IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 1 April 2024, accepted 3 May 2024, date of publication 8 May 2024, date of current version 20 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3397998

RESEARCH ARTICLE

Aggregation-Based Perceptual Deep Model for Scenic Image Recognition

YUAN HONG, YANG XU, AND MU HU
Department of Orthopedics, Ruijin Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai 201801, China
Corresponding author: Mu Hu (1611018@tongji.edu.cn)

ABSTRACT In the artificial intelligence (AI) community, accurately interpreting the semantics of complex sceneries is a critical component across various systems. This paper introduces an effective pipeline that intelligently merges multi-channel perceptual visual features to identify scenic images with intricate spatial layouts. Our focus is on developing a deep hierarchical model that proactively identifies where human gaze is directed in a scenery. Overall, our method includes three key modules. First, we employ the BING objectness descriptor for the swift and precise localization of objects or their elements across multiple scales within a scene. Meanwhile, an algorithm for the local-global fusion of features is formulated to represent each BING patch, integrating various low-level attributes from different channels. Second, to mimic the human process of identifying semantically or visually significant patches within a scenery, we employee an active learning algorithm to localize those scenic patches that are semantically or visually salient. They further constitute the so-called Gaze Shift Path (GSP). Finally, an aggregation-guided deep neural network is designed to calculate the deep GSP features, which are subsequently applied to a multi-label SVM to distinguish among various scenic categories. Empirical evaluations reveal that our method's categorization accuracy outperforms existing models on six generic scenic datasets by 2% $\sim$ 4.5%. Besides, we observe a higher stability of our method according to the repetitive experiments. Furthermore, our method demonstrates exceptional discriminative power on a specially compiled sports educational image collection, wherein the accuracy exceeds the second best performer by 8%. These results showed the huge potential to computationally discover human gaze behavior in different visual recognition tasks.

INDEX TERMS Perceptual, feature fusion, local-global, active learning, deep architecture.

I. INTRODUCTION

The accurate attribution of multiple labels to each scene plays a pivotal role in the architecture of contemporary artificial intelligence (AI) systems. This paper presents instances where such recognition is crucial: for instance, intelligent navigation systems require the computation of the shortest route from start to finish. This necessitates the integration of various scene-related attributes, including transportation network configurations, directional flows of streets, and the morphology of urban landscapes, to refine pathfinding algorithms. Furthermore, in the domain of public security frameworks, the extraction of diverse scene-aware characteristics, such as road markings and elevation changes, is fundamental to augment the real-time monitoring capabilities for pedestrians and vehicles. It is observed that vehicular accidents predominantly occur at intersections rather than on uninterrupted stretches of road. Through rapid and precise classification of scenic types, the deployment of multi-camera surveillance networks at critical junctions becomes feasible, allowing for the detailed observation of unusual vehicular and pedestrian activity. Overall, by enhancing the capability to swiftly and accurately identify distinct scenic categories, we facilitate the strategic placement of surveillance systems and the optimization of navigational algorithms, thereby significantly contributing to the safety and efficiency of urban infrastructure.

Within the scholarly domain, a multitude of algorithms for visual categorization and annotation tailored to scenic imagery of varying resolutions has been developed.

The associate editor coordinating the review of this manuscript and approving it for publication was Wenxin Liu.

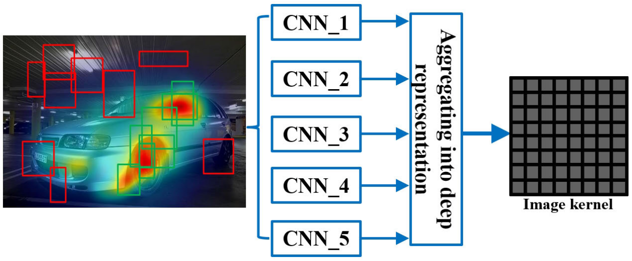

FIGURE 1. A Synopsis of Our Scenery Classification Approach through Perceptual Aggregation-based Deep Model (The red windows denote the BING object patches. The multiple BING patches are fed into the corresponding CNNs for calculating the deep features, which are further aggregated into an image-level deep features. Such image-level deep features are leveraged to train an image kernel machine for categorization.)

Prominent methodologies can be classified into three main categories: 1) Multiple Instance Learning (MIL) and CNN-guided region detection utilizing weak supervision techniques [49], [50]; 2) semantic-aware graph models designed for complex scene parsing [54], [55]; and 3) sophisticated hierarchical structures dedicated to the annotation of scenic photographs [51], [52], [53]. Despite these advancements, existing approaches often struggle to faithfully represent scenic images, attributed to several challenges:

- A plethora of visually appealing objects or their components are present within high-resolution scenic imagery, as shown in Fig. 1. Identifying semantic labels for these images necessitates a biologically-inspired algorithm that emulates human perception of salient regions. Crafting a deep learning solution capable of both identifying these regions and enhancing their visual representation presents multiple hurdles, including: i) delineating the gaze shifting path (GSP) as individuals sequentially focus on engaging image patches; ii) mitigating the effect of label noise inherent in large-scale training datasets; and iii) semantically integrating labels from the overall image down to specific patches within each scene;
- The accurate depiction of semantically or visually significant areas within a scene often relies on diverse low-level descriptors, each isolating scenic elements in a unique channel. Achieving a complementary fusion of these descriptors necessitates a method for intelligently determining the significance of each feature channel. However, formulating a mathematically tractable model for this purpose is challenging. Practical issues include: i) integrating local features from spatially adjacent regions within a scene; ii) maintaining global compositional integrity across various internal scenic areas; and iii) dynamically tuning channel weights to cater to distinct sets of scenic images.

Addressing the challenges outlined previously, we introduce an innovative scenery categorization framework that intricately and actively simulates human gaze dynamics. This approach entails representing each scenic image patch through the strategic amalgamation of various low-level features. Specifically, as shown in Fig. 2, our method contains three main modules. within a comprehensive collection of scenic images where labels may be compromised, we initially apply the popular Binarized Normed Gradients (BING) algorithm [58] to extract numerous object-centric patches across the dataset. Each patch is then characterized through a novel fusion algorithm for low-level features, which simultaneously encodes the local as well as the global topological structures of samples (module 1). Further advancing our methodology, we introduce a novel aggregation-based deep framework (as shown in Fig. 1) to emulate human gaze patterns in scene perception. This framework excels in calculating the Gaze Shift Path (GSP) and deriving a deep representation of GSPs (module 2). By leveraging the learned deep representations, we construct a kernelized machine. It is used to train a multi-label Support Vector Machine (SVM) tasked with the categorization of scenic imagery (module 3). Our empirical assessments, conducted across six public scenic datasets and a specially assembled sports education image collection, showed the overwhelming performance of the designed recognition pipeline.

