of our knowledge, current models do not fully capture the complexity of agricultural scenic images for several reasons:

- In practice, each agricultural image may contain numerous ground objects with complex spatial arrangements, making it challenging to effectively and efficiently interpret their underlying semantics. The primary difficulties include: (a) mathematically formulating the intricate spatial relationships between agricultural features such as crop fields, irrigation systems, and land use patterns, and (b) designing a deep learning architecture that transforms these spatial relations into consistent visual features. Additionally, integrating these spatial relationships into conventional classifiers (e.g., Bayesian classifier[8]) further complicates the process.

[1]College of Biological and Food Engineering, Huanghuai University, Zhumadian 463000, Hanan, China. [2]College of Computer Sciences, Beijing Technology and Business University, Beijing 102488, China. ✉email: liujunhe79@126.com

- The vast number of objects in each agricultural image necessitates extensive pixel-level annotation, which is labor-intensive. With advancements in learning from weak annotations, we can rely on image-level labels to infer regional semantics. However, using weakly-supervised, user-supplied labels introduces subjectivity and potential errors, creating challenges. Therefore, building a system that can withstand noisy labels is critical.
- An efficient classification system for agricultural images must accurately define the sample distributions within the feature space. However, initial sample distributions may be suboptimal due to inaccurate user-supplied labels. Ideally, a mathematical model capable of dynamically adjusting the real geometry among samples in the label refinement process is necessary. Designing such a multi-attribute optimization model that is feasible requires considerable expertise.

To address and mitigate these issues, we propose a biologically-inspired approach to agricultural image classification. Our method features two primary innovations: (1) the structured assembly of multiple visually and semantically important graphlets into a Gaze Shift Path (GSP), and (2) a novel matrix factorization (MF) technique that converts GSPs in an agricultural image into binary hash codes, allowing for the correction of potentially inaccurate semantic labels. A schematic of the proposed method is displayed in Fig. 2. Initially, from a diverse set of agricultural images, each containing some inaccurate semantic labels, object-based patches are extracted. These patches, representing features like crop types, irrigation systems, and land contours, are linked based on their spatial proximity to form graphlets. An active learning paradigm[1] is then developed to acquire a GSP that sequentially captures human attention to visually and semantically significant areas within each agricultural image (outlined in Sect 3.1). GSPs provide deeper insights than typical visual saliency maps because they capture actual gaze behaviors over agricultural scenes.

Next, a noise-resistant MF approach is applied to binarize the GSPs, facilitating rapid and accurate comparisons of graphlet pairs (explained in Sect "Advanced graphlet hashing technique"). This MF technique incorporates three key elements. As a result, binary codes for each GSP are generated. By binarizing all training GSPs, they are converted into kernel-induced vectors that serve as the foundation for training a multi-label SVM for visual classification (described in Sect "Image kernel for scenic image classification"). Extensive quantitative evaluations against state-of-the-art deep recognition models confirm the superior performance of our classifier in the context of agricultural scenic image classification.

Furthermore, to evaluate both qualitatively and quantitatively the role of Gaze Shift Paths (GSPs) in agricultural image classification, we performed a comparative analysis between the GSPs predicted by our model and those observed by 37 typical participants. The analysis revealed that our predicted GSPs matched those recorded from human participants in over 90% of cases. Additionally, GSPs collected from 33 Alzheimer's patients showed significant differences from both our model's predictions and the GSPs of healthy individuals. As a result, classification accuracy using GSPs from Alzheimer's patients was markedly reduced, demonstrating the impact of visual perception deficits on the interpretation of agricultural imagery.

This research contributes in three key ways: (1) the development of an advanced matrix factorization (MF) method that efficiently integrates multiple features to generate binary codes for GSPs, (2) the implementation of a robust visual recognition system that effectively categorizes agricultural aerial images on a large scale into various semantic groups, and (3) an extensive user study with 70 participants to rigorously evaluate the influence of GSPs on agricultural image classification.

The structure of this document is as follows: the next section reviews relevant literature related to our research. Section "Our methodology" details the proposed methods, including the construction of GSPs, our PerceptHashing strategy, and the implementation of a kernel-induced SVM classifier. Section "Experimental validation" outlines comprehensive experiments that validate the effectiveness of our approach. Section "Concluding remarks" summarizes our findings and suggests areas for further investigation.

Related work

Our approach to classifying scenic images intersects with two pivotal themes in AI: utilizing graph-based algorithm to analyzing visual data and implementing discrete hashing techniques.

Graphical models in image analysis: Many graphical models were designed for intricately capturing the interconnections between image components. For instance, Demirci et al.[9] developed methods to align vertices in annotated, noisy graphs. Felzenszwalb et al.[10] modeled deformable part relationships with a spring model to minimize correspondence costs. In another approach[11], graph vertices represent identifiable and ambiguous object parts, assigning category labels based on neighboring relations. Duchenne et al.[12] introduced a kernel for matching graphical representations to identify the labels of different objects. The authors[13] implemented a visual understanding framework that adjusts a graphical model that incorporates random grammars. Additionally, the authors[14] developed a multi-layer model breaking down objects into a structured And-Or graph. Zhang et al.[15] leveraged human pose keypoints to create a deep graph matching framework, applying the perspective-n-point algorithm[16] for precise object and pose mapping. Gao et al.[17] enhanced a deep model with a topology-aware quadratic constraint to fine-tune geometric and structural contexts. However, these graphical models are typically customized for specific datasets, highlighting the need for a versatile approach that can universally encode scenic images without pre-existing data dependencies.

Advancements in Discrete Hashing: Bronstein et al.[18] expanded on metric learning to adapt unimodal hashing into a multimodal context. Kumar et al.[19] broadened spectral hashing[20] to accommodate multiple modalities. Zhu et al. designed a framework where each feature mode is structured using a well-designed anchors graphical model, developing a common hashing feature space that integrates both intra- and inter-modal correlations through a generative approach. Yu et al.[21] proposed a hashing framework that utilizes a discriminative dictionary for enhanced media retrieval across multiple sources, encoding features through learned sparse codes. Despite these innovations, existing hashing models frequently struggle with noise in labels and lack the capability to

adaptively update data distributions to optimize hash code learning. Further explorations in remote sensing, as conducted by Shan et al.[22] and others, demonstrate the application of deep hashing algorithms specifically crafted for distinct retrieval and classification tasks within scenic imagery, underscoring the continuous need to refine these techniques to effectively handle complex image datasets.

