IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 July 2024, accepted 28 August 2024, date of publication 2 September 2024, date of current version 10 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3452981

RESEARCH ARTICLE

Perceptual Feature Integration for Sports Dancing Action Scenery Detection and Optimization

LINGJUN XIANG[1] AND XIANG GAO[2]
[1]Department of Physical Education and Research, Hunan Institute of Technology, Hengyang 421002, China
[2]Sports Training College, Guangzhou Sport University, Guangzhou 510500, China

Corresponding author: Lingjun Xiang (2015001997@hnit.edu.cn)

This work was supported in part by Hunan Province General Higher Education Teaching Reform Research Project (Xiang Jiao Tong [2023] No. 352) under Project HNJG-20231319, and in part by the Scientific Research Project of Hunan Provincial Department of Education (Xiang Jiao Tong [2023] No. 361) under Project 23C0405.

ABSTRACT Deciphering the complex semantics within varied dancing sceneries is crucial for a multitude of AI endeavors. It can facilitate applications like dancing action optimization and dancing education. In our research, we propose a sophisticated approach to discerning multi-faceted perceptual visual features for accurately identifying dancing scenic imagery with intricate spatial designs. Our work centers on crafting a deep hierarchical structure adept at simulating human gaze patterns, utilizing the BING metric to pinpoint objects and their components within scenes at different scales. To emulate human visual dynamics, we introduce a Robust Deep Active Learning (RDAL) methodology, systematically creating gaze shift paths (GSPs) and capturing their profound representations. A key novelty of RDAL is its resilience to inaccuracies in labeling, employing a strategically designed sparse penalty framework that facilitates the exclusion of non-informative or irrelevant deep GSP attributes. Furthermore, we propose a manifold-regularized feature selector (MRFS) to isolate premium deep GSP features, concurrently developing a linear SVM for dancing scene recognition. Our method's efficacy, validated through rigorous testing, not only showcased its enhanced performance across conventional scenic datasets but also highlighted the exceptional discriminating power of deep GSP features in a specialized dataset for recognizing different dancing actions. Finally, the dancing actions can be optimized using a probabilistic model.

INDEX TERMS Perceptual, dancing action, manifold-regularized, active learning, deep architecture.

I. INTRODUCTION

In the domain of sophisticated AI technologies, the capacity for assigning multiple descriptors to every scene plays a critical role. For example, enhancing the efficiency of route planning in intelligent navigation systems necessitates the utilization of various scene-centric details, such as the configuration of transportation networks, the alignment of roadways, and the traits of the cityscape. Additionally, contemporary public safety infrastructures depend on recognizing distinct elements within scenes, like road signage and inclines, to augment real-time surveillance of pedestrian and vehicular movements. It has been noted that intersections are notably more prone to vehicular incidents compared to straight segments of roads. Therefore, accurately distinguishing among various types of scenes allows for the focused implementation of multi-camera surveillance networks at crucial junctions, enabling comprehensive monitoring of atypical interactions between vehicles and pedestrians. Within the scope of visual categorization and labeling, cutting-edge algorithms have been crafted to depict the complexity of scenic imagery across different scales. These advanced methods include: 1) leveraging Multiple Instance Learning (MIL) and CNNs for region pinpointing via weak supervision [42], [43]; 2) employing semantic graph structures for detailed scene interpretation [47], [48]; and 3) constructing layered frameworks for the targeted labeling of scenic photographs [44], [45], [46]. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Wei-Wen Hu.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

