IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 February 2024, accepted 11 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3377657

RESEARCH ARTICLE

Perceptual Visual Feature Learning With Applications in Sports Educational Image Understanding

TENGSHENG LIU[1] **AND MINGHUI XU**[2]

[1]Department of Physical Education, Wuhan Institute of Technology, Wuhan 430070, China
[2]Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua Polytechnic, Jinhua 321017, China

Corresponding author: Tengsheng Liu (15090601@wit.edu.cn)

ABSTRACT Effectively understanding the semantics of sophisticated sceneries is a key module in plenty of artificial intelligence (AI) systems. In this article, we optimally fuse multi-channel perceptual visual features for recognizing scenic pictures with complex spatial configurations, focusing on formulating a deep hierarchical model to actively discover human gaze allocation. In detail, to uncover semantically/visually important patches within each scenery, we utilize the BING objectness descriptor to rapidly and accurately localize multi-scale objects or their components. Subsequently, a local-global feature fusion scenario is proposed to dynamically combine the multiple low-level features from multiple scenic patches. To simulate how humans perceiving semantically/visually important scenic patches, we design a robust deep active learning (RDAL) paradigm that sequentially derives gaze shift path (GSP) and hierarchically learns deep GSP features in a unified architecture. Notably, the key advantage of RDAL is the high tolerance of label noise by adding an elaborately-designed sparse penalty. That is, the contaminated and redundant deep GSP features can be implicitly abandoned. Finally, the refined deep GSP features are integrated into a multi-label SVM for recognizing sceneries of different categories. Empirical comparisons showed that: 1) our method performs competitively on six generic scenery set (average accuracy $2\% \sim 4.3\%$ higher than the second best performer), and 2) our deep GSP feature is particularly discriminative to our compiled sport educational image set (average accuracy 7.7% higher than the second best performer).

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

I. INTRODUCTION

Successfully recognizing the multiple labels belonging to each scenery is an essential component in plenty of modern AI infrastructures. Here we introduce some examples, for intelligent navigation, it is necessary to calculate the shortest path between the origin and the destination. In practice, we want multiple scenery-related features,such the transportation network topology, street direction, and urban terrain, to optimize the calculation. Additionally, in the existing public security systems, it is standard to extract different scene-aware features ,*e.g.*, road annotations and gradients, to enhance the real-time tracking of pedestrians and vehicles. Generally, car crashes are highly probably occurred at street intersections, while least likely occurred on flat roads. By fast and accurately identifying various scenic categories, in practice, we can install a multi-camera surveillance system near the road intersection for precisely inspecting abnormal vehicles and pedestrian behaviors.

In the literature, dozens of visual categorization/annotation algorithms were proposed for describing scenic imageries have various resolutions. Well-known models can be categorized as: 1) MIL (multiple instance learning)/CNN-guided region localization by leveraging weak supervision [41], [42]; 2) semantically-aware graph models for parsing [46], [47]; and 3) well-made hierarchical architectures for annotating

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

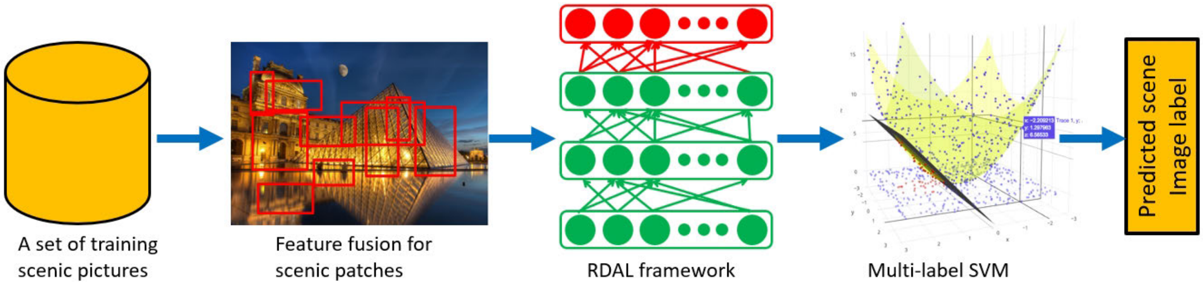

FIGURE 1. An overview of our designed scenery categorization by perceptual feature fusion.

scenic photos [43], [44], [45]. However, to our best knowledge, the current techniques fail to accurately represent scenic pictures because of the following factors:

- Practically, we notice that there exists many attractive objects or their parts in each high-resolution scenery, as exemplified in Fig. 1. To discover those semantic labels for each scenic picture, a biologically-inspired algorithm is required to simulate human perceiving the the visually prominent regions. Practically, building a deep learning algorithm to jointly obtain the visually prominent regions and refine the visual representation to the above regions is difficult. Some possible challenges are: i) computing the path when human beings sequentially allocating their gazes onto the attractive image patches (such as the GSPs as presented in Fig. 1), 2) avoiding the inherent noisy labels from the massive-scale training samples, and 3) semantically encoding labels at image-level into various image patches in each scenic picture;
- To our best knowledge, semantically/visually important regions within each scenery are practically described by different low-level descriptors, each captures scenic regions in a single channel. To fuse these low-level features complementarily, it is necessary to intelligently predict the weight of each feature channel. However, mathematically deriving a solvable weighting scenario is uneasy. Some practical difficulties are: i) how to incorporate the local feature of spatially neighboring regions inside each scenery, ii) how to preserve the global compositional feature among the multiple internal scenic regions, and iii) how to adaptively adjust the channel weights with respect to different scenic image set.

To tackle the aforementioned issues, a new scenery categorization pipeline is proposed by deeply and actively modeling human gaze behavior, wherein each scenic image patch is represented by optimally fusing a variety of low-level features. A description of the our work is shown in Fig. 1. Concretely, supposing we have a rich set of scenic images wherein the labels are potentially contaminated, we first deploy the well-known binarized norm gradients (BING) [50] for obtaining many object-aware patches inside all the scenic images. To represent each scenic patch, we formulate a low-level feature fusion algorithm that simultaneously encodes the local and global sample geometry structure. Subsequently, to stimulate human gaze allocation during scene image perception, we propose a novel robust deep active learning (RDAL) framework to jointly calculate human gaze shift path (GSP) and learn the deep GSP representation. Herein, one key advantage of the RDAL is that the contaminated and redundant image labels can be intelligently handled. Besides, RDAL is trained in a semi-supervised mode, *i.e.*, only a part of semantic labels are required. Finally, using the deep GSP features learned by RDAL, the training scenic pictures are combined into a image kernel machine, which is leveraged to train a multi-label SVM for scenery categorization. Empirical evaluations on both the six public scenery data set and our compiled sport educational image set showed our method's advantage.