Alzheimer's disease (AD) was characterized as a neurodegenerative disorder by the buildup of amyloid-beta plaques and neurofibrillary tangles, leading to progressive cognitive decline, as explained by Trejo-Lopez et al.[23]. These hallmark features were central to understanding AD pathology, and their presence had been crucial in diagnostic criteria and therapeutic research. Treatments like Donanemab and Lecanemab, designed to target amyloid-beta, showed potential in slowing early-stage disease progression. Mintun et al.[24] discussed the efficacy of Donanemab in clearing amyloid plaques, while Van Dyck et al.[25] highlighted Lecanemab's promise in reducing cognitive decline when administered early. Dubois et al.[26] emphasized the importance of incorporating biomarkers such as amyloid and tau proteins into diagnostic criteria, as they improved early detection when combined with clinical phenotypes. Porsteinsson et al.[27] stressed that evolving diagnostic practices needed to integrate biomarkers for early-stage identification to ensure timely intervention with treatments like Donanemab. Srivastava et al.[28] reviewed AD treatment approaches, noting that while no cure existed, symptomatic relief remained vital in managing the disease. They explored pharmacological and non-pharmacological strategies, highlighting the challenges in developing a definitive therapy. Tatulian[29] examined these challenges, acknowledging the failures of past therapies but remained hopeful about emerging, multifaceted approaches targeting various aspects of the disease. Dreylikh and Karaman[30] explored AD's multifactorial causes, including genetic predispositions and environmental factors, arguing that personalized treatment approaches were necessary to address individual variability in disease progression. Finally, Vaz and Silvestre[31] presented an overview of pharmacological advancements targeting not just amyloid-beta, but also tau protein and neuroinflammation, calling for continued innovation to develop therapies that effectively altered AD's progression and mitigated cognitive decline.

Our methodology

GSP derivation

In the domain of agricultural scenic photography, numerous objects or fragments thereof are visible within each image. Research in biological and psychological fields indicates that human observation typically concentrates on a few objects that are visually or semantically distinct within their field of view[32]. Predominantly, the human visual system filters out background elements, focusing primarily on striking foreground features. This understanding is integral to our method for classifying scenic images, where we implement a rapid object detection technique alongside a geometry-preserving active learning strategy to pinpoint key foreground elements.

We deploy the widely recognized BING operator[33] as a measure of object presence. Despite its use, many object patches remain identifiable within each image. Observational data suggest humans usually focus on fewer than ten objects per image. Our method mimics this selective attention by employing a sophisticated active learning approach, termed geometry-preserving active learning[1], which selects up to L essential object-aware patches in each scenery. This selection is influenced by both the geometric arrangement and the semantic attributes of the patches.

To assemble each graphlet, we utilize a random walk strategy[34] across object patches that are spatially adjacent, as defined by a multi-tier spatial pyramid. Adjacency is determined based on the proximity of cells, which themselves are defined by the patches' placements. A random patch is selected as the starting point for the random walk that builds the graphlet.

Each graphlet's vector representation[35] undergoes an esteemed active selection process[1], aiming to identify K graphlets that can effectively represent the broader image. This selection is designed to maximize the representational ability of the chosen graphlets, effectively reconstructing the remaining graphlets in the image. The inherent non-linear nature of this task necessitates the use of an iterative algorithm, which sequentially selects the most representative graphlets based on their capacity to encapsulate the image's features. These graphlets are then sequentially linked to construct a Gaze Shift Path (GSP), demonstrated on the right in Fig. 1.

Advanced graphlet hashing technique

For efficient and precise comparison of graphlets derived from scenic images that come with imprecise image tags, we developed a binarized matrix factorization (MF)[36]-based deep hashing technique tailored to effectively manage label inaccuracies. This approach maintains the critical integer characteristics of the matrix containing binarized labels. This is represented by:

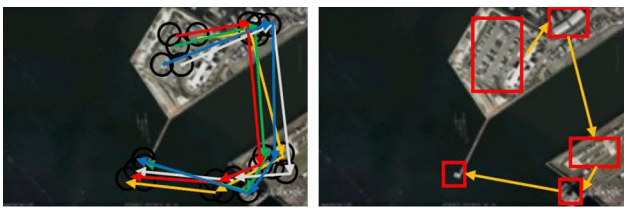

Fig. 1. Gaze shift paths captured by 5 volunteers from our computer science department (left), each indicated by arrows of different colors; GSP forecasted by our deployed active learning algorithm[1] (right).

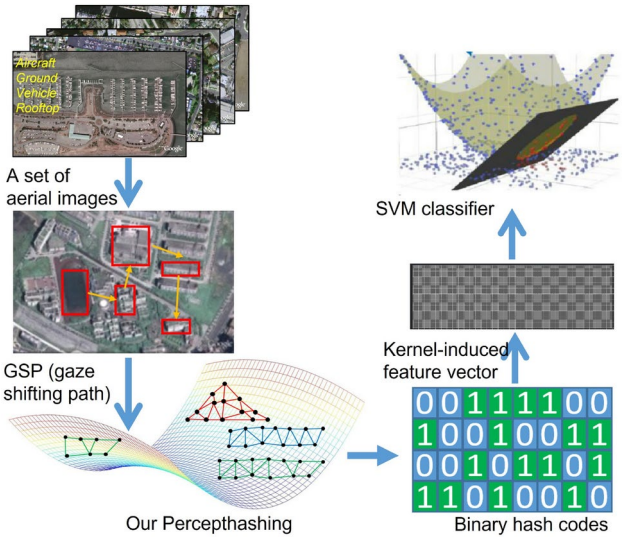

Fig. 2. The design of our scenic image classification system.

$$\min_{\mathbf{Q},\mathbf{R}} \mathscr{J}(\mathbf{U}, \mathbf{Q}\mathbf{R}^T) + \Theta(\mathbf{R}, \mathbf{Q}), \quad s.t., \quad \mathbf{Q} \in \{-1, 1\}, \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{c \times t}$ symbolizes the image-level labels and $\mathbf{Q} \in \mathbb{R}^{n \times t}$ corresponds to the scenery within our implicitly hidden space. $\mathscr{J}$ measures the matrix factorization losing value and $\Theta(\cdot)$ signifies the regularization component. As previously noted, the visible image-level labels $\mathbf{T}$ may be compromised, potentially leading to less-than-ideal factorization outcomes. To address this challenge, we focus on calculating the matrix containing image tags $\mathbf{L}$ using the explicit labels through the so-called sparse learning. Within this framework, the element $\mathbf{L}_{ij}$ acts as a marker that determines the association among the *i*-th scenic picture and the *j*-th tag at image-level. This setup leads to the below formulation:

$$\begin{aligned} &\min_{\mathbf{M},\mathbf{Q},\mathbf{R}} \mathscr{J}(\mathbf{M}, \mathbf{Q}\mathbf{R}^T) + \mathscr{J}_l(\mathbf{M}, \mathbf{U}) + \Theta(\mathbf{R}, \mathbf{Q}), \\ &s.t. \ \ \mathbf{L} \in \{-1, 1\}, \quad \mathbf{Q} \in \{-1, 1\}, \end{aligned} \tag{2}$$

Herein, $\mathscr{J}_l$ imposes a penalty for reconstructing the calculated tag matrix from the explicit and noisy labels.