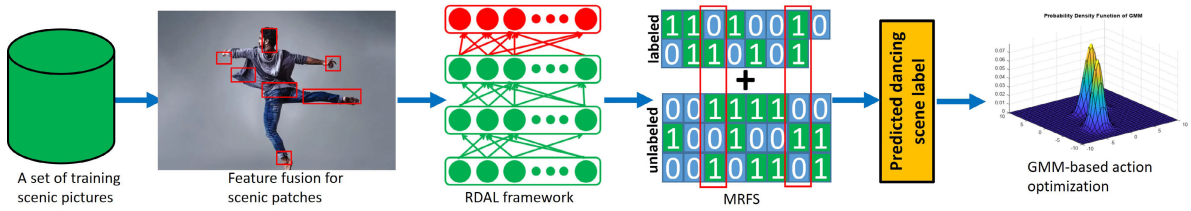

FIGURE 1. An overview of our designed dancing scenery categorization and optimzing by perceptual features integration.

these advancements still face challenges in capturing the full essence of dancing scenic images, encountering key obstacles:

- Identifying specific focal objects or segments within high-resolution dancing images requires an approach that mirrors the human visual system's capacity to zero in on visually or semantically important regions. The quest is to craft a model adept at discerning and representing these focal areas accurately. This includes capturing gaze shift paths (GSP) that mirror shifts in human attention throughout relevant parts of dancing images, navigating through label noise common in large-scale datasets, and embedding semantic labels at the patch level to authentically convey scene content;
- Significant dancing scene elements are frequently highlighted by varied low-level indicators, each spotlighting distinct aspects of the dancing scenery. Creating an effective integration of these indicators demands a structured approach to evenly distribute the impact of each feature set, made more challenging by the need to adjust feature channel importance in response to the diversity of dancing scenic datasets.

To address these challenges, our approach introduces a pioneering framework for dancing scene classification that incorporates deep and proactive analyses of human gaze behavior. Our method begins by applying the BING objectness measure [56] to identify object-related patches across a broad spectrum of scenic images, potentially including inaccurately labeled examples (Sec III-A). To better align our model with human visual mechanisms, we integrate a Robust Deep Active Learning (RDAL) system. RDAL precisely calculates the human gaze shift path (GSP) and its in-depth characterizations, tackling issues of label inaccuracies and redundancies head-on. RDAL employs a semi-supervised learning approach, utilizing a limited selection of semantic labels for initial training (Sec III-B). Subsequently, a manifold-regularized feature selection (MRFS) strategy is deployed to isolate highly distinct deep GSP attributes, upon which a linear SVM is formulated for dancing scene classification (Sec III-C). Extensive tests conducted on six publicly accessible datasets and our exclusive dancing action dataset have validated the effectiveness and superiority of our method, underscoring the advantages of integrating gaze-driven analysis into scene classification efforts. Moreover, we propose a Gaussian mixture model (GMM) for optimizing different dancing actions, wherein competitive performance can be observed (Sec IV-D).

This research holds two significant contributions: the creation of the Robust Deep Active Learning (RDAL) system, which intricately charts human gaze movements while extracting related visual features, and the development of a selective feature evaluation method, MRFS, that dynamically gauges the significance of diverse feature channels relative to deeply examined GSP attributes.

The organization of this document is: Sec II reviews the pertinent literature. The next section describes our sports dancing scenery detection/optimization framework, highlighting three fundamental elements: 1) the effective extraction of BING object patches from each dancing scenery, 2) the application of the RDAL technique for GSP extraction and deep learning, and 3) the use of a unified feature selection and classification framework. Sec IV showcases the empirical validations that underline our method's improved effectiveness. Sec V concludes the paper with a summary of findings and future perspectives.

II. PREVIOUS WORK REVIEW

The field of computer vision has experienced substantial progress through the incorporation of deep learning techniques specifically tailored for scene analysis. At pthe forefront of these advancements are hierarchical Convolutional Neural Networks (CNNs) and elaborate models, skilled in analyzing extensive image collections, especially noted with ImageNet [34]. Research such as [11] has demonstrated the high precision of these models in scene classification, utilizing subsets of ImageNet. Though originally broad in application, the intricate feature-extraction capabilities of ImageNet-CNNs have significantly contributed to various tasks in computer vision, ranging from video analysis to spotting anomalies. Over the last decade, enhancements in ImageNet-based CNNs have mainly concentrated on enlarging training sets and advancing model structures. Methods like selective search [35] have played a crucial role in assembling extensive, category-agnostic patch sets by merging search strategies with semantic annotations. R-CNNs [36] have stressed on fetching quality patch samples for in-depth image insights. Moreover, the creation of expansive, scene-oriented datasets for training [10] and employing pre-trained hierarchical CNNs for identifying and representing localized scene aspects [38] mark notable advancements. Recent developments have also explored multi-task and multi-resolution scene classification strategies. These include using manifold-based regularization [4] and blending low-rank