Totally speaking, the novelties of this work are three-fold. First, we formulate the RDAL framework that can actively learn human gaze behavior and deeply calculate gaze-guided visual feature simultaneously. Second, we deploy a advantageous feature integration technique to dynamically calculate the importance of different feature channel for each scenic patch. Third, we drive an iterative algorithm to solve the RDAL that is non-convex to its variables.

II. RELATED WORK

In computer vision community, lots of deep scene categorization models have been released. The hierarchical CNNs coupled with carefully-designed deep structures can effectively conduct scene recognition toward Internet-scale scene images like the well-known ImageNet [33]. In [10], researchers proposed to learn a massive-scale deep neural network by leveraging part of the ImageNet [33] image set. And they have received overwhelming categorization accuracy. In practice, we notice that, despite the fact that ImageNet-CNN are designed for generic visual modeling, the generated deep representations can enhance many computer vision tasks, such as video parsing and abnormal event detection. In the past decade, standard ImageNet-based CNN were updated from two aspects. On one hand, researchers proposed to effectively obtain a large collection of samples to improve the training of the multi-layer architecture. For example, the pervasively-used selective search [34] incorporates the attributes of enumerative search and semantics-level annotation into a unified framework. A succinct set of category-independent patch samples can be generated for deep learning. On the other hand, the authors [35]proposed the so-called region-level CNN (RCNN), aiming at effectively sampling a set of high quality patch samples. The authors [9] upgraded the CNN-guided scenery categorization by producing highly descriptive training data. That is, an Internet-scale scenery-related image set have been collected. Moreover, it is usually ineffective to train a deep visual architecture using the entire scenic picture or random scenic patches. In this way, the authors [37] deployed a pre-learned hierarchical CNN for producing local and representative scenic patches to optimize the deep scene categorization learning. Moreover, in [4], a multi-task and multi-resolution scenery categorization algorithm is proposed by maintaining the intrinsic feature distribution using a manifold-based regularizer. In [5], the author proposed a scenery semantic annotation framework, wherein a low-rank deep features is calculated to capture the category-based posterior probability. Meanwhile, a Markov probabilistic model is utilized to learn the contextual feature for each scenery. The authors [6] formulated a deep model by discovering the relationship between different deep layers. Afterward, an unlabeled learning model progressively learns the deep feature by leveraging the scenery geometrical feature. Further in [39], the authors seamlessly combine discriminative feature learning and weak label learning into a unified scene analytical model. A so-called stack discriminative sparsity autoencoder is designed for calculating the high-level visual representations.

Plenty of computational visual models were proposed for analyzing aerial photos. In [40], the authors provided a multi-modal learning algorithm to simultaneously annotate the HR aerial imagery. The authors [17] provided a novel multi-attention-based algorithm to calculate aerial photos' representation's weights. In conclusion, the above image-level visual models are practically utilized for classifying multi-resolution aerial images. They cannot optimally handle LR aerial image modeling because of the unavoidable blurred tiny but discriminative objects. To precisely capture discriminative objects with multiple scales, we require an effective region-level modeling technique. In this way, we can precisely localize those tiny/small objects inside each LR aerial photograph. In [55], the authors designed a so-called group sparsity regularizer for enhancing robustly recognize human faces. They proposed an upper-bounded function to upgrade the l_1-norm to seek sparsity. This can optimally tackle the negative influences of bias and outliers. Further in [36], the authors formulated the incomplete multi-view clustering into a incomplete similarity graphs upgradation and complete tensor representation learning task. To characterize an aerial image regionally, researchers [15] designed a multi-layer deep learner for detecting multi-scale important ground objects. In [52], researchers formulated a focal-loss-based deep model to accurately localize various cars within each LR&HR aerial photographs. In [54], the authors designed a geographic object detection model to handle HR aerial images by intelligently extracting intersections as well as roads. In [53], the authors proposed to combine feature engineering and soft-labels calculation to form an effective visual detector for modeling aerial images.

III. OUR SCENERY UNDERSTANDING PIPELINE

A. EXTRACTING OBJECT-AWARE PATCHES

In the literature, many surveys in visual cognition and psychology [48], [49] have revealed the fact that, when humans perceiving different sceneries, their gazes will first fixed onto those semantically/visually important regions. That is to say, only a few discriminative scenic regions will be selected for visual cognition. In practice, we believe that it is necessary to incorporate such human gaze allocation into scenery categorization. Here, an efficient object-aware patch detection as well as a robust deep active learning (RDAL) technique is designed to localize those semantically/visually important scenic patches for mimicking human visual perception.

In practice, human vision system tends to fix onto those semantically/visually important objects or their parts, such as vehicles and tall buildings. It is observable that, such objects couple with the spatial distributions significantly influence how human beings perceiving different scenic pictures. In order to identify objects or their parts that potentially attract human visual attention, we deploy the well-known BING [50] objectness measure for extracting a set of high quality object-aware patches within different sceneries. Herein, we claim three key competitiveness of BING. First, it achieves a super high object patch detection effectiveness with very low computation. Second, the GSP extraction can be substantially enhanced by producing s high quality set of object-level patch. Third, BING exhibits an optimal generalization capacity to unseen object categories. This makes the trained scenery categorization model highly adaptable to multiple data sets.

B. OPTIMALLY FUSING PATCH FEATURES

By enumeratively extracting the BING [50] object patches inside a scenic picture, we practically obtain a set of low-level features to characterize a scenic patch. Subsequently, we develop a multi-channel feature fusion algorithm that optimally combines these low-level visual features. In our work, a local-global feature fusion scheme is employed, which has three impressive attributes: 1) the patch local distribution in the low-level feature space should be maximally kept since each patch is usually visually similar to its neighbor, 2) the patch global distribution in the low-level feature space should also be well maintained since it determines the global scenery composition, and

3) the feature weights are dynamically tuned toward each scenic image set.

1) PATCH LOCAL DISTRIBUTION

Herein, we represent x_j^i as the visual feature from j-th scenic patch in the i-th feature channel. And we denote x_j^i as well as the L spatially adjacent ones as $\mathbf{X}_j^i = [x_j^i, x_{j1}^i, \cdots, x_{jL}^i]$. Meanwhile, $\mathbf{Y}_j^i = [y_j^i, y_{j1}^i, \cdots, y_{jL}^i]$ is used to represent the feature fusing result for $\mathbf{X}_j^i$. On this basis, we can formulate the task of maintaining the local distribution of L spatially adjacent scenic patches as follows:

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

Herein, r_j^i denotes a M-dimensional vector describing the correlation of scenic patch x_j^i and its spatially adjacent scenic patch, *i.e.*, $(r_j^i)_l = \exp\left(-\frac{||x_j^i - x_{jl}^i||^2}{t^2}\right)$, t represents the variance of a Gaussian distribution.