In our matrix hashing, the significance of maintaining the intrinsic sample distribution is well acknowledged[1], such as the local connectivity among neighboring samples. Concurrently, it is essential to develop a hash function that enhances the scalability of graphlet comparisons. Our calculated binarized codes for each scenery are derived using objective function: $\mathbf{t} = \mathrm{sgn}(f(x)\mathbf{W})$. Ultimately, the below formulation can be obtianed:

$$\begin{aligned} &\min_{\mathbf{I},\mathbf{W},f} \beta \sum\nolimits_{i=1}^{n} \mathscr{J}(\mathbf{h}^i, f(\mathbf{x}_i)\mathbf{W}) + \frac{\gamma}{2} \sum\nolimits_{i=1}^{n} \sum\nolimits_{j=1}^{n} \mathbf{M}_{ij} ||\mathbf{b}^i - \mathbf{b}^j||, \\ &s.t. \ \ \mathbf{I} \in \{-1, 1\}^{n \times L}, \end{aligned} \tag{3}$$

Equation (3) is reformulated in matrix notation as follows:

$$\begin{aligned} &\min_{\mathbf{I},\mathbf{W},f} \beta \mathscr{J}(\mathbf{I}, f(\mathbf{X}\mathbf{W})) + \gamma \mathrm{tr}(\mathbf{I}^T \mathbf{K} \mathbf{I}), \\ &s.t. \ \ \mathbf{I} \in \{-1, 1\}^{n \times L}, \end{aligned} \tag{4}$$

Herein, β and γ denote two positive coefficients that represent the weight of the respective items. L signifies the number of dimensions in the implicit hashing space. $\mathbf{K} \in \mathbb{R}^{n \times n}$ represents the Laplacian matrix, specifically, $\mathbf{K} = \mathbf{A} - \mathbf{M}$. Herein, matrix $\mathbf{A}$ denotes a diagonal one wherein the *ii*-th entity is calcuated by accmulating the i^{th} row of $\mathbf{M}$. As indicated in Eq. (4), this setup allows for the concurrent learning of the binarized codes for different sceneries.

To ensure sufficiently compatible ability between matrix factorization and binarized codes, we propose a shared space between the structural framework utilized by the designed MF as well as the hashing. This implies that our implicit feature space identified by binarized MF corresponds directly to the Hamming space computed. Consequently, the semantic insights garnered through the proposed method using the calculated matrix of tags can inform and direct the hashing operation. Formally, it is easy to set $\mathbf{I} = \mathbf{Q}$ and $L = t$, which leads us to the below expression:

$$\min_{\mathbf{L},\mathbf{R},\mathbf{I},\mathbf{W},f} \mathscr{J}(\mathbf{L},\mathbf{I}\mathbf{R}^T) + \mathscr{J}_l(\mathbf{L},\mathbf{U}) + \beta\mathscr{J}(\mathbf{I}, f(\mathbf{X}\mathbf{W})) + \frac{\gamma}{2}\mathrm{tr}(\mathbf{I}^T\mathbf{K}\mathbf{I}), s.t.\ \ \mathbf{L} \in \{-1,1\}^{n\times R},\ \ \mathbf{I} \in \{-1,1\}^{n\times L}, \tag{5}$$

Herein, R counts the scenic image classes.

It's important to note, the optimizing operation of (5) is dedicated to training the hashing operation and binarized codes using a pre-determined samples. It is built based on potentially imprecise image tags. This pre-set data graph remains static in the model training. This will be sub-optimal. Optimally, the data graph should be dynamically updated throughout the learning phase. In this way, we design a methodology for simultaneous cacluation of the data graph. Specifically, as we refine the noisy labels, it is crucial that the data graph $\mathbf{M}$ aligns well with the evolving model. Consequently, the value of Eq. (5) is enhanced as follows:

$$\begin{aligned}&\min_{\mathbf{L},\mathbf{R},\mathbf{I},\mathbf{W},\mathbf{M},f} \mathscr{J}(\mathbf{L},\mathbf{I}\mathbf{R}^T) + \mathscr{J}_l(\mathbf{L},\mathbf{U}) + \alpha\mathscr{J}(\mathbf{M},\mathbf{M}_0) + \\ &\beta\mathscr{J}(\mathbf{I}, f(\mathbf{X})\mathbf{W}) + \frac{\gamma}{2}\mathrm{tr}(\mathbf{I}^T\mathbf{K}\mathbf{I}) + \Theta(\mathbf{R},\mathbf{W}), \\ &s.t.\ \mathbf{L} \in \{-1,1\}^{n\times R}, \mathbf{I} \in \{-1,1\}^{n\times L}, \sum\nolimits_{j=1}^{n} \mathbf{M}_{ij} = 1,\end{aligned} \tag{6}$$

During the model updating, $\mathbf{K}$ is adjusted using $\mathbf{K} = \mathbf{A} - (\mathbf{M} + \mathbf{M}^T)/2$. $\mathbf{M}_0$ represents the input graph, which is derived from $\mathbf{T}$. The stated formulation effectively combines hash code calculation, semantic integration, and dynamic updates to the data graph within a cohesive framework.

To address the solution of Eq. (6), it is necessary to specify $\mathscr{J}$, $\mathscr{J}_l$, and Θ. For this purpose, we calculate $\mathscr{J}(a,b) = \frac{1}{2}(a-b)^2$. To mitigate the effects of imprecise image tags, we define $\mathscr{J}_l(a,b) = \mu|a-b|$. In our regularization, we apply $\Theta(\mathbf{X},\mathbf{Y}) = \frac{\lambda}{2}||\mathbf{X}||_F^2 + \frac{\eta}{2}||\mathbf{Y}||_F^2$. Accordingly, the objective function is refined to:

$$\begin{aligned}&\min_{\mathbf{L},\mathbf{R},\mathbf{I},\mathbf{W},\mathbf{M}} \frac{1}{2}||\mathbf{L} - \mathbf{I}\mathbf{R}^T|| + \mu||\mathbf{L} - \mathbf{U}||_1 + \frac{\alpha}{2}||\mathbf{M} - \mathbf{M}_0||_F^2 + \\ &\frac{\beta}{2}||\mathbf{I} - f(\mathbf{X})\mathbf{W}||_F^2 + \frac{\gamma}{2}\mathrm{tr}(\mathbf{I}^T\mathbf{K}\mathbf{I}) + \frac{\lambda}{2}||\mathbf{R}||_F^2 + \frac{\eta}{2}||\mathbf{W}||_{21}^2 \\ &s.t.\ \mathbf{L} \in \{-1,1\}^{n\times R}, \mathbf{I} \in \{-1,1\}^{n\times L}, \mathbf{M}^i > 0, \sum\nolimits_{j=1}^{n} \mathbf{M}_{ij} = 1,\end{aligned} \tag{7}$$

We note that the objective function outlined in Eq. (7) is not convex when considering the entire parameters simultaneously. Herein, we have developed to iteratively optimize it. Further details on the algorithmic approach can be accessed through the following link: https://docs.google.com/document/d/1ltjBlrVPvBkMthxntZWt0gMmJXoD00qD/edit?usp=share_link&ouid=101578137679720572579&rtpof=true&sd=true.