feature learning with Markov models for deeper semantic analysis [5], focusing on preserving feature distributions. Innovations in unsupervised learning for deep feature extraction from scenes [6], independent of annotated data, have paved new paths for model training. Integrated methods that combine discriminative feature learning with weak label approaches, utilizing dense sparse autoencoders for enhanced visual representations, have broadened the scope for scene analysis [40].

Aerial imagery analysis has significantly gained from advanced computational models utilizing state-of-the-art machine learning techniques. A multi-modal learning approach for annotating high-resolution aerial images was introduced in [41], heralding a leap forward in the domain. Meanwhile, [19] delved into a novel multi-attention mechanism for evaluating feature relevance within aerial images, proving its effectiveness across different image resolutions. Nonetheless, applying these models to low-resolution images faces challenges, especially in accurately identifying small, essential objects that appear blurred. Overcoming these challenges necessitates a focus on region-level modeling to ensure accurate object detection and localization within low-resolution aerial images. In enhancing facial recognition, [67] proposed a group sparsity regularizer to refine the l_1-norm for reducing bias and diminishing outlier effects. The issue of incomplete multi-view clustering was tackled by [37] through improving incomplete similarity graphs and creating detailed tensor representations. For detailed regional analysis of aerial photos, [17] suggested a multi-layered deep learning approach for object detection across scales. Aiming for precision in vehicle localization within both low and high-resolution aerial images, a deep learning model based on focal loss was presented in [58]. Additionally, [66] introduced a geographic object detection model for high-resolution images, focusing on extracting key features like intersections and roads. Finally, [65] combined feature engineering with soft-labeling techniques to develop a durable visual detection framework tailored for aerial imagery analysis.

III. FRAMEWORK FOR CLASSIFYING SCENERY

A. HIGHLIGHTING SEMANTICALLY ESSENTIAL AREAS

Studies within the realms of visual cognition and psychology [49], [50] have consistently demonstrated a fundamental aspect of human behavior: the predilection for concentrating on the most semantically or visually pivotal parts of a scene. Such research elucidates that human focus is selectively aimed at particular zones deemed crucial for understanding, rather than being uniformly dispersed across the entire visual field. In light of this insight, we propose an advanced technique that merges the detection of object-specific patches with a Robust Deep Active Learning (RDAL) strategy. Our goal is to pinpoint and scrutinize those dancing scene segments most likely to attract human attention.

Observational data validate that the human visual apparatus gives precedence to regions populated by semantically dense or visually compelling objects, such as dancer arms or feet, which significantly shape scene perception through their presence and layout. To adeptly delineate these key areas, we adopt the BING objectness metric [56], acclaimed for its capability to efficiently segregate distinct, object-related patches within varied scenery settings. The BING technique is distinguished for several reasons: its unrivaled accuracy in singling out pertinent patches with minimal processing requirements, its enhancement of the Gaze Shift Paths (GSP) identification process by supplying superior object-centric patches, and its exceptional adaptability to different object classes beyond its initial training scope. These characteristics guarantee the adaptability and effectiveness of our dancing scenery classification framework across a wide array of visual datasets.