Based on the derivations by us, we can reorganize the aforementioned objective function into the matrix form:

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

Herein, matrix $\mathbf{B} = [-\mathbf{e}_M^T, \mathbf{I}_M]^T \mathrm{diag}(r_j^i)[-\mathbf{e}_M^T, \mathbf{I}_M] \in \mathbb{R}^{(M+1)\times(M+1)}$. Noticeably, $\mathbf{e}_M = [1, \cdots, 1]$ denotes a M-dimensional vector, matrix $\mathbf{I}$ represents an $M \times M$ identity one, and matrix $\mathrm{diag}(r_j^i)$ denotes an $M \times M$ diagonal one and the jj-th entity is r_j^i.

In a mathematical view, locally optimizing the H features is represented as:

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

Herein, κ_i measures the importance of each channel of feature.

2) PATCH GLOBAL DISTRIBUTION

As we introduced above, incorporating the global geometry of object-aware patches within a scenic picture is an important task. Herein, we hypothesize that $\mathbf{Y}_j^i = \mathbf{Y}\mathbf{A}_j^i$, and matrix $\mathbf{A}_j^i = \mathbb{R}^{N\times(M+1)}$ represents the selecting one. It reflects that scenic patches are distributed locally to the whole scenic patches inside an image. In this way, we can upgrade objective function (3) as follows:

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

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:

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

Herein, $\mathbf{C}^i$ represents a diagonal matrix. Each entity is calculated as: $\mathbf{C}_{jj}^i = \sum_l [\mathbf{S}^i]_{jl}$, $\mathbf{S}$ denotes an $N \times N$ matrix, that is, $[\mathbf{S}^i]_{uv} = \exp(-\frac{||x_u - x_v||^2}{t^2})$, and N represents how many scenic patches inside a scenic picture. Notably, $\mathbf{C}^i$ denotes the unnormalized Laplacian matrix [13]. To accelerate the calculation, we introduce a normalization step to $\mathbf{D}^i$, that is,

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

Herein, $\mathbf{D}_n^i$ represents a normalized $\mathbf{D}^i$.

In total, our local to global feature fusion scenario can be formulated into the below objective function:

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

We observe that, the minimization of (7) makes $\kappa_i = 1$. Here, we simply select multiple highly informative features. In practice, using a hard constraint is a sub-optimal choice. This is because we need multiple features be simultaneously utilized for scenery categorization. Aiming at this objective, we apply the trick in [16]. More specifically, we have the following setup $\kappa_i \leftarrow \kappa_i^o$ and $o > 1$. Herein, the ideal κ_i toward multi-channel features have to be dynamically adjusted. In theory, each channel contributes uniquely toward the resulting fused feature for optimally describing each scenic patch.

C. ROBUST DEEP ACTIVE LEARNING (RDAL)

There exists an abundance of object patches ($10^2 \sim 10^4$) extracted by leveraging BING [50]. However, in practical scenarios, human attention is generally directed to a few objects within each scene. In order to reflect such actively perceiving each scenery, a novel robust deep active learning (RDAL) approach has been designed. This approach aims to jointly identify L scenic patches for GSP construction, based on which we compute the deep GSP representation. RDAL cooperatively fuzes the following aspects: 1) scenery spatial composition, 2) semantic descriptiveness of object patches, and 3) those potentially contaminated semantic labels.

1) SPATIAL COMPOSITION OF DIFFERENT SCENERIES

It is widely recognized that an effective scenery categorization algorithm is expected to describe scenic spatial compositions, specifically the relative positioning of foreground and background regions. For quantifying this character, it is reasonable to assume that a scenic patch is represented by its spatially adjacent ones. In the representing process, we can weight the importance of an object patch by optimizing the below formulation:

$$\begin{aligned} &\arg\min_{\mathbf{E}} \sum\nolimits_{i=1}^{N} ||z_i - \sum\nolimits_{j=1}^{N} \mathbf{F}_{ij} z_j|| \\ &s.t. \sum\nolimits_{j=1}^{N} \mathbf{F}_{ij} = 1, \mathbf{F}_{ij} = 0 \ \ if \ \ z_i \notin \mathcal{N}(z_j), \end{aligned} \tag{8}$$

Here, $\{z_1, \cdots, z_N\} \in \mathbb{R}^{N\times A}$ denotes deeply-learned features calculated from the N scenic patches obtained through BING [50] in each scenery. Here, A represents the dimension of each scenic patch's deep representation, and the matrix $\mathbf{F}_{ij}$ indicates the importance of the i-th scenic patch to recover the j-th scenic patch. Additionally, $\mathcal{N}(z_i)$ encompasses the spatial adjacent ones with respect to scenic patch z_i.

2) SEMANTIC REPRESENTATIVENESS OF SCENIC PATCHES

In addition to spatial encoding different sceneries, our selected scenic patches' semantic representativeness in building GSPs is also significant. Utilizing the reconstruction error defined in (8), the reconstructed scenic patches can be represented as $g_1, \cdots, g_N$. Thereafter, the L selected scenic patches are identified by minimizing the subsequent equation:

$$\eta(g_1, \cdots, g_N) = \sum\nolimits_{i=1}^{L} ||g_{q_i} - g_{q_i}||^2 + \tau \sum\nolimits_{i=1}^{N} ||g_i - \sum\nolimits_{j=1}^{N} \mathbf{F}_{ij} g_j||^2, \tag{9}$$

Here, τ weights the regularizer, and $g_{q_1}, \cdots, g_{q_K}$ represents the set of L scenic patches selected by our RDAL. The first term particularly minimizes the cost to fix the coordinates of the selected samples. Simultaneously, the last term ensures that the semantically reconstructed scenic patches highly similar to the input. Overall, minimizing (16) yields a collection of scenic patches that precisely reflect human visually/semantically perceiving diverse scenes.

We define matrices $\mathbf{A} = [z_1, \cdots, z_N]$ and $\mathbf{H} = [g_1, \cdots, g_N]$, and denote matrix Δ as an $N \times N$ diagonal one encoding the selected scenic patches. In this context, $\Delta_{ii} = 1$ if $i \in \{q_1, \cdots, q_L\}$ and 0 otherwise. This allows upgrading the objective function (16) in the following:

$$\eta(\mathbf{Q}) = \mathrm{tr}((\mathbf{H} - \mathbf{A})^T \Delta (\mathbf{H} - \mathbf{A})) + \tau \mathrm{tr}(\mathbf{H}^T \mathbf{L} \mathbf{H}), \tag{10}$$

Herein, we have $\mathbf{L} = (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F})$. For optimizing (17), we set $\eta(\mathbf{H})$'s gradient to zero, and thereby we have:

$$\Delta(\mathbf{H} - \mathbf{A}) + \tau \mathbf{L}\mathbf{H} = 0. \tag{11}$$

In this context, we can compute the rebuilt scenic patches as follows:

$$\mathbf{H} = (\tau \mathbf{L} + \Delta)^{-1} \Delta \mathbf{A}. \tag{12}$$

By leveraging the reconstructed scenic patches, we can upgrade the reconstruction error as:

$$\eta(z_{q_1}, \cdots, z_{q_K}) = ||\mathbf{Z} - \mathbf{G}||_F^2 = ||\mathbf{Z} - (\tau \mathbf{K} + \Delta)^{-1} \Delta \mathbf{Z}||_F^2 = ||(\tau \mathbf{K} + \Delta)^{-1} \tau \mathbf{K} \mathbf{Z}||_F^2, \tag{13}$$

Herein, $|| \cdot ||_F^2$ represents the Frobenius norm for a matrix.