In addition to leveraging shallow feature engineering, our hash learning framework integrates deep features using a multi-layered deep architecture, thereby naturally enhancing Eq. (7). Specifically, $f(\mathbf{x})$ represents the output from the uppermost layer, and $\{\mathbf{W}_i\}$ denotes transformation matrices that apply across multiple deep layers[35]. Various multi-layer models, like the CNN[8], can be leveraged for calculating hierarchical features directly from pixels of scenic images. The components $\mathbf{L}$, $\mathbf{R}$, $\mathbf{I}$, $\{\mathbf{W}_i\}$, and $\mathbf{M}$ are carefully refined during the learning process. Variables of the model are adjusted, with network training protocols based on methodologies from our previous research[35]. If we finished training our deep model, for any new GSP $\mathbf{x}^*$, we calculate the binarized codes using $\mathbf{b}^* = \mathrm{sgn}(f(\mathbf{x}^*)\prod_{i=1}^{F}\mathbf{W}_i)$, where F counts the multiple layers.

Utilizing the binarized codes generated for each object patch, when constructing a Gaze Shift Path (GSP) that includes K graphlets, these graphlet-level binary codes are concatenated into an extended binary vector that effectively represents the entire GSP.

Image kernel for scenic image classification

As noted earlier, numerous graphlets are derived from each scenic image, which are then translated into binary hash codes. We recognize two challenges: 1) the quantity of graphlets varies across different scenic images, and 2) the hash code dimensions differ depending on the size of the graphlets extracted. These variations make it impractical to input the hash codes directly to a conventional recognizer for semantic classification purposes. Aiming at it, we utilize a kernelized quantizing technique to create a consistent vector for each scenery. For a given scenic image, we obtain graphlets to construct the Gaze Shift Path (GSP), which are then converted into binary hash codes. Subsequently, the GSP from the i-th scenic image is used to compute a kernelized vector $\mathbf{v}_i = \{v_{i1}, v_{i2}, ..., v_{iN}\}$, where N counts the training samples. Specifically, the j-th component of $\mathbf{e}_i$ is determined as follows:

$$\mathbf{e}_{ij} \propto \exp(-\frac{1}{SS'}\sum\nolimits_{i=1}^{S_i}\sum\nolimits_{j=1}^{S_j} d_J(\mathbf{b}_i,\mathbf{b}_j)), \tag{8}$$

Herein, S and S' denote the count of graphlets extracted from pairwise scenic images; $d_J(\cdot,\cdot)$ is used to calculate the Jaccard similarity between the binary hash codes of these graphlets.

Utilizing the quantized feature vectors derived above, we learn a multi-label SVM to recognize the scenic images. Specifically, when distinguishing between sceneries from two different categories, the process involves training a binary classifier, that is,

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

Herein, l_i represents the category label assigned to the each scenic image during training; β defines the hyperplane that discriminates between sceneries belonging to both categories; C is a regularization parameter that mediates the trade-off of achieving a low mistake on the training data and maintaining model simplicity to avoid overfitting; and N_{ab} counts the scenic images from both categories during training.

For a test scenic image represented by a calculated vector $\mathbf{e}^* \in \mathbb{R}^{N_{ab}}$, its category is predicted by:

$$\mathrm{sgn}(\sum_{t=1}^{N_{ab}} d_t l_t k(\mathbf{e}_i, \mathbf{e}^*) + b), \tag{10}$$

Herein, the bias term b is calculated as $1 - \sum_{t=1}^{N_{ab}} d_t l_t k(\mathbf{e}_i, \mathbf{e}_s)$. $\mathbf{e}_s$ is the vector associated with the '+1' class. During the testing phase, binary classification is performed $C(C-1)/2$ times for all category pairs. The final classification is determined through a voting mechanism, whereby $\mathbf{v}^*$ is assigned the label from the category set that accumulates the most votes.

Experimental validation

Dataset and experimental setup

Our study rigorously assesses the PerceptHashing method through multiple experimental phases. Initially, we compare our approach against well-established image classification models to demonstrate its competitiveness. This is followed by a step-by-step model justification that scrutinizes the contribution of each module within our methodology. Additionally, we conduct both qualitative and quantitative evaluations of Gaze Shift Paths (GSPs) derived from both typical observers and individuals with Alzheimer's.

The dataset used in our experiments is extensive, comprising approximately 2.3 million agricultural scenic images, segmented into 0.63 million for the training phase, 1.22 million for testing the model, and 0.45 million for validation purposes. These images were meticulously gathered from sources such as Google, Apple, and Bing Maps through an automated web crawler that operated continuously for over 3500 hours. Following collection, an initial labeling process was conducted by 82 volunteers, all with expertise in agriculture and photography. This group, consisting of 50 males and 32 females, manually annotated 14.7% of the images from each significant agricultural region, using 47 different image-level labels specific to agricultural landscapes. Subsequently, a multi-label SVM model was trained to assign labels to the remaining untagged images, which were then reviewed and corrected manually by the volunteers. Approximately 26% of the SVM's initial misclassifications were rectified during this manual review. This comprehensive agricultural image dataset is planned for public release upon formal acceptance of the research.

Compared testing of classification performance

This section presents the performance evaluation of our scenic image classification approach, assessing both its effectiveness and efficiency against a diverse array of models. We initially benchmark our technique against specialized deep architectures tailored for scenic image classification and further extend the comparison to encompass leading deep generic object and scene recognition models.

We commence by evaluating our approach against seven specialized deep visual classification models[37–43], which are designed with intrinsic understanding of various aerial image categories. The implementations for[37,38,41,42] are readily accessible; we utilize these to perform comparative evaluations under the parameters: $\mu = 0.12$, $\beta = 0.25$, $\gamma = 0.2$, $\lambda = \eta = 0.15$. For models[39,40,43] where implementations are not publicly available, we have reconstructed these using Python to closely match their documented performance.

Also, we compare our technique against ten contemporary multi-layer visual recognition models known for their efficacy in object categorization: SPP-CNN[44], CleNet[45], DFB[46], ML-CRN[47], MLGCN[48], SGM[49] and MLT[50]. Given the relevance of scenic image classification to scene categorization, we also benchmark against three advanced scene classification models. Only the implementation of[51] was unavailable, leading us to recreate it in C++. We strive to achieve fidelity to the results these models originally reported. Furthermore, the proposed technique is made comparison with six recently published models for scene and image categorization algorithms[22,52–56].

For the models implemented in-house, experimental configurations are as follows: - For[39], we adopt ResNet-152[57] as the foundation, modified for multi-label tasks except for the terminal fully-connected layer, which remains set to 17 units, initialized with weights from ImageNet-trained ResNet-152[58]. - For[40], LSTM layers are initialized with random values within [– 0.2, 0.2]. - In[43], domain adaptation is applied using RSSCN7[42] datasets, with ResNet-108[57] be leveraged as the backbone. - For[51], we retrain the object bank[59] with 18 refined scenic image categories using average-pooling and liblinear for SVM processing, employing 7-fold cross-validation for robustness assessment. As the results presented in Table 1, our method's performance is highly competitive.