B. DEPLOYING ROBUST DEEP ACTIVE LEARNING (RDAL)

By applying the BING algorithm [56], a vast array of object-oriented patches, ranging from the hundreds to thousands, can be identified within diverse dancing scenic contexts. It's imperative to acknowledge, however, that human attention is often drawn to a limited set of features in a dancing scene, reflecting a more discerning pattern of observation. In response to this phenomenon, we deploy an innovative Robust Deep Active Learning (RDAL) strategy designed to select an optimal subset of dancing scenic patches, denoted as L, for the creation of Gaze Shift Paths (GSP) and the extraction of their deep representations. The RDAL approach distinctively amalgamates key factors: the geometric configuration of dancing scenes, the inherent semantic value of certain object patches, and the obstacle presented by inaccurate semantic labels. This integrated tactic guarantees a holistic and precise dancing scene portrayal, in line with human visual preferences.

1) EVALUATING THE SPATIAL DYNAMICS OF DANCING SCENERY

The efficacy of scenery classification hinges on a thorough comprehension of the scene's spatial dynamics, including the arrangement of elements in the foreground and background. This comprehension requires a methodology that gauges the importance of each dancing scenic patch based on its spatial relation to adjacent patches. Throughout this process, the relevance of individual object patches is determined via an optimization method that accounts for the spatial links between patches. Such an approach affords a detailed representation of dancing scenic structures, pivotal for accurate dancing scenery classification.

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

Within this framework, the collection $z_1, \cdots, z_N \in \mathbb{R}^{N \times A}$ represents the array of deep features derived from N identified scenic patches through the utilization of the BING

algorithm [56]. Here, A denotes the complexity or depth of the feature representation for each patch, and the matrix $\mathbf{F}_{ij}$ measures the impact of the i-th patch on the reconstruction of the j-th patch. Additionally, $\mathcal{N}(z_i)$ encompasses the patches in close spatial proximity to the i-th patch, underlining the scene's internal spatial coherence.

2) SEMANTIC SIGNIFICANCE OF DANCING SCENIC PATCHES

In addition to the spatial arrangement of each dancing scene, the intrinsic semantic content of the selected patches is integral for the formulation of Gaze Shift Paths (GSPs). Utilizing the reconstruction discrepancy as specified in (1), the reformulated scenic patches are represented as $g_1, \cdots, g_N$. The determination of the L patches endowed with the highest semantic content is directed by the optimization of the ensuing equation:

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

In this scenario, τ serves as the regularization parameter, and the set $g_{q_1}, \cdots, g_{q_L}$ includes the L scenic patches identified by our RDAL method. The goal of this equation is to reduce the difference, focusing on the spatial characteristics of the selected patches. It also preserves the semantic coherence of the dancing scenic patches being reconstructed, ensuring they closely resemble their original forms. Successfully minimizing (2) yields a collection of dancing scenic patches that accurately reflect human scene perception in terms of both visual and semantic aspects.

For the objective of analysis, we introduce matrices $\mathbf{A} = [z_1, \cdots, z_N]$ and $\mathbf{H} = [g_1, \cdots, g_N]$. Meanwhile, matrix Δ, an $N \times N$ diagonal matrix, denotes the selection of dancing scenic patches, with Δ_{ii} assigned a value of 1 for $i \in \{q_1, \cdots, q_L\}$, symbolizing the chosen patches, and 0 for all other instances. This setup allows for the enhancement of the objective function (2) as follows:

$$\eta(\mathbf{Q}) = \mathrm{tr}((\mathbf{H} - \mathbf{A})^T \Delta (\mathbf{H} - \mathbf{A})) + \tau \mathrm{tr}(\mathbf{H}^T \mathbf{L} \mathbf{H}), \tag{3}$$

In this context, the formulation $\mathbf{L} = (\mathbf{I} - \mathbf{F})^T (\mathbf{I} - \mathbf{F})$ is established. To optimize (3), the gradient of $\eta(\mathbf{H})$ is equated to zero, leading to the following condition:

$$\Delta(\mathbf{H} - \mathbf{A}) + \tau \mathbf{L} \mathbf{H} = 0. \tag{4}$$

In this scenario, the reconstruction of scenic patches is determined as follows:

$$\mathbf{H} = (\tau \mathbf{L} + \Delta)^{-1} \Delta \mathbf{A}. \tag{5}$$

Utilizing the reconstructed scenic patches, the reconstruction error can be refined as follows:

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

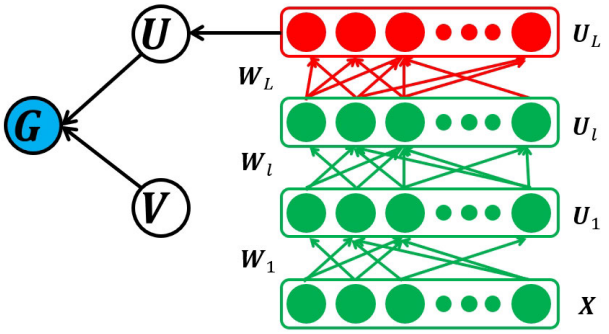

FIGURE 2. Architecture of the intricately and semantically encoded Gaze Shift Path (GSP).

3) THE RDAL METHODOLOGY

Our strategy incorporates a multi-layer method to uncover the visual elements characterizing each dancing scene, employing a profound learning framework to decode the complex features associated with various dancing sceneries. As depicted in Fig. 2, the RDAL architecture utilizes an R-level deep structure to methodically decompose the semantic label matrix $\mathbf{G}$ into a sequence of $R + 1$ matrices, inclusive of $\mathbf{V}$ and $\mathbf{U}_R$ to $\mathbf{U}_1$. This layered approach enables the precise extraction of deep scenic features and the nuanced representation of new dancing scenic visuals, starting with $\mathbf{U}_1 = \mathbf{W}_1 \mathbf{X}$ at the initial layer.

At the heart of the RDAL framework is the utilization of linear combinations in a sequence to intricately develop the latent attributes that are fundamental to each dancing scenery. This core principle allows for a thorough depiction of dancing scene characteristics without resorting to complex equations. Consequently, the multi-layered architecture of the deep learning model is efficiently outlined:

$$\begin{aligned} \mathbf{G} &\leftarrow \mathbf{P}\mathbf{Q}_R, \\ \mathbf{Q}_R &= \mathbf{U}_R \mathbf{P}_{R-1}, \\ &\cdots \\ \mathbf{Q}_1 &= \mathbf{U}_1 \mathbf{Y}, \end{aligned} \tag{7}$$

Within our model, $\mathbf{U}_i$ is designated as the transformation matrix for the i-th layer, while $\mathbf{P}$ denotes the matrix of semantic labels, which remain indirectly observed. The matrix $\mathbf{Q}_i$ illustrates the dancing scene's representation at the i-th deep layer. Furthermore, $\mathbf{Y}$ is comprised of y_i, representing the B-dimensional comprehensive feature for each scenic patch. Under the Robust Deep Active Learning (RDAL) framework, the most profound representation achieved at the highest layer is represented by $\mathbf{Q} = \mathbf{Q}_L$. As specified in (7), the training segment of our deep learning structure focuses on deriving the latent factor $\mathbf{P}$ along with the sequence of transformation matrices $\mathbf{U}_R, \cdots, \mathbf{U}_1$, facilitating a layered and intricate insight into scene dynamics.

In conclusion, the comprehensive deep-model-driven active learning process can be mathematically depicted as

follows:

$$\min_{\mathbf{P},\Delta\mathbf{U}_1,\cdots,\mathbf{U}_R} \frac{1}{2}||\mathbf{F}-\mathbf{PQ}||_F^2+\frac{\alpha}{2}||\mathbf{P}||_F^2+\frac{\alpha}{2}\sum\nolimits_{i=1}^{R}||\mathbf{U}_i||_F^2 + \frac{\beta}{2}||\mathbf{U}||_{2,1}, \tag{8}$$

Within this framework, the matrix $\mathbf{F} \in \mathbb{R}^{R\times N}$ encapsulates the semantic labels, where $\mathbf{F}_{ij} = 1$ signifies the association of the i-th scenic image with the j-th label, and $\mathbf{F}_{ij} = 0$ denotes no such link. Here, R represents the total count of distinct semantic labels, α serves as the regularization parameter to mitigate overfitting risks, and β ensures sparsity across the columns of U_i. Given the potential for visual features to be interrelated, duplicative, or adversely affected by noise, adopting a sparse representation through the l_{21}-norm becomes crucial. This approach effectively filters out low-quality, noisy features.