3) OUR RDAL FRAMEWORK

To semantically learn the visual descriptors within each scenery, we hierarchically compute the hidden scenery-related feature using a deep model. As illustrated in Fig. 2, for an R-layer deep architecture, the proposed RDAL decomposes the matrix of semantic labels $\mathbf{G}$ into $R+1$ factor matrices: $\mathbf{V}$, $\mathbf{U}_R, \cdots, \mathbf{U}_1$. For convenient derivation of deep features of each scenery and the representation of new scenic images, the first deep layer calculates features based on the following equation: $\mathbf{U}_1 = \mathbf{W}_1 \mathbf{X}$. Importantly, in the RDAL, we propose the fundamental idea rather than presenting a sophisticated formulation. That is, we deploy multiple linear combinations to deeply engineer the latent scenery features. Theoretically, the hierarchial multi-layer architecture can be expressed as:

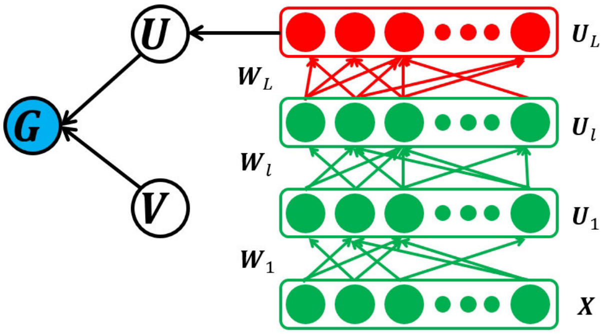

FIGURE 2. Structure of the designed deeply and semantically GSP encoding.

$$\begin{aligned} &\mathbf{G} \leftarrow \mathbf{P}\mathbf{Q}_R, \\ &\mathbf{Q}_R = \mathbf{U}_R \mathbf{P}_{R-1}, \\ &\quad \cdots \\ &\mathbf{Q}_1 = \mathbf{U}_1 \mathbf{Y}, \end{aligned} \tag{14}$$

Here, $\mathbf{U}_i$ represents the i-th layer's transformation matrix, $\mathbf{P}$ denotes the matrix containing the unobservable semantic labels, and $\mathbf{Q}_i$ is the calculated scene representation matrix from the i-th deep layer. $\mathbf{Y}$ is a matrix comprises y_i, which is the B-dimensional fused feature from the i-th scenic patch. For our proposed RDAL, the deep representation corresponding to the top layer $\mathbf{Q}$ can be characterized by $\mathbf{Q} = \mathbf{Q}_L$. Following (14), for the training of our deep model, we focus on learning a factor $\mathbf{P}$ as well as R transformation matrices $\mathbf{U}_R, \cdots, \mathbf{U}_1$.

In summary, the entire deep-model-guided active learning can be mathematically represented as:

$$\min_{\mathbf{P}, \Delta \mathbf{U}_1, \cdots, \mathbf{U}_R} \frac{1}{2} ||\mathbf{F} - \mathbf{P}\mathbf{Q}||_F^2 + \frac{\alpha}{2} ||\mathbf{P}||_F^2 + \frac{\alpha}{2} \sum\nolimits_{i=1}^{R} ||\mathbf{U}_i||_F^2 + \frac{\beta}{2} ||\mathbf{U}||_{2,1}, \tag{15}$$

Herein, matrix $\mathbf{F} \in \mathbb{R}^{R \times N}$ contains the semantic labels, with $\mathbf{F}_{ij} = 1$ indicating that the i-th scenic image is labeled by j, and $\mathbf{F}ij = 0$ otherwise. Meanwhile, R counts the unique semantic labels, α acts as the regularizer weight to prevent overfitting, and β maintains the column-wise sparsity in Ui. Recognizing that visual features may be correlated, redundant, or even contaminated, it is necessary to incorporate a sparse model with the $l21$-norm. This effectively avoids the low quality noisy features. We present the solution of (15) in the supplementary document.

It is essential to highlight that, unlike the first two visual feature (scenery spatial composition and patch-level semantic description), our proposed RDAL framework is executed in a semi-supervised mode. That is, model training requires only a few semantic labels, as indicated in (15). It is advantageous for modeling lots of images where many semantic labels may be missing since manual labeling is intractable.

By learning the deep GSP representation for each scenic image, following [24], a multi-label SVM is learned for scenery categorization.

4) KERNEL-INDUCED SVM

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:

$$\alpha_i \propto \exp\left(-d(y(\mathcal{P}_j^*), y(\mathcal{P}_j^i))\right), \tag{16}$$

In this formulation, $d(\cdot,\cdot)$ is used to represent the Euclidean distance between pairs of vectors, where y representations the deep visual feature extracted from each GSP. The parameter N counts the training scenic images.

Utilizing the feature vector derived as mentioned, we proceed to train a multi-class SVM [27] 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.

$$\max_{\beta\in\mathbb{R}^{N_{pq}}} \ \omega(\beta) = \sum\nolimits_{i=1}^{N_{pq}} \beta_i - \frac{1}{2}\sum\nolimits_{i=1}^{N_{pq}} \gamma_i\gamma_j l_i l_j k(\alpha_i, \alpha_j)$$
$$s.t. \quad 0 \le \gamma_i \le C, \sum\nolimits_{i=1}^{N_{pq}} \gamma_i l_i = 0, \tag{17}$$

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. EMPIRICAL ANALYSIS

Herein, we evaluate the efficacy of our scene classification model by RDAL through four experimental evaluations. We begin by presenting the experimental configurations and introducing six benchmark scene datasets. Subsequently, we conduct a comparative analysis with various shallow and multi-layer recognizers. Next, we examine the impact of key variables in our approach. Lastly, we leverage the deep GSP feature learned by our model to improve education-related sport scenery categorization.

In this work, our categorization model and all the compared baseline models are implemented using Python 3.10. The computational platform includes four Nvidia A100 GPUs, an Intel Xeon w9-3495X, and 256GB main memory.

A. DATA SETS AND SETTING

To extensively evaluate our categorization model, we conduct experiments on six diverse scenic image sets, including two standard as well as multiple recent ones. Example images from the experimental scenery sets are illustrated in Fig.3. The two standard data sets are Scene-15 [11] and 67 [12].