	37	38	39	40	41	42	43	SPP+CNN	CleNet
Average	0.651 ± 0.013	0.623 ± 0.011	0.643 ± 0.013	0.651 ± 0.013	0.631 ± 0.014	0.671 ± 0.012	0.671 ± 0.013	0.632 ± 0.015	0.633 ± 0.013
	DFB	**MLCRN**	**ML-GCN**	**SGM**	**MLT**	51	60	61	**Ours**
Average	0.623 ± 0.013	0.656 ± 0.013	0.641 ± 0.014	0.643 ± 0.013	0.621 ± 0.012	0.603 ± 0.011	0.615 ± 0.015	0.623 ± 0.014	**0.693 ± 0.012**

Table 1. Comparative analysis with various recognition algorithms. We conducted multiple evaluations, repeating each test 20 times and duly reporting the standard deviations for each.). Significant values are in bold.

	T1	T2
P1	− 3.643%	− 4.076%
P2	− 1.942%	− 4.461%
P3	− 0.911%	− 3.476%
P4	− 2.146%	na
P5	− 4.331%	na
P6	− 3.251%	na
P7	− 2.165%	na

Table 2. Performance variations (increases "+" and decreases "-") upon substitution of each core module.

Evaluation of key components in our scenic image classification framework

In this segment of our research, we evaluate the indispensable roles of two fundamental components in our scenic image classification framework: the Gaze Shift Path (GSP) construction and the binary hashing process for generating code. We conduct a simulated failure analysis by replacing each component with a less effective alternative to measure the impact on classification accuracy, using the well-regarded SUN dataset[62].

To probe the influence of the GSP construction, we replace our object detection method with three different systems: the commonly used objectness measure[63] (labelled "S11"), the Multi-Scale Combinatorial Grouping (MCG) for object proposals[64] (S12), and AttentionMask for object detection[65] (S13). Additionally, we explore the relevance of the appearance of object patches and their spatial configurations by removing terms $\mathbf{G}_1$ (S14) and $\mathbf{G}_2$ (S15) from the model. We also test the replacement of our geometry-preserving active learning approach with RankNet[66] (S16) and a graph-based ranking system[67] (S17). The changes in classification accuracy are documented in Table 2, where each configuration "Sij" represents the interaction of "Ti" and "Pj". Notably, substituting our BING[33] approach with a standard objectness measure[63] results in a substantial decrease in accuracy. Likewise, neglecting the graphlet structure significantly lowers performance, reinforcing the value of using graphlets to differentiate between scenic image categories.

Next, we evaluate the binary hashing component by altering three crucial attributes. In the first place, we remove the noises reduction component (S21), substituting $\mathbf{L}$ with the original label matrix $\mathbf{T}$ and omitting the regularization term $\mu||\mathbf{L}-\mathbf{T}||_1$. Secondly, we lessen the restrictions on binary codes in $\mathbf{H}$ while maintaining other conditions unchanged (S22). At last, the multi-layer feature engineering is simplified from a complex multi-layered to a basic single-layer transformation (S23), employing a unified transformation matrix $\mathbf{Z}$. The results, displayed in Table 1, show that removing the noise reduction and advanced feature learning significantly reduces accuracy by more than 3.1%. Additionally, relaxing the binary code constraints decreases accuracy by 4.573% and increases testing time by 316%, highlighting the critical role of these components in ensuring the robustness and efficiency of scenic image classification.

Comparative study of gaze shift paths in Alzheimer's patients

This segment of our research delves into the analysis of Gaze Shift Paths (GSPs) as observed among both normal observers and Alzheimer's patients to ascertain their effects on the accuracy of image classification. We recruited a total of 70 participants divided into two distinct groups for this study.

The first group consisted of 37 normal observers, all of whom are either Ph.D. or master's students. This group includes 25 males as well as 12 females, all with a background in photography and image composition.

The second group comprised 33 individuals diagnosed with Alzheimer's disease, sourced from the Hangzhou Seventh People's Hospital.The selection of the 33 participants with Alzheimer's disease was meticulously conducted at Hangzhou Seventh People's Hospital, a leading institution renowned for its expertise in neurology and geriatric care. The participants were recruited through various channels, including local memory clinics and Alzheimer's support groups, ensuring broad outreach to potential candidates. This approach facilitated a diverse sample reflective of the local population while also allowing for the identification of individuals at various stages of the disease. Adherence to stringent diagnostic criteria was paramount; assessments included comprehensive cognitive evaluations, such as the Mini-Mental State Examination (MMSE), which provided a standardized measure of cognitive impairment. Additionally, advanced neuroimaging techniques, including MRI and PET scans, were utilized to identify Alzheimer's-specific biomarkers, such as amyloid plaques and neurodegeneration, thus confirming the diagnosis with a high degree of accuracy. The patients are categorized

by disease progression with 11 in the early stages, 13 in the middle stages, and 9 in the late stages of Alzheimer's, ranging in age from 51 to 68. The demographic breakdown includes 23 males and 10 females.

Inclusion criteria mandated that participants had a confirmed diagnosis of mild to moderate Alzheimer's disease and demonstrated the capacity to provide informed consent, or had a legal representative who could do so on their behalf. To enhance the validity of the findings, individuals with significant comorbid psychiatric or neurological disorders that could confound results were excluded. This rigorous selection process ensured that the study focused on a homogeneous group of Alzheimer's patients, thereby enhancing the reliability of the outcomes. The sample size of 33 was strategically chosen based on power analyses indicating that this number would yield sufficient statistical power to detect meaningful effects in cognitive assessments, while also considering logistical factors such as funding, recruitment timelines, and the feasibility of conducting longitudinal follow-ups. Ethical considerations were central to the study's design; all participants, or their legal representatives, provided informed consent after being thoroughly briefed on the study's purpose, procedures, potential risks, and benefits. This study also received approval from the institutional review board (IRB), adhering to ethical standards for research involving vulnerable populations. Overall, the recruitment and selection process at Hangzhou Seventh People's Hospital was designed to create a robust and representative sample of participants with Alzheimer's disease, aligning with the study's objectives to investigate cognitive decline and evaluate potential therapeutic interventions effectively. This comprehensive approach not only enhances the credibility of the findings but also ensures that the voices of individuals affected by Alzheimer's disease are central to the research.