It's important to underscore that, in contrast to the initial two visual features namely, the spatial organization of dancing scenery and the semantic delineation at the patch level. RDAL strategy is implemented within a semi-supervised framework. This means that the model's training leverages only a limited set of semantic labels, as specified in (8). This method is particularly beneficial for processing numerous images where comprehensive semantic labeling is unfeasible due to the prohibitive demands of manual annotation.

C. MANIFOLD FEATURE SELECTION AND CLASSIFICATION

While the deeply analyzed Gaze Shift Path (GSP) features are rich in information, they introduce complexities that must be navigated to improve efficacy: 1) The feature set can expand to a prohibitive dimensionality, notably with a significant K, resulting in the accumulation of numerous significant image patches. This expansion, especially to dimensions around $128K$, can lead to the curse of dimensionality during the training of classifiers, necessitating a further compression of the deep GSP feature dimensions. 2) The limited availability of labeled samples, due to constrained annotation efforts, calls for a model adept at leveraging both labeled and unlabeled data for learning. Our aspiration is not limited to merely reducing dimensionality but also encompasses the learning of a classifier for scenic imagery. This ambition motivates the adoption of a semi-supervised feature selection method described hereafter. 3) The matrix $\mathbf{C} = [c_1, \cdots, c_N] \in \mathbb{R}^{N\times T}$ is defined to contain the deep GSP features for all training samples. For ease of explanation, it's assumed that the first L scenic images are labeled ($\mathbf{C}_L = [c_1, \cdots, c_L]$), while the remaining are unlabeled ($\mathbf{C}_U = [c_{L+1}, \cdots, c_N]$). The label matrix for these L labeled instances is represented by $\mathbf{M} = [m_1, \cdots, m_L]$.

By applying manifold learning principles [8], we proceed to construct a semi-supervised Support Vector Machine (SVM) framework tailored for this scenario.

$$\|\Phi\|_2^2 = \sum\nolimits_{i=1}^{N}\sum\nolimits_{j=1}^{N}(\phi(c_i)-\phi(c_j))^2\mathbf{N}_{ij} = \Phi^T\mathbf{A}\Phi, \tag{9}$$

In this setup, $\mathbf{N}_{ij}$ denotes the strength of connectivity between samples c_i and c_j, serving as an indicator of their resemblance. The SVM's decision function for the entire sample set is represented as $\Phi = [\phi(c_1), \cdots, \phi(c_N)]$. For constructing the graph Laplacian, a key concept in graph theory, we employ $\mathbf{A} = \mathbf{T} - \mathbf{N}$, where $\mathbf{T}$ is a diagonal matrix with each diagonal element, $\mathbf{T}_{ii}$, being the aggregate of the weights $\mathbf{N}_{ij}$ corresponding to sample i, detailed by $\mathbf{T}_{ii} = \sum_{j=1}^{N}\mathbf{A}_{ij}$. As outlined in [12], the norm $||\cdot||_2^2$ evaluates the continuity of the decision function across the dataset of c_i.