- Scene-15: This data set encompasses 15 categories, with 13 released by Feifei [14]. Each scenery category consists of 200 $\sim$ 400 scenic pictures, with an average resolution of 320 × 250. The images are mostly sourced from COREL, individual pictures, and Google.
- Scene-67: This data set includes a rich set of indoor scenic pictures, compiled based on three sources: 1) Picasa and Altavsta, 2) photograph sharing website, and 3) the LabelMe images.

Besides, we four four more recent scenic picture sets, namely ZJU aerial imagery [3], ILSVRC-2010 [33], SUN [7], and Places [9]. Moreover, we also introduce an non-public scenery set compiled by ourselves. This data set contains massive-scale sport educational images (called MSEI), that is, the sport sceneries are leveraged for educational purpose. In detail, this dataset comprises 92,0000 images gathered from nice sports types, namely basketball, football, volleyball, outdoor golf, athletics, table tennis, rowing, baseball and equestrian. The snapshot of our collected dataset is presented in Fig. 4. Specific statistics about our dataset can be found in Table 1.

TABLE 1. Details of our sport educational image set.

Sport name	Training image #	Validation 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

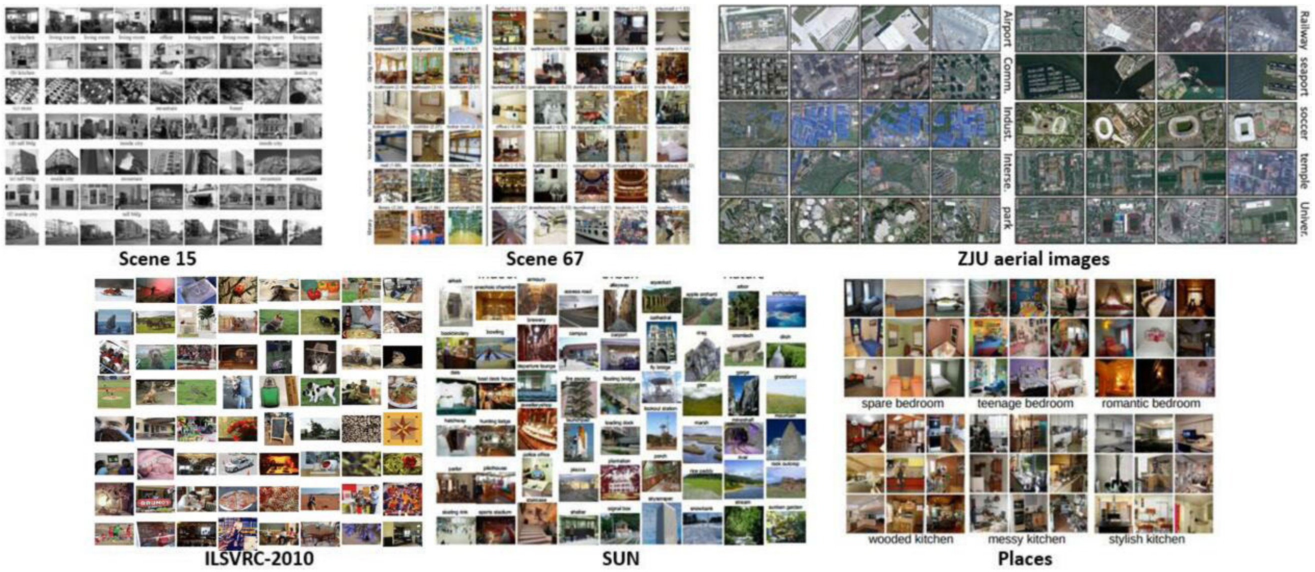

FIGURE 3. Some example pictures from the above six scene data sets.

FIGURE 4. Some example pictures from our sport educational image set.

For the Scene-15, we follow the default settings [11], wherein 100 scenic images from each category are for training and the rest are for testing. For the Scene-67, following [12], 80 scenic pictures from each category are employed for model training, whereas the rest are for model evaluation. For the ZJU aerial imagery [3], we randomly select half of the aerial photos within each category for training and the rest are for model testing. For the ILSVRC-2010 [33], there exists approximately 1.2 million training scenic images, 50,000 validation scenic images, and 150,000 testing scenic images. For the SUN [7], we randomly select 50 samples for training and another 50 samples for testing, following the experimental setups in the publication. For the Places [9], it contains 1,803,460 training images in total. For each category, the training image number varies from 3,068 to 5,000. Meanwhile, the evaluation set contains 900 images per class.

Before carefully testing the baseline algorithms, we provide an overview of our approach's empirical setups: 1) object patches: The BING [50] scenic patches are consistently set at 1000 toward the six scenic image sets. This ensures effectively localizing all the potential objects. 2) spatial neighbors: spatial neighbor number (L) is fixed at five. 3) low-level features: we employ three low-level features for representing each object patch: a 16-D color moment [56], a 64-D HOG [57], a 160-D edge and color histogram [8]. 4) GSP's internal regions: the number of GPS's internal regions, denoted as K, is fixed at five. This setup aligns with the observation that humans practically attend to at most five salient regions in a scene. 5) patch-level deep feature: the dimension of our deep patch-wise feature is fixed to 212.

B. COMPARISON WITH OTHER RECOGNITION MODELS

1) SCENERY CATEGORIZATION TASK

To begin with, our perception-guided scenery categorization model undergoes empirical comparison with four commonly utilized shallow classification models. These models are: 1) fixed-length walk kernel (FWK) and its tree kernel version (FTK) [18]. 2) multi-resolution histogram (MRH) [25]. 3) kernel machine learning by SP, along with three variants: LLC-SP [19], SC-SP [20], and OB-SP [21]. 4) image representation by super vector (SV) [22] and supervised image coding(SSC) [23]. In the comparative analysis, the settings for each algorithm are standardized as follows: The lengths of FWK and FTK are tuned within the range of two to ten. For MRH, the scene images are pre-processed by RBF-based smoothing, which is calculated using 12 gray scales. For SPM and the upgraded versions, all the training scenic pictures are decomposed into one SIFT descriptors extracted using 16 × 16 grids. Subsequently, a 400-sized codebook is trained by leveraging kmeans clustering.

In light of the remarkable performance achieved by multi-layer recognizers recently, we carried out a comparative

TABLE 2. Averaged categorization accuracies on the compared models on the aforementioned data sets.

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(MKL)	Ours(Softmax)			
Scene-15	82.8%	86.6%	90.2%	91.2%	**93.4%**	92.1%			
Scene-67	71.3%	72.1%	71.5%	74.7%	**76.7%**	72.9%			
ZJU Aerial	81.1%	80.1%	78.6	81.2%	**84.3%**	82.6%			
ILSVRC-2010	40.5%	41.1%	40.8%	41.1%	**44.2%**	42.7%			
SUN397	50.4%	51.2%	52.2%	52.0%	**56.3%**	53.2%			
Places205	49.3%	48.2%	48.3%	49.3%	**52.1%**	50.1%			
MSEI	59.7%	61.5%	62.5%	64.7%	**72.4%**	71.6%			

TABLE 3. Derivations on the compared models on the aforementioned data sets.