When collecting gaze-shifting paths using an eye tracker, several vision conditions must be considered to ensure accurate data. Adequate and consistent lighting is essential to minimize reflections, while participant comfort should be prioritized to avoid visual strain, allowing breaks if needed. Standardizing screen size and distance from participants is crucial, as these factors can influence gaze behavior. The complexity and type of visual stimuli presented should vary to assess different gaze patterns effectively. Proper calibration and regular validation of the eye tracker are vital for accurate tracking, along with clearly defined tasks, such as visual searches or reading. Additionally, individual differences-such as age, visual acuity, and familiarity with technology-should be accounted for, as these can significantly impact gaze behavior. By carefully managing these conditions, researchers can collect reliable data that reflects participants' visual attention and cognitive processing accurately. Besides, as double checking by Hangzhou Seventh People's Hospital, the 33 Alzheimer's patients for our user study are only with Alzheimer's disease. They don't have other diseases that may affect the user study.

For our analysis, we utilized a novel human eye tracking equipment, as shown in Fig. 3, to capture the gaze patterns of participants. This method provided precise data on how each person's gaze is distributed across different visual stimuli, particularly highlighting the contrast in visual perception capabilities between healthy individuals and those affected by Alzheimer's disease, and its subsequent impact on their ability to classify images. Some example eye-tracking results are shown in Figs. 4 and 5.

To objectively measure the differences in GSPs among various groups, we employ a method to assess the overlap between pairs of GSPs, designated as L_1 and L_2. The similarity between these GSPs is calculated using the following formula:

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

Herein, nP denotes the pixel numberin each respective scenic image area, and $\mathrm{nP}(L_1 \cap L_2)$ indicates the intersection of pixels between two GSPs. According to our findings, the average overlap percentage between the GSPs of normal and abnormal individuals is approximately 63.547%, reflecting their distinct visual processing abilities.

While it is challenging to definitively predict whether the observed differences between healthy participants and Alzheimer's patients are solely attributable to Alzheimer's disease or age-related vision impairments, it's important to note that the Alzheimer's patients in this study were screened to ensure they had no vision diseases. This rigorous assessment helps strengthen the argument that the differences observed are more likely related to the effects of Alzheimer's rather than confounding factors such as undiagnosed vision impairments.

Fig. 3. A depiction of the eye tracker used in our study.

Fig. 4. A comparison of gaze shift paths (GSPs): Visualized are GSPs captured by five normal observers, each labeled using distinct colors, alongside a GSP from Alzheimers patients highlighted with yellow circles, and a GSP computed by us, denoted using red.

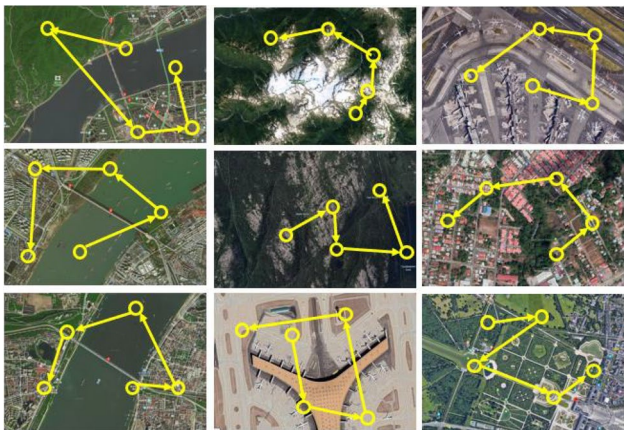

Fig. 5. GSPs tracked from an Alzheimer's patient across various scenic images.

To ensure that Alzheimer's patients in the study had no underlying vision diseases, a comprehensive pre-screening process was conducted by Hangzhou Seventh People's Hospital based on our requirement, which included collecting detailed ocular health histories to identify any previous diagnoses of vision-related conditions. Standard visual acuity tests were performed using a Snellen chart, along with thorough eye examinations that involved retinal assessments via ophthalmoscopy, intraocular pressure measurements to check for glaucoma, and visual field testing to evaluate peripheral vision. Additional tests included color vision assessments with Ishihara plates, contrast sensitivity evaluations, and refraction assessments to determine the need for corrective lenses. Clear exclusion criteria were established to rule out participants with any identified vision impairments or eye diseases, ensuring that only those with healthy vision were included in the study. This rigorous approach strengthens the argument that the differences observed between healthy participants and Alzheimer's patients are more likely attributable to the disease itself rather than confounding visual factors.

Experiment on ADHD children

This study aims to investigate differences in visual attention patterns between 16 children diagnosed with ADHD from Hangzhou Seventh People's Hospital, aged 6-12, and a matched control group using eye-tracking technology. Participants will include 4 children aged 6-7, 6 aged 8-9, 4 aged 10-11, and 2 aged 12, all screened for visual impairments and comparable IQ levels. Using our adopted eye tracker, participants will engage with dynamic scenes and complex static images, identifying target objects amidst distractors. Metrics such as fixation duration, number of fixations, saccade patterns, and response accuracy will be recorded. The procedure involves a brief introduction to the tasks in a quiet, well-lit environment, with breaks provided to minimize fatigue. Data will be analyzed using statistical software to compare gaze metrics and identify significant differences between ADHD and control groups. It is observable that children with ADHD will exhibit shorter fixation durations, increased fixation counts, and more frequent saccades, reflecting challenges in maintaining attention. Statistically, based on our fixation accuracy calculation in Eq. (11), the average overlap percentage between the GSPs of normal and ADHD patients are approximately 61.032%. This shows the significantly different visual processing abilities. Moreover, such overlap percentage between Alzheimer's patients and ADHD patients are about 43.009%, which means that these two types of diseases influences visual perception differently. Such findings may enhance understanding of attentional mechanisms in ADHD and inform targeted interventions to improve focus and learning strategies for affected children.

Concluding remarks

This research is driven by the increasing application of biologically-inspired models in artificial intelligence, particularly in the field of agricultural analysis. We introduce a state-of-the-art method for agricultural image classification that efficiently processes Gaze Shift Paths (GSPs) for binary representation, even when faced with noisy category labels. Our approach utilizes BING objectness measures to create graphlets that capture the essential spatial features of agricultural elements such as crop fields, water bodies, and farm structures within each image. These graphlets form the foundation for our GSPs, which are refined using an active learning algorithm. A subsequent matrix factorization (MF) method, designed to be resilient to label noise, integrates these features into a deep hashing scheme for robust classification. Extensive testing on a large-scale agricultural image dataset has demonstrated the high effectiveness of our methodology.

One notable limitation of our current model is the extensive computational resources required during the training phase of the deep hashing process. Future work will focus on optimizing and parallelizing the core components to improve training efficiency. Additionally, we plan to systematically explore the impact of GSP size variations on classification accuracy and aim to identify a concise set of class-related graphlets tailored for various intelligent agricultural systems.

Data availibility

The datasets used and/or analysed during the current study available from the corresponding author (Jun-He Liu) on reasonable request.