Nevertheless, our approach may encounter a limitation in the form of discrepancies between generated GSPs and natural human gaze patterns. To overcome this challenge, we plan to conduct comprehensive user studies to compare our GSPs with actual human gaze data. The aim is to refine our Low-rank Active Learning (LRAL) algorithm to more accurately replicate human visual behavior, thereby improving the quality of architectural categorization results.

In summary, the innovations introduced in this research are twofold. Firstly, we develop an aggregation-based deep learning model capable of actively learning and precisely modeling human gaze behavior, while concurrently extracting gaze-guided visual features. Secondly, we implement a sophisticated feature fusion approach that dynamically assesses and integrates the significance of various feature channels for each scenic image patch, enhancing the accuracy and relevance of the extracted features.

II. RELATED WORK IN SCENE CATEGORIZATION

The field of computer vision has seen the advent of numerous deep learning models for scene categorization. Hierarchical Convolutional Neural Networks (CNNs), augmented with intricately devised deep structures, have demonstrated efficacy in recognizing scenes from vast image collections such as the well-known ImageNet dataset [41]. A significant advancement was introduced in [5], where researchers developed a massive-scale deep neural network utilizing a subset of ImageNet [41], achieving remarkable accuracy in categorization tasks. Despite their generalist design, CNNs trained on ImageNet have proven to bolster a variety of computer vision applications, including video parsing and anomaly detection. Over the last decade, enhancements to standard ImageNet-based CNNs have occurred in two primary dimensions. Firstly, efforts have been made to amass

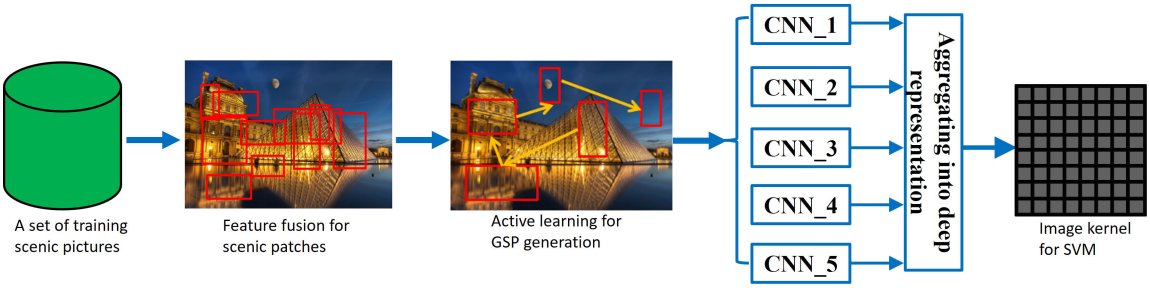
A set of training scenic pictures
Feature fusion for scenic patches
Active learning for GSP generation
CNN_1
CNN_2
CNN_3
CNN_4
CNN_5
Aggregating into deep representation
Image kernel for SVM

FIGURE 2. The pipeline of our scenery categorization.

extensive sample collections to bolster the training of multi-layer architectures. Techniques like selective search [42] have merged enumerative search principles with semantic-level annotation, facilitating the generation of category-agnostic patch samples conducive to deep learning. Secondly, the introduction of region-level CNNs (R-CNNs) [43] marked a strategic move towards sampling high-quality patch samples. Further contributions include [14], where researchers enriched CNN-based scene categorization by curating a vast collection of Internet-scale, scenery-related images. However, training deep visual models using entire scenic images or arbitrary patches has often been deemed inefficient. To address this, [45] utilized a pre-trained hierarchical CNN to identify local, representative scenic patches, thereby refining deep scene categorization learning. Moreover, [4] introduced a multi-task and multi-resolution algorithm for scene categorization that preserves intrinsic feature distribution with a manifold-based regularizer. In [7], a novel framework for scene semantic annotation was proposed, leveraging low-rank deep features to estimate category-specific posterior probabilities and employing a Markov probabilistic model to capture contextual features. The relationship between different deep layers was explored in [8], where an unlabeled learning model was developed to iteratively learn deep features based on the geometric attributes of scenes. Lastly, [47] innovatively integrated discriminative feature and weak label learning within a comprehensive scene analysis model, introducing a stacked discriminative sparsity autoencoder for computing advanced visual representations.

The field of computational vision has seen a plethora of models developed for the analysis of aerial imagery. For instance, a multi-modal learning technique for annotating high-resolution (HR) aerial images was introduced in [48], while a novel multi-attention-based algorithm for evaluating representations of aerial photos was proposed in [27]. These image-level visual models have been effectively applied to classify aerial images of varying resolutions. However, they face limitations in modeling low-resolution (LR) aerial images, primarily due to the challenge of accurately identifying small yet critical objects that appear blurred. To address the need for detecting discriminative objects across multiple scales, an efficient region-level modeling technique is essential. Such an approach enables the precise localization of diminutive objects within each LR aerial photograph. In the pursuit of robust face recognition, a group sparsity regularizer was designed in [62], introducing an upper-bounded function to enhance the sparsity-seeking capabilities of the l_1-norm, effectively mitigating bias and outlier impacts. Additionally, the challenge of incomplete multi-view clustering was tackled in [44] by transforming it into a task of upgrading incomplete similarity graphs and learning a complete tensor representation. For regional characterization of aerial images, a multi-layer deep learning model focused on identifying significant ground objects across scales was developed in [23]. Furthermore, a focal-loss-based deep learning model for precise vehicle localization within both LR and HR aerial photographs was formulated in [59]. A geographic object detection model, capable of intelligently extracting roads and intersections from HR aerial images, was presented in [61]. Lastly, an innovative visual detector combining feature engineering with soft-label calculations for aerial image analysis was proposed in [60].

In [16], the authors introduce a Self-Guided Separation Network for the classification of remote-sensing scenes. Uniquely, this approach leverages background information beyond the primary target in the image, employing a target-background separation strategy as a supportive tool for decision-making. This technique enhances the ability to differentiate between samples that share similar targets but have distinct backgrounds. Additionally, the network enhances the variety of feature focuses across its branches using contrastive regularization, boosting the distinction between target and background information. In [17], the researchers explore both the local and global structures of scene images, merging these insights to enhance scene recognition accuracy for both indoor and outdoor environments. In [18], Zhao and colleagues develop an advanced efficient multisample contrastive network that assimilates knowledge from multiple samples. This involves creating a dynamic dictionary, updated through momentum, to identify positive and negative sample pairs throughout the dataset. A contrastive module is then used to aggregate similarity and discriminative knowledge between samples, with the insights gained subsequently integrated into the main classifier through knowledge distillation. In [25], the authors introduced a methodical strategy to address the challenge of benchmarking in the context of saliency detection. They distinguished between saliency models, maps, and metrics by adopting principles from Bayesian decision theory. This approach defined a saliency model as a probabilistic framework for predicting fixation densities, while a saliency map was conceptualized as a prediction tailored to a specific metric, optimized to enhance expected performance based on the model's density predictions. Additionally, the authors formulated these optimal saliency maps tailored to several widely utilized saliency metrics. Many saliency models struggle to effectively capture the significant mutual information between an image's content and the locations of viewer fixations. This challenge is addressed through a transfer learning approach, utilizing the DeepGaze I model, which employs features learned from object recognition tasks to predict where viewers will focus. Building on this, the authors introduce a newer version, "DeepGaze II" [26], which transforms an image into the high-dimensional feature space provided by the VGG network. Following this transformation, a straightforward readout network is employed to generate a density prediction of fixation points. This readout network undergoes initial training on the SALICON dataset and is subsequently fine-tuned using the MIT1003 dataset, enhancing its predictive accuracy.