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 (MKL)	Ours (Softmax)			
Scene-15	0.012	0.014	0.013	0.014	0.009	0.011			
Scene-67	0.017	0.012	0.013	0.013	0.007	0.009			
ZJU Aerial	0.014	0.013	0.014	0.012	0.008	0.007			
ILSVRC-2010	0.013	0.013	0.014	0.011	0.009	0.007			
SUN397	0.012	0.013	0.014	0.013	0.009	0.008			
Places205	0.013	0.012	0.014	0.012	0.008	0.006			
MSEI	0.014	0.011	0.015	0.014	0.006	0.009			

study with a collection of deeply-learned scene recognition models. In detail, we assess the below deep models: ImageNet CNN (IN-CNN) [10], R-CNN [35], meta object CNN (M-CNN) [37], deep mining CNN (DM-CNN) [26], and spatial pyramid pooling CNN (SPP-CNN) [28]. With the sole exception of [37], the deep recognition models' source codes are public, facilitating a direct assessment without any modifications to the parameters. For [37], we started by by selecting 192 to 384 region proposals from each of the six image sets, generated by MCG [29]. We fix the the regional visual representation at a dimension of 4096 from the FC7 layer from the combined CNN [9]. Next, we produce 400 superpixels in each scenery by leveraging the well-known SLIC [2]. The superpixels are optimized through either pre-specified linear LDA (SP-LDA) or the selection of the 120 visually attractive patches calculated by GBV [1] (SP-GBV). For our method, we combining multiple low-level features, then the RDAL selects semantically/visually salient superpixels (referred to as GSPs) for building Graph-based Superpixels (GSPs). They are integrated for calculating the kernel machine to classify sceneries. The performance of our BING-based rectangular patches and superpixels is detailed in Tables 2 and 3. Remarkably, the BING-guided rectangular patches outperform superpixels, indicating the higher descriptiveness. Last but not least, a comparative study is conducted with multiple recent scenery categorization models by Mesnil et al. [30], Xiao et al. [31], and Cong et al. [32].

TABLE 4. Averaged precisions on the compared models on the aforementioned data sets.

Data set	FWK	FTK	MRH	PM	LLC-SP	SC-SP	OB-SP	SV	SSC
Scene-15	64.2%	66.4%	58.1%	62.3%	72.5%	74.4%	68.3%	74.1%	78.6%
Scene-67	35.2%	24.3%	30.1%	38.5%	42.1%	39.2%	39.7%	41.2%	44.2%
ZJU Aerial	59.4%	60.2%	53.3%	63.8%	70.4%	69.6%	69.2%	68.5%	73.3%
ILSVRC-2010	28.5%	26.3%	25.4%	27.4%	32.0%	31.3%	32.6%	31.4%	33.5%
SUN397	14.3%	13.1%	12.6%	16.3%	33.1%	33.1%	35.4%	32.2%	37.1%
Places205	19.1%	18.9%	17.2%	21.7%	27.5%	24.1%	26.4%	24.4%	26.6%
MSEI	43.2%	42.5%	43.0%	42.6%	44.8%	49.5%	42.2%	46.4%	47.3%
Data set	IN-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	SP-S	SP-GBV	SP-LDA	Mesnil
Scene-15	64.3%	79.3%	82.1%	80.5%	83.4%	83.6%	82.1%	81.7%	81.6%
Scene-67	51.3%	57.2%	61.6%	60.7%	56.4%	68.4%	61.2%	60.8%	60.3%
ZJU Aerial	66.3%	72.4%	71.6%	72.6%	71.5%	72.8%	71.5%	72.5%	69.6%
ILSVRC-2010	31.5%	33.4%	35.2%	34.4%	36.1%	36.3%	35.8%	36.2%	35.9%
SUN397	43.4%	41.6%	44.8%	42.4%	46.4%	45.8%	43.3%	45.4%	44.7%
Places205	35.1%	37.5%	39.3%	40.3%	42.1%	43.6%	43.2%	41.6%	42.8%
MSEI	46.3%	44.7%	43.5%	46.7%	48.3%	46.3%	53.7%	57.5%	58.4%
Data set	Xiao	Cong	Fast R-CNN	Faster R-CNN	Ours(MKL)	Ours(Softmax)			
Scene-15	73.1%	78.2%	81.3%	83.6%	**88.1%**	85.7%			
Scene-67	63.6%	64.5%	66.3%	66.2%	**71.1%**	67.4%			
ZJU Aerial	73.3%	72.6%	71.6	74.5%	**78.8%**	74.6%			
ILSVRC-2010	35.1%	37.3%	34.9%	36.4%	**40.1%**	37.4%			
SUN397	45.4%	46.1%	46.6%	46.4%	**51.5%**	47.6%			
Places205	42.4%	43.5%	41.9%	42.0%	**47.5%**	41.4%			
MSEI	53.5%	52.7%	55.4%	58.3%	**66.4%**	62.7%			

TABLE 5. Averaged recalls on the compared models on the aforementioned data sets.

Data set	FWK	FTK	MRH	PM	LLC-SP	SC-SP	OB-SP	SV	SSC
Scene-15	43.6%	45.6%	66.3%	48.6%	30.9%	32.7%	35.5%	38.0%	31.3%
Scene-67	45.4%	48.4%	57.4%	58.6%	50.7%	50.4%	28.4%	34.4%	39.4%
ZJU Aerial	44.7%	50.4%	60.6%	57.6%	41.2%	42.3%	40.2%	42.6%	38.8%
ILSVRC-2010	40.4%	41.3%	48.3%	49.5%	48.4%	53.2%	54.9%	58.5%	55.5%
SUN397	55.4%	50.4%	57.5%	53.0%	54.6%	58.3%	52.7%	59.4%	57.2%
Places205	30.3%	32.5%	34.6%	30.3%	35.7%	37.9%	33.6%	40.6%	40.4%
MSEI	65.3%	65.8%	67.8%	63.4%	65.8%	62.4%	67.5%	65.3%	67.3%
Data set	IN-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	SP-S	SP-GBV	SP-LDA	Mesnil
Scene-15	28.7%	20.5%	27.6%	20.5%	27.4%	18.4%	26.7%	28.7%	29.3%
Scene-67	59.6%	58.6%	48.6%	49.5%	54.3%	43.4%	46.7%	50.2%	48.5%
ZJU Aerial	47.5%	38.5%	39.1%	38.4%	39.4%	35.6%	38.2%	37.5%	30.5%
ILSVRC-2010	68.4%	62.4%	64.6%	68.5%	63.1%	68.6%	66.8%	62.7%	67.1%
SUN397	60.4%	67.4%	68.8%	67.3%	65.1%	62.4%	61.5%	64.8%	63.7%
Places205	64.5%	65.3%	62.3%	70.2%	65.4%	63.6%	67.4%	68.2%	63.8%
MSEI	64.3%	66.8%	68.4%	64.6%	64.3%	65.3%	64.6%	58.4%	53.2%
Data set	Xiao	Cong	Fast R-CNN	Faster R-CNN	Ours(MKL)	Ours(Softmax)			
Scene-15	36.5%	33,1%	40.6%	39.3%	**48.4%**	46.4%			
Scene-67	50.6%	46.5%	47.5%	43.2%	**63.2%**	60.3%			
ZJU Aerial	67.4%	67.1%	58.5	65.4%	**71.2%**	66.9%			
ILSVRC-2010	74.3%	72.4%	75.0%	74.1%	**83.4%**	78.3%			
SUN397	64.4%	63.2.1%	67.0%	60.3%	**77.4%**	71.7%			
Places205	65.7%	68.5%	66.5%	68.3%	**83.2%**	77.4%			
MSEI	64.3%	68.1%	63.4%	61.5%	**73.3%**	69.5%			