Received: 26 April 2024; Accepted: 24 October 2024
Published online: 11 November 2024

References

1. Zhang, Lijun et al. Active learning based on locally linear reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**(10), 2026–2038. https://doi.org/10.1109/TPAMI.2011.20 (2011).
2. Zhou, Sharon et al. DeepWind: Weakly supervised localization of wind turbines in satellite imagery. In *Annual Conference on Neural Information Processing Systems* (2009).
3. Cao, Liujuan et al. Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning. *Pattern Recogn.* **64**, 417–424. https://doi.org/10.1016/j.patcog.2016.10.033 (2017).
4. Shu, Tianmin, Xie, Dan, Rothrock, Brandon, Todorovic, Sinisa & Zhu, Song-Chun. Joint inference of groups, events and human roles in aerial videos. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2015.7299088 (2015).
5. Porway, Jake, Wang, Qiongchen & Zhu, Song-Chun. A hierarchical and contextual model for aerial image parsing. *Int. J. Comput. Vis.* **88**(2), 254–283. https://doi.org/10.1007/s11263-009-0306-1 (2010).
6. Zheng, Zhuo, Zhong, Yanfei, Wang, Junjue & Ma, Ailong. Foreground-aware relation network for geospatial object segmentation in high spatial resolution remote sensing imagery. *Comput. Vis. Pattern Recog.* https://doi.org/10.1109/CVPR42600.2020.00415 (2020).
7. Kemker, Ronald, Salvaggio, Carl & Kanan, Christopher. Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning. *ISPRS J. Photogramm. Remote. Sens.* **145**, 60–77. https://doi.org/10.1016/j.isprsjprs.2018.04.014 (2018).
8. Krizhevsky, Alex, Sutskever, Ilya & Hinton, Geoffrey E. ImageNet classification with deep convolutional neural networks. *Annual Conference on Neural Information Processing Systems*. https://doi.org/10.1145/3065386 (2012).
9. Fatih Demirci, M., Shokoufandeh, Ali, Keselman, Yakov, Bretzner, Lars & Dickinson, Sven. Object recognition as many-to-many feature matching. *Int. J. Comput. Vis.* **69**(2), 203–222. https://doi.org/10.1007/s11263-006-6993-y (2006).
10. Felzenszwalb, Pedro F. & Huttenlocher, Daniel P. Pictorial structures for object recognition. *Int. J. Comput. Vis.* **61**(1), 55–79. https://doi.org/10.1023/B:VISI.0000042934.15159.49 (2005).
11. Yong Jae Lee. Object-graphs for context-aware category discovery. *International Conference on Aquatic Invasive Species*. https://doi.org/10.1109/CVPR.2010.5540237 (2009).
12. Duchenne, Olivier, Joulin, Armand & Ponce, Jean. A graph-matching kernel for object categorization. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/ICCV.2011.6126445 (2011).
13. Lin, Liang et al. Object categorization with sketch representation and generalized samples. *Pattern Recogn.* **45**(10), 3648–3660. https://doi.org/10.1016/j.patcog.2012.03.017 (2012).
14. Lin, Liang, Tianfu, Wu., Porway, Jake & Zijian, Xu. A stochastic graph grammar for compositional object representation and recognition. *Pattern Recogn.* **42**(7), 1297–1307. https://doi.org/10.1016/j.patcog.2008.10.033 (2009).