III. OUR ALGORITHM

A. INTEGRATING OBJECT-AWARE PATCHES INTO SCENE CATEGORIZATION

Research in visual cognition and psychology [56], [57] has consistently shown that human observers tend to focus on semantically or visually significant regions when viewing diverse sceneries. This suggests that only select discriminative regions are considered during visual processing. Recognizing the importance of mimicking this selective attention in scene categorization, we have developed a methodology that includes efficient detection of object-aware patches and a geometry-preserved deep active learning approach to identify semantically or visually crucial scenic patches, thereby emulating human visual perception. The human visual system is naturally drawn to semantically or visually critical objects or their components, such as vehicles and skyscrapers, which, along with their spatial distribution, significantly influence the perception of different scenes. To pinpoint objects or components likely to capture human attention, we utilize the BING [58] objectness metric to extract a collection of high-quality object-aware patches from various scenes. The BING approach offers three principal advantages that are crucial for our scene categorization model. Firstly, it demonstrates exceptional efficiency in detecting object patches with minimal computational demand. Secondly, the quality of Gaze Shift Path (GSP) extraction is significantly improved by identifying a superior set of object-level patches. Thirdly, BING's outstanding generalization capability across unseen object categories ensures the adaptability of our scene categorization model across diverse datasets, enhancing its utility in practical applications.

B. OPTIMAL INTEGRATION OF PATCH FEATURES

Through the systematic extraction of object patches using the BING [58] algorithm within scenic images, we successfully gather a collection of low-level features for each scenic patch. Based on this, a novel multi-channel feature fusion scenario is designed to intelligently incorporate the low-level visual features. The algorithm is designed using in a local-global fusion architecture, which has three advantageous attributes: 1) Preservation of the local distribution in the low-level feature space is encoded, as scenic patches practically share visual similarities with their spatial neighbors. 2) Maintenance of the global distribution within the low-level feature space is significant as well, as it represents the overall composition of the scenery. 3) Feature weights are adaptively tuned toward each particular set of scenic images, ensuring the fusion process is optimal toward each unique each scenic dataset. This algorithm not only enhances the descriptiveness of visual features extracted from each scenic patch but also ensures that both local and global scenic context are optimally captured, thereby facilitating the accuracy of scene categorization.

1) MAINTAINING PATCH LOCAL DISTRIBUTION

For this objective, we define x_j^i as the visual feature extracted from the j-th scenic patch within the i-th feature channel. Additionally, x_j^i and its L spatially neighboring features are collectively represented as $\mathbf{X}j^i = [x_j^i, xj1^i, \cdots, x_{jL}^i]$. Similarly, $\mathbf{Y}j^i = [y_j^i, yj1^i, \cdots, y_{jL}^i]$ denotes the outcomes of feature fusion for $\mathbf{X}_j^i$. By leveraging the above definitions, the objective of preserving the local distribution among L spatially adjacent scenic patches can be formulated as follows:

$$\arg\min_{\mathbf{Y}_j^i} \sum\nolimits_{l=1}^{L} ||y_j^i - y_{jl}^i||^2 (r_j^i)_l, \tag{1}$$

In this framework, r_j^i is defined as an M-dimensional vector that quantifies the correlation between the scenic patch x_j^i and its spatially neighboring patch. Specifically, the l-th element of r_j^i, denoted as $(r_j^i)l$, is calculated using the expression $\exp\left(-\frac{||x_j^i - xjl^i||^2}{t^2}\right)$, where t denotes the standard deviation of a Gaussian distribution.

Following some mathematical derivations, we can transform the above objective function into a matrix representation as follows:

$$\arg\min_{\mathbf{Y}_j^i} \mathrm{tr}(\mathbf{Y}_j^i \mathbf{B}_j^i (\mathbf{Y}_j^i)^T), \tag{2}$$

Within this formulation, the matrix $\mathbf{B} = [-\mathbf{e}_M^T, \mathbf{I}_M]^T \mathrm{diag}(r_j^i)$ $[-\mathbf{e}_M^T, \mathbf{I}_M]$ is defined in the space $\mathbb{R}^{(M+1)\times(M+1)}$. It is important to note that $\mathbf{e}_M = [1, \cdots, 1]$ is an M-dimensional vector consisting entirely of ones, $\mathbf{I}_M$ signifies an $M \times M$

identity matrix, and $\mathrm{diag}(r_j^i)$ is an $M \times M$ diagonal matrix with its diagonal entries being the elements of r_j^i.

From a mathematical perspective, the local optimization of the H features can be described as follows:

$$\arg\min_{\mathbf{Y}=\{\mathbf{Y}_j^i\}_{i=1}^H,\kappa} \sum_{i=1}^{H} \kappa_i \mathrm{tr}(\mathbf{Y}_j^i \mathbf{B}_j^i (\mathbf{Y}_j^i)^T), \tag{3}$$

In our model, κ_i quantifies the significance of each feature channel.

2) PATCH GLOBAL DISTRIBUTION

Building on our earlier discussion, integrating the global spatial structure of object-aware patches within a scenic image is crucial. We propose that $\mathbf{Y}_j^i = \mathbf{Y}\mathbf{A}_j^i$, where the matrix $\mathbf{A}_j^i \in \mathbb{R}^{N\times(M+1)}$ acts as a selector matrix. This setup indicates that scenic patches are locally correlated to the entire set of patches within the image, suggesting a distributed approach to understanding scene composition. Consequently, the objective function previously defined can be modified as follows:

$$\begin{aligned} &\arg\min_{\mathbf{Y}=\{\mathbf{Y}\}_{i=1}^H,\kappa} \sum_{i=1}^{H} \kappa_i \mathrm{tr}(\mathbf{Y}\mathbf{A}_j^i \mathbf{B}_j^i (\mathbf{A}_j^i)^T \mathbf{Y}^T) \\ &= \arg\min_{\mathbf{Y}=\{\mathbf{Y}\}_{i=1}^H,\kappa} \sum_{i=1}^{H} \kappa_i \mathrm{tr}(\mathbf{Y}\mathbf{D}^i \mathbf{Y}^T), \end{aligned} \tag{4}$$

We notice that $\mathbf{B} = [-\mathbf{e}_L^T, \mathbf{I}_L]^T \mathrm{diag}(r_j^i)[-\mathbf{e}_L^T, \mathbf{I}_L]$, by reorganizing (4), the following equation can be received:

$$\mathbf{D}^i = \mathbf{C}^i - \mathbf{S}^i, \tag{5}$$

In this model, $\mathbf{C}^i$ is defined as a diagonal matrix where each entry, $\mathbf{C}_{jj}^i$, is derived by summing over l as follows: $\mathbf{C}_{jj}^i = \sum_l [\mathbf{S}^i]_{jl}$. Here, $\mathbf{S}$ is an $N \times N$ matrix with elements $[\mathbf{S}^i]_{uv} = \exp(-\frac{||x_u - x_v||^2}{t^2})$, encapsulating the interaction between pairs of scenic patches. Herein, N counts scenic patches in each scenic image. Importantly, $\mathbf{C}^i$ serves as the unnormalized Laplacian matrix [21], playing a crucial role in our analysis.