Examining Tables 2 and 3, we perform a quantitative comparison across the aforementioned deep or flat visual recognizers. Each experiment undergoes 20 repetitions, wherein the corresponding standard deviations are presented. As shown, our method performs the best in both the classification accuracy and stability. More importantly, on our compiled MSEI image set, the RDAL performs overwhelmingly, *i.e.*, the categorization precision is over 8% higher than the second best performer. Besides, more details results are shown in Tables 4, 5, 7, and 6.

C. PERFORMANCE BY ADJUSTING PARAMETERS

Totally, our designed perception-guided deep recognizer incorporates several important parameters impacting the effectiveness of scenery categorization. Herein, we assess our method's performance by tuning these parameters, and suggest optimal setups based on the outcomes. Herein, we test three parameters: i) L, counting the neighbors for rebuilding an object patch, ii) K, counting the selected object patches in a GSP, and iii) weights for regularization terms α, β, γ. Our experimentation is conducted on the Scene-15 [11] because of the impractical time consumption of experiments on larger data sets. Next, L counts the neighbors utilized for rebuilding a scenic patch. Maintaining the locality of object patches during our designed feature fusion is crucial. Herein, we vary L from one to 15 progressively. Correspondingly, the average recognition precisions across the 15 scenery categories are reported. As illustrated in Fig. 6, the precision

TABLE 6. Averaged F_1 scores on the compared models on the aforementioned data sets.

Data set	FWK	FTK	MRH	PM	LLC-SP	SC-SP	OB-SP	SV	SSC
Scene-15	52.7%	55.3%	61.7%	54.9%	40.6%	42.3%	51.4%	53.3%	44.2%
Scene-67	39.8%	36.0%	43.4%	48.3%	46.2%	44.6%	32.7%	37.6%	41.5%
ZJU Aerial	37.3%	55.0%	56.5%	60.5%	34.7%	47.5%	43.8%	53.7%	47.1%
ILSVRC-2010	37.1%	37.2%	38.4%	35.0%	40.1%	41.6%	39.8%	39.2%	37.3%
SUN397	20.4%	18.9%	18.9%	20.9%	40.4%	40.7%	42.1%	42.9%	45.6%
Places205	23.8%	24.0%	24.6%	25.9%	31.1%	30.7%	29.9%	31.9%	31.6%
MSEI	53.1%	53.3%	54.0%	52.8%	54.5%	54.6%	53.8%	55.1%	55.6%
Data set	IN-CNN	R-CNN	M-CNN	DM-CNN	SPP-CNN	SP-S	SP-GBV	SP-LDA	Mesnil
Scene-15	40.6%	39.3%	54.1%	39.4%	54.7%	51.2%	50.3%	54.3%	55.8%
Scene-67	55.2%	57.9%	54.6%	55.0%	55.3%	54.0%	53.6%	55.2%	54.4%
ZJU Aerial	56.1%	51.4%	53.9%	54.3%	54.3%	52.8%	53.8%	52.9%	46.6%
ILSVRC-2010	49.8%	46.5%	47.9%	49.2%	49.3%	49.1%	47.9%	49.2%	48.9%
SUN397	50.5%	53.1%	55.6%	54.0%	55.2%	53.7%	52.7%	54.5%	54.2%
Places205	45.2%	49.0%	49.0%	53.1%	52.2%	52.6%	54.4%	53.6%	52.5%
MSEI	54.6%	55.2%	55.9%	55.1%	55.9%	56.2%	58.6%	57.9%	55.8%
Data set	Xiao	Cong	Fast R-CNN	Faster R-CNN	Ours(MKL)	Ours(Softmax)			
Scene-15	77.7%	82.1%	85.4%	87.2%	**90.6%**	88.9%			
Scene-67	57.0%	54.8%	56.4%	53.7%	**67.0%**	63.5%			
ZJU Aerial	70.3%	69.8%	64.9%	69.6%	**74.8%**	70.5%			
ILSVRC-2010	51.0%	52.9%	53.3%	53.1%	**58.1%**	55.0%			
SUN397	53.2%	54.1%	56.1%	53.4%	**62.4%**	58.8%			
Places205	51.6%	55.1%	53.5%	54.7%	**63.1%**	56.7%			
MSEI	57.6%	59.4%	59.1%	59.8%	**69.5%**	65.9%			

TABLE 7. Confusion matrix of our method on the MSEI.

	Basketball	Football	Volleyball	Golf	Athletics	Hockey	Tennis	Rowing	Equestrian
Basketball	65.2%	4.3%	5.3%	3.5%	4.0%	3.1%	2.3%	3.2%	9.1%
Football	3.1%	80.4%	2.4$	1.9%	3.3%	1.4%	2.2%	3.1%	2.2%
Volleyball	2.6%	4.2%	70.3%	4.1%	1.8%	2.5%	5.3%	5.1%	4.1%
Golf 3.4%	3.8%	4.2%	1.5	74.6%	2.8%	2.6%	2.8%	1.4%	2.9%
Athletics	4.4%	3.7%	4.6%	3.1%	68.5%	3.1%	2.5%	6.9	3.2%
Hockey	3.4%	2.7%	4.1%	1.4%	2.2%	76.9%	2.9%	2.1%	4.3%
Tennis	2.4%	1.8%	3.2%	2.3%	1.8%	2.4%	80.4%	4.3%	1.4%
Rowing	4.6%	5.3%	3.4%	2.7%	3.5%	2.9%	3.6%	69.6%	4.4%
Equestrian	3.5%	5.6%	1.9%	2.8%	3.7%	4.5%%	3.2%	4.0%	70.8%

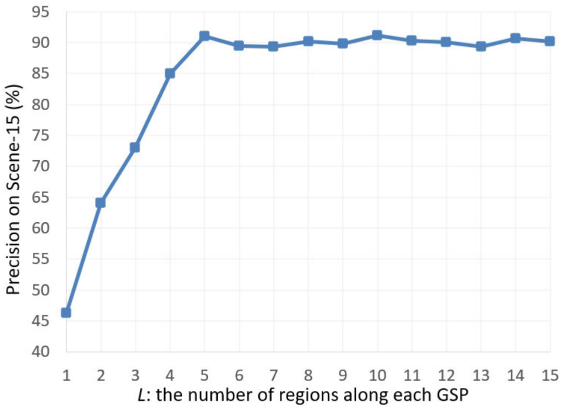

FIGURE 5. Categorization precision by adjusting *L*.