15. Zhang, Shaobo, Zhao, Wanqing, Guan, Ziyu, Peng, Xianlin & Peng, Jinye. Keypoint-graph-driven Learning Framework for object pose estimation. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR46437.2021.00112 (2021).
16. Lepetit, V., Moreno-Noguer, F. & Fua, P. An accurate o(n) solution to the Pnp problem. *Int. J. Comput. Vis.* **81**(2), 155. https://doi.org/10.1007/s11263-008-0152-6 (2009).
17. Gao, Quankai, Wang, Fudong, Xue, Nan, Jin-Gang, Yu. & Xia, Gui-Song. Deep graph matching under quadratic constraint. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR46437.2021.00503 (2021).
18. Bronstein, Michael M. & Bronst, Alexander M. Data fusion through cross-modality metric learning using similarity-sensitive hashing. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2010.5539928 (2010).
19. Kumar, S. & Udupa, R. Learning hash functions for cross-view similarity search. *International Joint Conference on Artificial Intelligence* (2011).
20. Weiss, Y., Torralba, A., Fergus, R, Spectral hashing. In *Annual Conference on Neural Information Processing Systems* (2008).
21. Zhou, Yu. et al. Discriminative coupled dictionary hashing for fast cross-media retrieval. *ACM Special Interest Group on Information Retrieval*. https://doi.org/10.1145/2600428.2609563 (2014).
22. Tang, Xu. et al. Efficient multiscale transformer and cross-level attention learning for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–15. https://doi.org/10.1109/TGRS.2022.3194505 (2022).
23. Trejo-Lopez, J. A., Yachnis, A. T. & Prokop, S. Neuropathology of Alzheimer's disease. *Neurotherapeutics* (2023).
24. Mintun, M. A., Duggan Evans, A. C. & Lowe, C. Donanemab in early Alzheimer's disease. *N. Engl. J. Med.* (2021).
25. Van Dyck, C. H. et al. Lecanemab in early Alzheimer's disease. *N. Engl. J. Med.* (2023).
26. Dubois, B., Villain, N., Frisoni, G. B. & Rabinovici, G. D. Clinical diagnosis of Alzheimer's disease: recommendations of the International Working Group. *Lancet* (2021).
27. Porsteinsson, A. P., Isaacson, R. S. & Knox, S. Diagnosis of early Alzheimer's disease: Clinical practice in 2021 (Springer, 2021).
28. Srivastava, S., Ahmad, R. & Khare, S. K. Alzheimer's disease and its treatment by different approaches: A review. *Eur. J. Med. Chem.* (2021).
29. Tatulian, S. A. Challenges and hopes for Alzheimer's disease. Drug Discov. Today (2022).
30. Dreylikh, Z. & Karaman, R. Comprehensive review on Alzheimer's disease: Causes and treatment. *Molecules* (2020).
31. Vaz, F. & Silvestre, S. Alzheimer's disease: Recent treatment strategies. *Eur. J. Pharmacol.* (2020).
32. van Ede, Freek, Chekroud, Sammi R. & Nobre, Anna C. Human gaze tracks the focusing of attention within the internal space of visual working memory. *J. Vis.* https://doi.org/10.1167/19.10.133b (2019).
33. Cheng, M.-M. et al. BING: Binarized normed gradients for objectness estimation at 300fps. *Comput. Vis. Media* **5**(1), 3–20. https://doi.org/10.1109/CVPR.2014.414 (2019).
34. Diestel, Reinhard & Theory, Graph. *Springer-Velag* https://doi.org/10.1007/978-3-662-53622-3 (2005).
35. Zhang, Luming et al. Bioinspired scene classification by deep active learning with remote sensing applications. *IEEE Trans. Cybern.* **52**(7), 5682–5694. https://doi.org/10.1109/TCYB.2020.2981480 (2021).
36. Koren, Yehuda, Bell, Robert M. & Volinsky, Chris. Matrix factorization techniques for recommender systems. *Computer* **42**(8), 30–37. https://doi.org/10.1109/MC.2009.263 (2009).
37. Kyrkou, Christos & Theocharides, Theocharis. EmergencyNet: efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.* **13**, 1687–1699. https://doi.org/10.1109/JSTARS.2020.2969809 (2020).
38. Kyrkou, C. & Theocharides, T. Deep-learning-based aerial image classification for emergency response applications using unmanned aerial vehicles. *Comput. Vis. Pattern Recogn. Workshops*. https://doi.org/10.1109/CVPRW.2019.00077 (2019).
39. Hua, Y. et al. Learning multi-label aerial image classification under label noise: A regularization approach using word embeddings. *Int. Geosci. Remote Sens. Symposium*. https://doi.org/10.1109/IGARSS39084.2020.9324069 (2020).
40. Hua, Y. et al. Multi-label aerial image classification using a bidirectional class-wise attention network. *Jt. Urban Remote Sens. Event.* https://doi.org/10.1109/JURSE.2019.8808940 (2019).
41. Mark, D. Satellite image classification with deep learning. *IEEE Appl. Imagery Pattern Recogn. Workshop*. https://doi.org/10.1109/AIPR.2017.8457969 (2017).
42. Sun, Huiming et al. Convolutional neural networks based remote sensing scene classification under clear and cloudy environments. *Int. Conf. Comput. Vis.* https://doi.org/10.1109/ICCVW54120.2021.00085 (2021).
43. Song, Shaoyue et al. Domain adaptation for convolutional neural networks-based remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **16**(8), 1324–1328. https://doi.org/10.1109/LGRS.2019.2896411 (2019).
44. He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing & Sun, Jian. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**(9), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824 (2015).
45. Lee, Kuang-Huei., He, Xiaodong, Zhang, Lei & Yang, Linjun. CleanNet: Transfer learning for scalable image classifier training with label noise. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2018.00571 (2018).
46. Wang, Yaming, Morariu, Vlad I. & Davis, Larry S. Learning a discriminative filter bank within a CNN for fine-grained recognition. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2018.00436 (2018).
47. Caglayan, Ali. Exploiting multi-layer features using a CNN-RNN approach for RGB-D object recognition. In *European Conference on Computer Vision Workshops*. https://doi.org/10.1007/978-3-030-11015-4-51 (2018).
48. Chen, Zhao-Min., Wei, Xiu-Shen., Wang, Peng & Guo, Yanwen. Multi-label image recognition with graph convolutional networks. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2019.00532 (2019).
49. Chen, Tianshui, Muxin, Xu., Hui, Xiaolu, Hefeng, Wu. & Lin, Liang. Learning semantic-specific graph representation for multi-label image recognition. *Int. Conf. Comput. Vis.* https://doi.org/10.1109/ICCV.2019.00061 (2019).
50. Lanchantin, J. et al. General multi-label image classification with transformers. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR46437.2021.01621 (2021).
51. Mesnil, Grégoire. et al. Unsupervised learning of semantics of object detections for scene categorizations. *Adv. Intell. Syst. Comput.* https://doi.org/10.1007/978-3-319-12610-4-13 (2015).
52. Miao, Wang, Geng, Jie & Jiang, Wen. Multigranularity decoupling network with pseudolabel selection for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–13. https://doi.org/10.1109/TGRS.2023.3244565 (2023).
53. Chen, Junsong, Yi, Jizheng, Chen, Aibin & Jin, Ze. EFCOMFF-Net: A multiscale feature fusion architecture with enhanced feature correlation for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **61**, 1–17. https://doi.org/10.1109/TGRS.2023.3255211 (2023).
54. Xie, Weiying et al. Co-compression via superior gene for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens,* **61**, 1–12. https://doi.org/10.1109/TGRS.2023.3247872 (2023).
55. Shen, Junge, Cao, Bin, Zhang, Chi, Wang, Ruxin & Wang, Qi. Remote sensing scene classification based on attention-enabled progressively searching. *IEEE Trans. Geosci. and Remote Sens.* **60**, 1–13. https://doi.org/10.1109/TGRS.2022.3186588 (2022).
56. Lv, Pengyuan, Wenjun, Wu., Zhong, Yanfei, Fang, Du. & Zhang, Liangpei. SCViT: A spatial-channel feature preserving vision transformer for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **60**, 1–12. https://doi.org/10.1109/TGRS.2022.3157671 (2022).
57. He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing & Sun, Jian. Deep residual learning for image recognition. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2016.90 (2016).
58. Deng, Jia et al. ImageNet: A large-scale hierarchical image database. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2009.5206848 (2009).

59. Li, L.-J. et al. A high-level image representation for scene classification and semantic feature sparsification. In *Annual Conference on Neural Information Processing Systems* (2010).
60. Herranz, Luis, Jiang, Shuqiang & Li, Xiangyang. Scene recognition with CNNs: Objects, scales, and dataset bias. *Comput. Vis. Pattern Recogn.* https://doi.org/10.1109/CVPR.2016.68 (2016).
61. Li, Yunsheng, Dixit, Mandar & Vasconcelos, Nuno. Deep scene image classification with the MFAFVNet. *Int. Conf. Comput. Vis.* https://doi.org/10.1109/ICCV.2017.613 (2017).
62. Xiao, Jianxiong, Ehinger, Krista A., Hays, James, Torralba, Antonio & Oliva, Aude. SUN database: Exploring a large collection of scene categories. *Int. J. Comput. Vis.* **119**(1), 3–22. https://doi.org/10.1007/s11263-014-0748-y (2016).
63. Alexe, Bogdan, Deselaers, Thomas & Ferrari, Vittorio. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**(11), 2189–2202. https://doi.org/10.1109/TPAMI.2012.28 (2012).
64. Tuset, Jordi Pont, Arbelaez, Pablo, Barron, Jonathan T., MarquÃs, Ferran & Malik, Jitendra. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(1), 128–140. https://doi.org/10.1109/TPAMI.2016.2537320 (2017).
65. Wilms, Christian & Frintrop, Simone. AttentionMask: Attentive, efficient object proposal generation focusing on small objects. *Asian Conf. Comput. Vis.* https://doi.org/10.1007/978-3-030-20890-5-43 (2018).
66. Christopher, J. C. B., et al. Learning to rank using gradient descent. In *International Conference on Machine Learning*. https://doi.org/10.1145/1102351.1102363 (2005).
67. Bin, Xu. et al. EMR: A scalable graph-based ranking model for content-based image retrieval. *IEEE Trans. Knowl. Data Eng.* **27**(1), 102–114. https://doi.org/10.1109/TKDE.2013.70 (2015).

Author contributions

Han Wang and Jun-He Liu contributed to the basic idea and theoretical derivations. Yi Yang contributed to the equation check and main idea discussion.

Funding

This study was funded by the Henan Province key research and development project (231111320300), the Henan Province Science and Technology Research Project (242102110140), the National Scientific Research Project Cultivation Fund of HuangHuai University (XKPY-2023002), Research Center for the Improvement and Cultivation of Bulk Edible Mushroom in Henan Province.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.-H.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024