To enhance computational efficiency, we apply a normalization procedure to $\mathbf{D}^i$, outlined as follows:

$$\mathbf{D}_n^i = (\mathbf{C}^i)^{-1/2}\mathbf{D}^i(\mathbf{C}^i)^{-1/2}, \tag{6}$$

In this context, $\mathbf{D}_n^i$ denotes the normalized version of $\mathbf{D}^i$.

Our approach to feature fusion, which transitions from local to global considerations, is encapsulated in the following objective function:

$$\begin{aligned} &\arg\min_{\mathbf{Y},\kappa} \sum_{j=1}^{H} \kappa_i \mathrm{tr}(\mathbf{Y}\mathbf{D}_n^j \mathbf{Y}^T), \\ &s.t., \ \mathbf{Y}\mathbf{Y}^T = \mathbf{I}, \sum_{j=1}^{H} \kappa_j = 1, \kappa_j > 0. \end{aligned} \tag{7}$$

Observations from minimizing equation (7) indicate that $\kappa_i = 1$, which leads to the selection of multiple highly informative features. However, employing a hard constraint in this context is considered sub-optimal since the goal is to utilize a combination of features for effective scenery categorization. To address this, we adopt a strategy as suggested in [24], specifically modifying the setup to $\kappa_i \leftarrow \kappa_i^o$ with $o > 1$. This modification implies that the ideal value of κ_i for multi-channel features needs dynamic adjustment. Theoretically, each feature channel should contribute distinctively to the composite feature, ensuring an optimal representation of each scenic patch.

C. AGGREGATION-BASED DEEP MODEL

Based on the patch-level fused feature, we leverage the geometry-preserved active learning algorithm [10] to selected multiple visually/semantically salient scenic patches inside each scenery, they are then seqentially linked to from a Gaze Shift Path (GSP). Upon deriving the GSP from each scenic imagery, we construct a sophisticated deep learning framework aimed at learning and integrating these paths into an kernel-induced feature vector, facilitating enhanced scenery classification. This framework is structured around two main elements: 1) An Adaptive Spatial Pooling (ASP)-enhanced CNN, tailored for detailed scene region analysis, and 2) A method for amalgamating regional hierarchical visual representations into a cohesive feature at image-level.

Element 1: In the community of image understanding, preserving the input size as well as the aspect ratio of each image has been acknowledged as crucial for accurately modeling scene spatial dynamics [11]. It has also been established that objects with arbitrary shapes convey richer semantic information about the scene compared to standard rectangular patches [9]. To address these insights, we refine the conventional five-layer CNN architecture [15] to accommodate inputs of varying shapes and sizes. This enhancement is achieved through the integration of an Adaptive Spatial Pooling (ASP) layer [11], which introduces the flexibility of adjusting pooling dimensions to suit input regions of any geometry, thereby preserving the integrity of scene representation.

Each deep CNN, as illustrated in Fig. 1, begins with processing a bunch of salient scenic patches identified via our adopted active learning technique. To enhance the generality of each patch, we introduce random jitter and apply horizontal/vertical flipping with a probability of 0.5. The architecture progresses through four stages, encompassing convolution, Adaptive Spatial Pooling (ASP), and local response normalization, culminating in a fully connected layer with 2048 units. Subsequently, the network divides into a fully connected layer with H units of 256 dimensions, representing H hidden topics related to scenery, such as "beach" and "forest". The utilization of shared lower layers helps in minimizing parameter count while leveraging the foundational CNN structure common across low-level features.

Element 2: As depicted in Fig. 1, for a GSP comprising a set of sequentially connected regions of arbitrary shapes, we extract an L-dimensional deep representation for a scenic patch utilizing the aforementioned regional CNN model. The above visual features are subsequently aggregated to form a comprehensive feature descriptor of the GSP.

We define $\Phi = \{\phi_i\}_{i\in[1,K]}$, where $\phi_i \in \mathbb{R}^M$ represents the each region's deep representation along the GSP, and $\mathcal{S}_m$ as the ensemble of the m-th feature dimension across all $\phi_i \in \Phi$, i.e., $\mathcal{S}_m = \{\psi_{mj}\}_{j\in[1,K]}$. The aggregation process employs a collection of statistical functions, $\Pi = \{\pi_u\}_{u\in[1,U]}$, including minimum, maximum, mean, and median, applied to the regional deep representations. The outcomes from Π are merged and then processed based on a fully-connected module, yielding an L-dimensional vector that encapsulates the deep representation of the GSP, thereby enhancing scene categorization through a methodologically grounded fusion of local to global visual features.

$$\mathcal{F}(\Psi) = \mathbf{P} \times (\oplus_{u=1}^{U} \oplus_{m=1}^{M} \pi_u(\mathcal{S}_m)), \tag{8}$$

where $\mathbf{P} \in \mathbb{R}^{L\times UM}$ denotes a matrix that includes deep aggregation layer's parameters. We set U to four as there exists four different operations in $\mathcal{S}$. Within this structure, $\mathbf{P} \in \mathbb{R}^{L\times UM}$ is the parameter matrix of the aggregation module, where U is set to four to match the four statistical functions in $\mathcal{S}$. The operator $\oplus$ represents vector concatenation, merging the UM-dimensional vectors $\phi_u(\mathcal{S}_m)$ into an extended vector.

1) TRAINING THE MULTI-LAYER AGGREGATION NETWORK

In our forward propagation stage, the output o_i from the i-th neuron in the statistic layer is calculated as $o_i = \sum_{a=1}^{K}\sum_{m=1}^{M} r_{am\to i}o'_{am}$, with $r_{am\to i}$ representing the "contribution" of neuron d_{km} to neuron i in the statistic layer. The error η_i received by the i-th neuron at this layer allows us to determine the back-propagated error η'_{km} for neuron d_{km}, calculated as $\eta'_{km} = \sum_i r_{km\to i}\eta_{km}$. The overall multi-layer deep model is optimized through standard error back-propagation [5], employing stochastic gradient descent for error minimization.