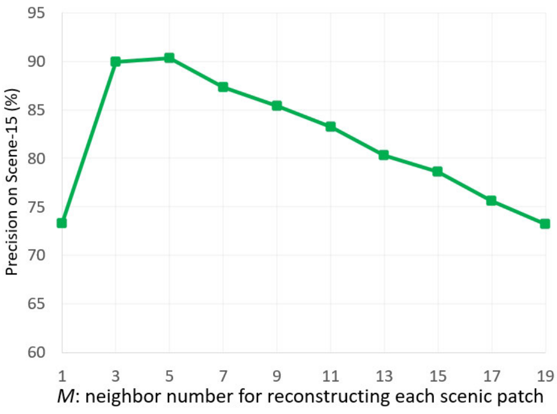

FIGURE 6. Categorization precision by adjusting *M*.

increases and reaches its peak when L falls within the range of three to five. Subsequently, the accuracy steadily declines. It means that having three or five spatially adjacent scenic patches is can effectively reconstruct each scenery. In our observations on Scene-15, we noticed that, following the extraction of scenic patches, each patch tends to be spatially neighboring to an average of three to five scenic patches. This implies that employing three to five neighbors toward a target patch is adequate. Additionally, as illustrated in Fig. 6, if an excessive number of potentially irrelevant scenic patches is considered, the reconstruction precision decreases. Also, it spends more time. Thirdly, we examine the impact of α, β, and γ on categorizing sceneries. We maintain all of them at 0.1 and adjust each one individually. In particular, we vary α from 0 to 0.95. As shown in Table 8, scene classification accuracies consistently increase and reache its peak at $\alpha = 0.25$. Subsequently, the performance experiences a significant decline. The potential reason is that enhancing one

TABLE 8. Categorization precision by adjusting α.

α	Accuracy	α	Accuracy
0	72.41%	0.55	72.43%
0.05	80.34%	0.6	68.12%
0.1	81.89%	0.65	67.43%
0.15	85.32%	0.7	63.21%
0.2	88.21%	0.75	58.35%
0.25	90.08%	0.8	55.66%
0.3	86.54%	0.85	51.35%
0.35	86.02%	0.9	47.42%
0.4	85.23%	0.95	45.32%
0.45	83.46%	1	44.33%
0.5	76.23%		

TABLE 9. Categorization precision by adjusting β.

β	Accuracy	β	Accuracy
0	73.47%	0.55	78.63%
0.05	82.03%	0.6	76.32%
0.1	84.33%	0.65	75.35%
0.15	85.65%	0.7	73.05%
0.2	87.16%	0.75	73.43%
0.25	88.73%	0.8	74.53%
0.3	90.21%	0.85	71.67%
0.35	86.32%	0.9	70.44%
0.4	83.55%	0.95	68.32%
0.45	82.73%	1	67.49%
0.5	80.35%		

TABLE 10. Categorization precision by adjusting γ.

γ	Accuracy	γ	Accuracy
0	76.37%	0.55	77.94%
0.05	81.63%	0.6	76.21%
0.1	87.30%	0.65	74.63%
0.15	88.48%	0.7	73.58%
0.2	90.30%	0.75	72.32%
0.25	87.64%	0.8	74.64%
0.3	85.64%	0.85	72.46%
0.35	84.52%	0.9	72.25%
0.4	83.32%	0.95	71.03%
0.45	82.51%	1	70.65%
0.5	81.66%		

term helps the overfitting phenomenon. However, excessive emphasis on this term will apparently diminish the influences of sparisity control and the semantics of scenic patches. Based on this, we set $\alpha = 0.25$. Simultaneously, we report the scenery categorization by tuning β and γ individually. The performances are detailed in Tables 9 and 10. Similar to the evaluation of α, we respectively determine the optimal β and γ as 0.3 and 0.2.

V. CONCLUSION

Effectively categorizing scenes into distinct classes has significant value in many AI applications. This work introduces a novel method, called the robust deep active learning (RDAL), which learns a descriptive image kernel by jointly uncovering and representing human gaze shifting. Starting with an large collection of scene images, we employ a local to global feature fusion to combine dfiferent features for characterizing each region. Subsequently, the RDAL algorithm is employed to identify visually and semantically attractive regions within each scenic picture, constructing a gaze shifting path (GSP), and calculating its deep representation. Lastly, the deep GSP representations are encoded into a kernelized vector for scenery recognition. Plenty of testing results showed the effectiveness of our biologically-guided deep categorization pipeline.

Noticeably, this paper cannot handle other technical challenges like the spectral discrepancy between aerial photos. In the future, we plan to build a unified and comprehensive low-resolution aerial photo understanding system supporting many modules, each of which can handle one of the aforementioned problems.

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] 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.

[6] 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.

[7] 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, pp. 1–5.

[8] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proc. ACM Workshops Multimedia*, Nov. 2000, pp. 11–17.

[9] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. NIPS*, 2014, pp. 44–51.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 711–723.

[11] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, 2006, pp. 4–12.

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

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

[14] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 51–56.

[15] 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.

[16] 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. 1–8.

[17] 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.

[18] Z. Harchaoui and F. Bach, "Image classification with segmentation graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 102–110.

[19] 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.

[20] 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.

[21] 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. 22–31.

[22] 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. 47–55.

[23] 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. 62–71.

[24] D. Song and D. Tao, "Biologically inspired feature manifold for scene classification," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 174–184, Jan. 2010.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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, pp. 321–331.

[36] 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, pp. 53–71.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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, pp. 43–50.

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. ACM MM*, 2010, pp. 7–18.

[52] 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.

[53] Y. Yu, X. Yang, J. Li, and X. Gao, "Object detection for aerial images with feature enhancement and soft label assignment," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5624216.

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

[55] 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.

[56] M. Stricker and M. Orengo, "Similarity of color images," in *Storage and Retrieval of Image and Video Databases*, 1995.

[57] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Vis. Pattern Recognit. (CVPR)*, 2005, pp. 13–20.

TENGSHENG LIU is currently a Faculty Member of the Department of Physical Education, Wuhan Institute of technology. His research interests include multimedia, computer vision, and image processing.

MINGHUI XU is currently a Faculty Member of Jinhua Polytechnic. Her research interests include artificial intelligence and machine learning.