2) KERNEL-INDUCED FEATURE VECTOR

Given that each scenic picture is characterized by a GSP in $\mathbb{R}^2$, traditional classifiers such as SVMs, which require 1-D vector features, face a challenge in directly categorizing scenes based on these paths. To address this, we introduce a kernel machine that transforms the multidimensional paths into 1-D vectors.

The effectiveness of the image kernel-induced feature depends on calculating distances between scenic pictures based on their GSPs. For each scenic picture, its paths $\mathcal{P}^*$ are transformed into vectors $\vec{a} = [\alpha_1, \alpha_2, \cdots, \alpha_N]$, with each element defined as:

$$\alpha_i \propto \exp\left(-\frac{1}{j^2 \cdot T^2}\sum\nolimits_{j=1}^{T} d(y(\mathcal{P}_j^*), y(\mathcal{P}_j^i))\right), \tag{9}$$

In this formulation, $d(\cdot,\cdot)$ is used to represent the distance of pairwise vectors, where y representations the deep visual feature extracted in an GSP. The parameter N counts the training scenic images, while T refers to the quantity of regions along each path. Additionally, $\mathcal{P}_j^*$ and $\mathcal{P}_j^i$ correspond to the j-th regions within the paths $\mathcal{P}^*$ and $\mathcal{P}^i$, respectively.

Utilizing the feature vector derived as mentioned, we proceed to train a multi-class SVM [6] for scene categorization. Given R distinct scenery categories, our approach involves the training of C_R^2 binary SVM classifiers to distinguish between scenes from the p-th and q-th categories by establishing a specific binary SVM for each pair.

$$\max_{\beta\in\mathbb{R}^{N_{pq}}} \omega(\beta) = \sum\nolimits_{i=1}^{N_{pq}} \beta_i - \frac{1}{2}\sum\nolimits_{i=1}^{N_{pq}} \gamma_i\gamma_j l_i l_j k(\alpha_i, \alpha_j)$$
$$s.t. \quad 0 \le \gamma_i \le C, \sum\nolimits_{i=1}^{N_{pq}} \gamma_i l_i = 0, \tag{10}$$

In this scenario, $\gamma_i \in \mathbb{R}^N$ represents the deep feature for the i-th training scenic image, with l_i representing its class label (where $+1$ corresponds to the p-th category and -1 to the q-th category). The variable α describes the hyperplane that distinguishes between scenic images belonging to the p-th category and those in the q-th category. The parameter $C > 0$ is utilized to balance the complexity of the model against the proportion of scenic images that cannot be discriminated, while N_{pq} counts the training scenic images from either the p-th or the q-th category. In practice, given $\mathcal{R}$ distinct scenic categories in total, we will produce $(\mathcal{R}-1)\mathcal{R}/2$ binary SVMs to differentiate between the entire $\mathcal{R}$ categories.

IV. TESTING OUR METHOD

In this section, we assess the performance of our scene classification framework, which leverages the designed aggregation-based CNNs, across four distinct experimental setups. Initially, we outline the experimental design and introduce six benchmark datasets for scene classification. Following this, we engage in a comparative study against a range of both shallow and deep learning-based recognizers. We then investigate the influence of critical parameters within our model. Finally, we demonstrate the application of the deep GSP features extracted by our model to enhance the classification of education-related sports scenes.

A. DATASETS AND EXPERIMENTAL SETUP

Our categorization model is thoroughly tested across six varied scenic image collections, which encompass both established benchmarks and more contemporary datasets. Representative images from these experimental scenery collections are displayed in Fig.3. Among these, the two foundational datasets employed are Scene-15 [19] and MIT Indoor Scene-67 [20].

Our evaluation encompasses a broad spectrum of scenic image datasets, detailed as follows:

- Scene-15: This dataset includes 15 diverse categories, with 13 initially introduced by Li and Perona [22]. Each category contains between 200 to 400 scenic images, averaging a resolution of 320 × 250. The images are primarily sourced from COREL, individual collections, and Google.
- Scene-67: This collection features a comprehensive array of indoor scenes, aggregated from three primary

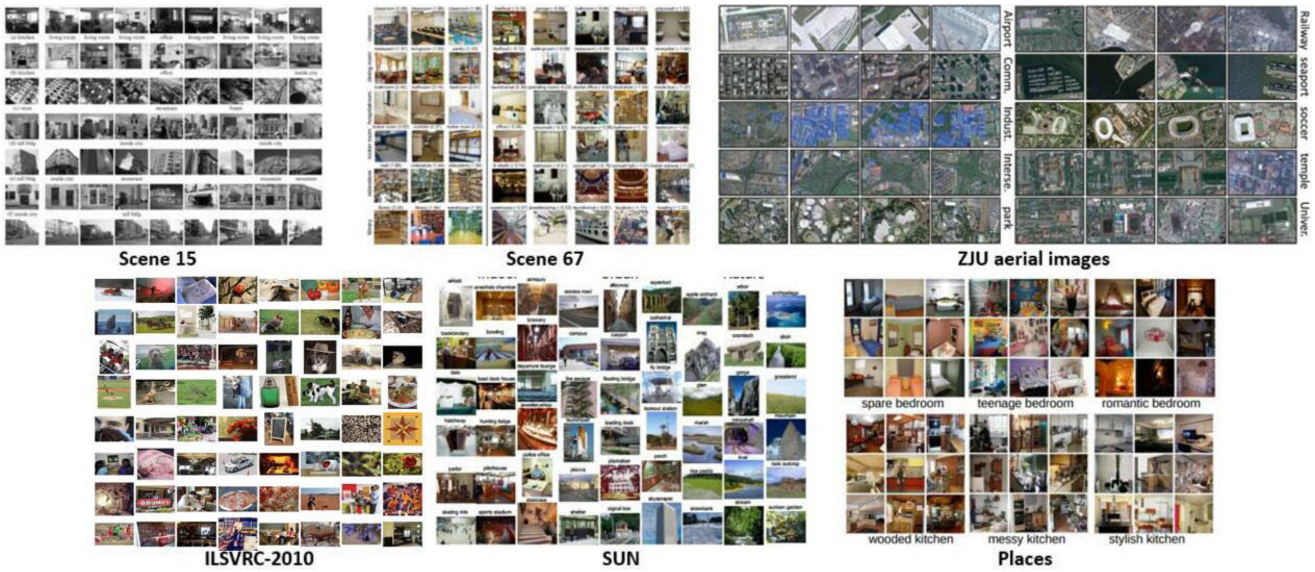

FIGURE 3. Sample images from the six scene datasets mentioned above.

sources: 1) Picasa and Altavista, 2) various photography sharing websites, and 3) LabelMe images.

- Additionally, we incorporate four more recent scenic image sets into our evaluation: ZJU Aerial Imagery [3], ILSVRC-2010 [41], SUN [46], Places [14].

Furthermore, we introduce a proprietary dataset, the Massive-Scale Sport Educational Images (MSEI), consisting of sports scenes utilized for educational purposes. This dataset is composed of approximately 920,000 images across nine sports categories: basketball, football, volleyball, outdoor golf, athletics, table tennis, rowing, baseball, and equestrian. A visual snapshot of this dataset is illustrated in Fig. 4, with detailed statistics provided in Table 1.

TABLE 1. Details of our sport educational image set.

Sport name	Training image #	Training image #
Basketball	83021	23121
Football	74343	25330
Volleyball	73021	29843
Golf	83243	24394
Athletics	65993	34220
Baseball	68321	21203
Tennis	73421	25436
Rowing	74355	24453
Equestrian	82103	20032

FIGURE 4. Sample images from our sport educational image collection.

Before delving into the comparative analysis with baseline algorithms, we detail the empirical configurations of our methodology: 1) Object Patches: Utilizing the BING [58] algorithm, we standardize the number of scenic patches to 1000 across all six scenic image datasets to ensure comprehensive localization of potential objects. 2) Spatial Neighbors: The number of spatial neighbors (L) for each patch is consistently set to five, facilitating a balanced local context assessment. 3) Low-Level Features: For each object patch, we incorporate three types of low-level features to capture comprehensive visual characteristics: a 16-dimensional color moment [63], a 64-dimensional Histogram of Oriented Gradients (HOG) [64], and a 160-dimensional combined edge and color histogram [13]. 4) GSP's Internal Regions: The quantity of internal regions within a Gaze Shifting Path (GSP), denoted by K, is established at five. This decision is based on empirical evidence suggesting that human attention typically focuses on up to five prominent regions within a scene. 5) Patch-Level Deep Feature: The dimensionality of our deep feature, extracted at the patch level, is standardized at 212, ensuring a detailed yet manageable representation for each patch. These settings aim to strike an optimal balance between capturing intricate details and ensuring computational efficiency, while aligning with the observed tendencies of human visual focus and the way objects are depicted in scenic imagery.

TABLE 2. **Averaged categorization accuracies on the compared models on the aforementioned data sets (we repeatedly experiment each baseline algorithm 10 times, based on which the average categorization accuracies are reported. the results showed the overwhelming performance of our perceptual scenery categorization model.)**

Data set	FWK	FTK	MRH	PM	LLC-SP	SC-SP	OB-SP	SV	SSC
Scene-15	72.1%	75.4%	67.2%	77.6%	81.3%	82.1%	77.1%	82.1%	87.4%
Scene-67	41.6%	41.8%	34.2%	44.5%	48.5%	47.7%	48.6%	47.3%	51.3%
ZJU Aerial	66.8%	68.3%	62.5%	73.3%	78.4%	78.1%	78.1%	78.3%	82.6%
ILSVRC-2010	32.1%	30.7%	27.4%	32.4%	38.4%	36.3%	37.2%	37.2%	38.4%
SUN397	15.3%	15.6%	14.2%	22.3%	39.3%	39.5%	38.0%	35.5%	40.2%
Places205	22.1%	22.2%	20.6%	27.5%	31.2%	32.3%	31.6%	31.3%	32.2%
MSEI	47.5%	48.2%	50.6%	47.3%	51.1%	54.1%	47.5%	51.3%	52.7%
Data set	IN-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	SP-S	SP-GBV	SP-LDA	Mesnil
Scene-15	83.1%	87.4%	87.3%	89.3%	92.3%	90.5%	86.2%	87.1%	86.4%
Scene-67	57.2%	68.1%	72.3%	68.4%	65.3%	76.2%	71.5%	72.1%	71.8%
ZJU Aerial	75.2%	79.1%	78.2%	81.0%	78.2%	81.2%	80.3%	81.1%	80.6%
ILSVRC-2010	35.7%	38.4%	40.4%	40.6%	41.3%	41.4%	40.4%	40.5%	40.5%
SUN397	48.1%	47.2%	51.2%	48.7%	52.1%	51.7%	50.5%	51.0%	50.5%
Places205	40.7%	43.7%	44.8%	45.9%	48.3%	49.9%	48.4%	48.1%	49.4%
MSEI	52.4%	50.5%	51.4%	53.5%	55.7%	52.6%	58.1%	61.3%	62.1%
Data set	Xiao	Cong	Fast R-CNN	Faster R-CNN	Ours	[16]	[17]	[18]	
Scene-15	82.8%	86.6%	90.2%	91.2%	**93.8%**	87.5%	84.3%	88.6%	
Scene-67	71.3%	72.1%	71.5%	74.7%	**77.5%**	72.3%	73.4%	77.4%	
ZJU Aerial	81.1%	80.1%	78.6	81.2%	**84.9%**	82.5%	81.0%	83.2%	
ILSVRC-2010	40.5%	41.1%	40.8%	41.1%	**44.2%**	44.5%	42.6%	43.8%	
SUN397	50.4%	51.2%	52.2%	52.0%	**58.3%**	53.2%	54.3%	51.8%	
Places205	49.3%	48.2%	48.3%	49.3%	**57.1%**	50.6%	51.3%	49.6%	
MSEI	59.7%	61.5%	62.5%	64.7%	**72.9%**	63.6%	61.2%	64.4%	

B. COMPARISON WITH OTHER RECOGNITION MODELS

1) SCENERY CATEGORIZATION TASK EVALUATION

Our perception-guided scenery categorization model is benchmarked against four widely recognized shallow classification algorithms: 1) Fixed-Length Walk Kernel (FWK) and its Tree Kernel version (FTK) [28]. 2) Multi-Resolution Histogram (MRH) [34]. 3) Kernel Machine Learning by Spatial Pyramid (SP), with variants including LLC-SP [29], SC-SP [30], and OB-SP [31]. 4) Image Representation by Super Vector (SV) [32] and Supervised Image Coding (SSC) [33].

In our comparative analysis, algorithm configurations are standardized. FWK and FTK lengths are adjusted between two and ten. MRH employs RBF-based smoothing with 12 gray scales for scenic image preprocessing. For SP and its variants, SIFT descriptors are extracted from 16 × 16 grids across all training images, followed by the construction of a 400-sized codebook via k-means clustering.

Considering the success of multi-layer recognition models, we extend our analysis to include several deep learning-based scene recognition models: ImageNet CNN (IN-CNN) [5], R-CNN [43], Meta Object CNN (M-CNN) [45], Deep Mining CNN (DM-CNN) [35], and Spatial Pyramid Pooling CNN (SPP-CNN) [36]. Except for [45], the source codes for these deep models are available, allowing for direct evaluation with unchanged parameters. For [45], we began by extracting 192 to 384 region proposals per image set using MCG [37], fixing the visual representation dimension at 4096 based on the FC7 layer outputs from a combined CNN [14]. Additionally, 400 superpixels per scene are generated using SLIC [2], optimized via SP-LDA or by selecting 120 visually significant patches as identified by GBV [1] (SP-GBV).

Our method integrates multiple low-level features, with the active learning framework identifying semantically or visually significant superpixels (GSPs) to form Graph-based Superpixels. These GSPs contribute to the kernel machine for scene classification. Performance comparisons between our BING-based rectangular patches and superpixels are presented in Tables 2 and 3, revealing the superior descriptive power of BING-guided rectangular patches over superpixels. Additionally, we conduct comparisons with recent scenery categorization models by Mesnil et al. [38], Xiao et al. [39], and Cong et al. [40], showcasing the robustness and efficacy of our approach.

Reviewing the data in Tables 2 and 3, we conduct a detailed quantitative analysis comparing the performance of the previously mentioned deep learning and traditional visual recognition models. Each experiment is replicated 20 times to ensure reliability, with the resulting standard deviations also reported. Our findings indicate that our approach not only achieves the highest classification accuracy but also exhibits superior stability across evaluations. Notably, within our specially curated Massive-Scale Sport Educational Images (MSEI) dataset, the aggregation-based deep architecture distinctly outperforms its closest competitor by more than 8% in categorization precision, underscoring the effectiveness and adaptability of our model in handling complex visual categorization tasks.

C. PARAMETER OPTIMIZATION IN PERCEPTION-GUIDED DEEP RECOGNITION

Our perception-guided deep recognition model is characterized by several critical parameters that significantly influence the accuracy of scenery categorization. In this analysis, we evaluate the impact of these parameters on

TABLE 3. Standard derivations on the compared models on the aforementioned data sets (we repeatedly experiment each baseline algorithm 10 times, based on which the standard derivations are reported.)

Data set	FWK	FTK	MRH	SP	LLC-SP	SC-SP	OB-SP	SV	SSC
Scene-15	0.013	0.012	0.012	0.015	0.016	0.017	0.011	0.013	0.012
Scene-67	0.014	0.013	0.015	0.014	0.014	0.013	0.013	0.014	0.014
ZJU Aerial	0.014	0.015	0.016	0.015	0.016	0.015	0.014	0.013	0.014
ILSVRC-2010	0.014	0.013	0.013	0.013	0.014	0.013	0.012	0.013	0.014
SUN397	0.012	0.014	0.014	0.013	0.014	0.015	0.016	0.013	0.015
Places205	0.013	0.014	0.015	0.014	0.016	0.014	0.016	0.015	0.017
MSEI	0.015	0.011	0.015	0.013	0.009	0.012	0.013	0.014	0.013
Data set	IN-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	SP-S	SP-GBVS	SP-LDA	Mesnil
Scene-15	0.016	0.013	0.014	0.014	0.015	0.013	0.014	0.013	0.015
Scene-67	0.013	0.015	0.013	0.013	0.014	0.013	0.015	0.013	0.012
ZJU Aerial	0.013	0.014	0.015	0.014	0.013	0.014	0.013	0.016	0.014
ILSVRC-2010	0.015	0.013	0.014	0.013	0.015	0.018	0.013	0.015	0.012
SUN397	0.013	0.014	0.015	0.012	0.014	0.012	0.014	0.014	0.015
Places205	0.012	0.014	0.012	0.013	0.013	0.014	0.013	0.012	0.013
MSEI	0.014	0.012	0.014	0.012	0.014	0.015	0.012	0.017	0.015
Data set	Xiao	Cong	Fast R-CNN	Faster R-CNN	Ours	[16]	[17]	[18]	
Scene-15	0.012	0.014	0.013	0.014	0.008	0.009	0.007	0.008	
Scene-67	0.017	0.012	0.013	0.013	0.008	0.006	0.007	0.009	
ZJU Aerial	0.014	0.013	0.014	0.012	0.008	0.0011	0.010	0.007	
ILSVRC-2010	0.013	0.013	0.014	0.011	0.011	0.007	0.007	0.009	
SUN397	0.012	0.013	0.014	0.013	0.010	0.008	0.007	0.006	
Places205	0.013	0.012	0.014	0.012	0.009	0.007	0.009	0.008	
MSEI	0.014	0.011	0.015	0.014	0.008	0.006	0.008	0.006j	

model performance and propose optimal configurations based on empirical results. Specifically, we examine three key parameters: i) L, the number of neighbors considered for reconstructing an object patch in active learning; and ii) K, the number of object patches selected within a Gaze Shifting Path (GSP). Due to the extensive computational demands associated with larger datasets, this parameter tuning exercise is carried out using the Scene-15 dataset [19], offering a practical balance between thorough assessment and computational feasibility.

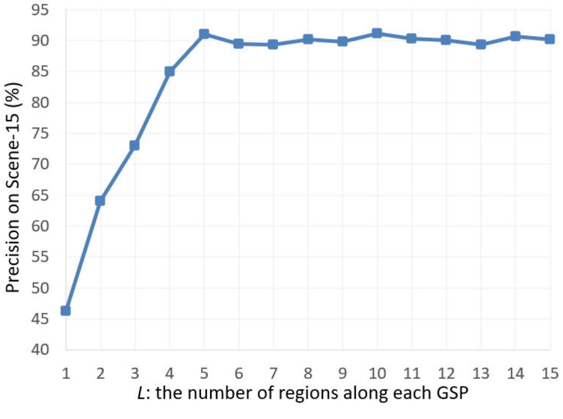

FIGURE 5. Categorization precision by adjusting L.

Next, L represents the count of neighboring patches used for reconstructing a scenic patch, a crucial aspect in preserving the locality of object patches within our feature fusion process. We incrementally adjust L from one to 15 and document the average recognition accuracies across 15 scenery categories. As depicted in Fig. 5, recognition precision improves and peaks when L is set between three to five, after which accuracy begins to decline. This indicates that utilizing three to five spatially adjacent scenic patches optimally reconstructs each scenery. Specifically, in our analysis of the Scene-15 dataset, we observed that scenic patches typically have three to five spatial neighbors. This observation suggests that a range of three to five neighbors for each target patch suffices for effective reconstruction. Furthermore, Fig. 6 demonstrates that including too many potentially irrelevant scenic patches not only reduces reconstruction accuracy but also increases computational time.

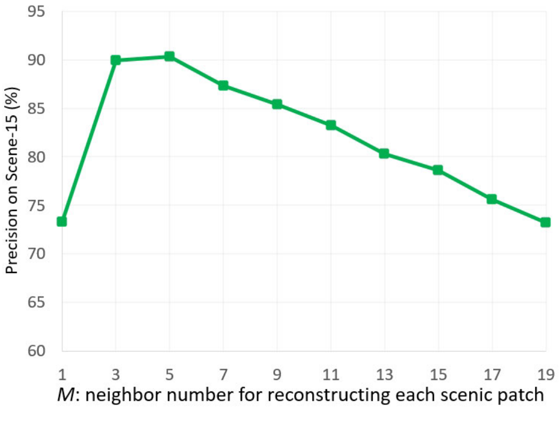

FIGURE 6. Categorization precision by adjusting M.

V. SUMMARY

The ability to accurately classify scenes into distinct categories holds considerable importance across a spectrum of artificial intelligence (AI) applications. In this study, we have introduced an innovative approach, termed aggregation-based CNNs, which adeptly learn a descriptive image kernel by simultaneously discovering and representing human gaze dynamics. Beginning with a comprehensive dataset of scenic images, our methodology employs a local-to-global feature fusion strategy to integrate various features for each region's characterization. The active learning technique is then applied to pinpoint both visually and semantically significant regions within each scene, thereby constructing a Gaze Shifting Path (GSP) and deriving its deep representation. These deep GSP features are subsequently transformed into a kernelized vector format, facilitating effective scene categorization. Our extensive experimental evaluations validate the robustness and efficiency of this biologically-inspired deep learning pipeline in scene categorization tasks.

REFERENCES

[1] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. NIPS*, 2006, pp. 545–553.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

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

[4] X. Lu, X. Li, and L. Mou, "Semi-supervised multitask learning for scene recognition," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1967–1976, Sep. 2015.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–9.

[6] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[7] X. Li, L. Mou, and X. Lu, "Scene parsing from an MAP perspective," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1876–1886, Sep. 2015.

[8] Y. Yuan, L. Mou, and X. Lu, "Scene recognition by manifold regularized deep learning architecture," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 10, pp. 2222–2233, Oct. 2015.

[9] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. ACM MM*, 2010.

[10] 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.

[11] L. Mai, H. Jin, and F. Liu, "Composition-preserving deep photo aesthetics assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 497–506.

[12] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010.

[13] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proc. ACM Workshops Multimedia*, Nov. 2000.

[14] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014.

[15] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. D. Bourdev, "PANDA: Pose aligned networks for deep attribute modeling," in *Proc. CVPR*, 2014.

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

[17] A. Khan, A. G. Chefranov, and H. Demirel, "Building discriminative features of scene recognition using multi-stages of inception-ResNet-v2," *Appl. Intell.*, vol. 53, no. 15, pp. 18431–18449, 2023.

[18] Y. Zhao, J. Liu, J. Yang, and Z. Wu, "EMSCNet: Efficient multisample contrastive network for remote sensing image scene classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5605814.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 1–8.

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

[21] U. von Luxburg, "A tutorial on spectral clustering," Max Planck Inst. Biol. Cybern., Germany, Tech. Rep. TR-149, 2006.

[22] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. CVPR*, 2005, pp. 524–531.

[23] C. Wang, X. Bai, S. Wang, J. Zhou, and P. Ren, "Multiscale visual attention networks for object detection in VHR remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 310–314, Feb. 2019.

[24] M. Wang, X.-S. Hua, X. Yuan, Y. Song, and L.-R. Dai, "Optimizing multi-graph learning: Towards a unified video annotation scheme," in *Proc. ACM Multimedia*, 2007, pp. 862–871.

[25] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proc. ECCV*, vol. 16, 2018, pp. 798–814.

[26] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," 2016, *arXiv:1610.01563*.

[27] W. Cai and Z. Wei, "Remote sensing image classification based on a cross-attention mechanism and graph convolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[28] Z. Harchaoui and F. Bach, "Image classification with segmentation graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007.

[29] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

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

[31] 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. NIPS*, 2010, pp. 1–9.

[32] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," in *Proc. ECCV*, 2010, pp. 1–14.

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

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

[35] Y. Li, L. Liu, C. Shen, and A. van den Hengel, "Mid-level deep pattern mining," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 971–980.

[36] 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.

[37] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 328–335.

[38] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent, "Unsupervised learning of semantics of object detections for scene categorizations," in *Proc. PRAM*, 2015, pp. 209–224.

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

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

[41] 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.*, Jun. 2009, pp. 248–255.

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

[43] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014.

[44] C. Zhang, H. Li, W. Lv, Z. Huang, Y. Gao, and C. Chen, "Enhanced tensor low-rank and sparse representation recovery for incomplete multi-view clustering," in *Proc. AAAI*, 2023.

[45] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in *Proc. ICCV*, 2015.

[46] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

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

[48] D. Hong, L. Gao, N. Yokoya, J. Yao, J. Chanussot, Q. Du, and B. Zhang, ''More diverse means better: Multimodal deep learning meets remote-sensing imagery classification,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 5, pp. 4340–4354, May 2021.

[49] S. Zhou, J. Irvin, Z. Wang, E. Zhang, J. Aljubran, W. Deadrick, R. Rajagopal, and A. Ng, ''DeepWind: Weakly supervised localization of wind turbines in satellite imagery,'' in *Proc. CVPR*, 2009.

[50] L. Cao, F. Luo, L. Chen, Y. Sheng, H. Wang, C. Wang, and R. Ji, ''Weakly supervised vehicle detection in satellite images via multi-instance discriminative learning,'' *Pattern Recognit.*, vol. 64, pp. 417–424, Apr. 2017.

[51] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4095–4104.

[52] 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. Workshops (CVPRW)*, Jun. 2016, pp. 680–688.

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

[54] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, ''Joint inference of groups, events and human roles in aerial videos,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4576–4584.

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

[56] J. M. Wolfe and T. S. Horowitz, ''What attributes guide the deployment of visual attention and how do they do it?'' *Nature Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, Jun. 2004.

[57] N. D. B. Bruce and J. K. Tsotsos, ''Saliency, attention, and visual search: An information theoretic approach,'' *J. Vis.*, vol. 9, no. 3, p. 5, Mar. 2009.

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

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

[60] Y. Yu, X. Yang, J. Li, and X. Gao, ''Object detection for aerial images with feature enhancement and soft label assignment,'' *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 34.

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

[62] C. Zhang, H. Li, C. Chen, Y. Qian, and X. Zhou, ''Enhanced group sparse regularized nonconvex regression for face recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2438–2452, May 2022.

[63] M. Stricker and M. Orengo, ''Similarity of color images,'' in *Proc. Storage Retr. Image Video Databases*, 1995.

[64] N. Dalal and B. Triggs, ''Histograms of oriented gradients for human detection,'' in *Proc. CVPR*, 2005.

YUAN HONG is currently a Faculty Member with Shanghai Jiao Tong University School of Medicine. His research interests include artificial intelligence, media understanding, computer vision, and machine learning.

YANG XU is currently a Faculty Member with Shanghai Jiao Tong University School of Medicine. His research interests include artificial intelligence and machine learning.

MU HU is currently with Shanghai Jiao Tong University School of Medicine. His research interests include multimedia and image understanding.