Considering a linear SVM, wherein the decision function for a given sample c_i is $\phi(c_i) = \mathbf{q}^T c_i - b$, this regularization approach is refined to $||\Phi||_2^2 = \mathbf{q}^T\mathbf{C}^T\mathbf{ACq}$. It is important to note that the bias b is excluded from the manifold regularization computation. With these frameworks in mind, the semi-supervised SVM model can be precisely formulated as follows:

$$\begin{aligned}&\min_{\mathbf{w},b,\xi}\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{U}\xi_i+\frac{\lambda}{2}\mathbf{w}^T\mathbf{C}^T\mathbf{ACw},\\ &\text{s.t. } y_i(\mathbf{w}^T x_i - b)\geq 1-\xi_i, \xi_i\geq 0,\quad i=1,\ldots,U,\end{aligned} \tag{10}$$

In this structure, ξ_i is the marginal error, with $\lambda \geq 0$ serving as the balance parameter adjusting the impact of both previously mentioned regularizers. This objective function enables the preservation of sample distribution integrity, particularly maintaining the spatial positioning of samples within the dataset.

Our semi-supervised Support Vector Machine (SVM) model further incorporates the aspect of semi-supervised feature selection (FS). For this endeavor, we define $\mathbf{h} = (h_1, \cdots, h_T)^T$, where each $h_i \in \{0, 1\}$ denotes the inclusion or exclusion of a feature. To streamline this process, a diagonal matrix $\mathbf{E}(\mathbf{h}) = \text{diag}(h_1, \cdots, h_T)$ is employed, which filters the representation of input samples through $\mathbf{CE}(\mathbf{h})$. Given the selection of A features, the constraint $\mathbf{h}^T\mathbf{e} = A$ confirms the designated number of feature selections. Leveraging this approach, (3) is further refined as follows:

$$\begin{aligned}&\min_{\mathbf{w},b,\xi,\mathbf{h}}\frac{1}{2}\|\mathbf{w}\|_2^2 + C\sum_{i=1}^{U}\xi_i+\frac{\lambda}{2}\mathbf{w}^T\mathbf{E}(\mathbf{h})^T\mathbf{C}^T\mathbf{ACE}(\mathbf{h})\mathbf{w},\\ &\text{s.t. } y_i(\mathbf{w}^T x_i - b)\geq 1-\xi_i, \xi_i\geq 0,\quad i=1,\ldots,U,\\ &\mathbf{h}^T\mathbf{1} = A.\end{aligned} \tag{11}$$

This equation is reformulated into a min-max optimization problem, as elaborated in [12].

By summarizing the discussions aforementioned, the pipeline of the dancing action scenery detection and optimization is provided in Alg. 1.

IV. EVALUATION OF EXPERIMENTAL PERFORMANCE

This section is dedicated to evaluating our dancing scene classification framework, which utilizes Robust Deep Active Learning (RDAL). We conduct this evaluation through four distinct experimental settings. We begin by describing the

Algorithm 1 Our Proposed Dancing Action Sceneries Detection and Optimization

input: A rich set of training dancing scenic pictures, the salient object patch number L, the iteration number of RDAL, and parameter τ and t;
output: The learned dancing scene classifier and GMM;

1) Use the BING algorithm to generate a rich set of BING object patches;
2) Leverage our RDAL to select L visually/semantically salient object patches, based on which the GSP are extraction and its deep features are calculated simultaneously using (8);
3) Utilize our manifold feature selector to acquire high-quality deep GSP features for dancing scenery classification by (11);
4) Train the GMM for dancing action optimization by (12).

experimental configurations and introduce six benchmark datasets that form the basis of our assessment. Subsequently, a comparative analysis is performed, where the efficacy of our RDAL-based model is contrasted against an array of scene recognition models, spanning both shallow and deep learning approaches. Following this comparative study, we explore the key factors that contribute to the success of our RDAL strategy. Lastly, we illustrate the utility of deep Gaze Shift Path (GSP) features, derived from our approach, in the specific domain of sports scenery classification, showcasing their impact on improving scene categorization.

A. DATASETS AND EXPERIMENTAL CONFIGURATION

Our evaluation of the categorization framework is carried out through extensive testing on six diverse scenic image datasets, which include both established benchmarks and more recent compilations. Representative images from these datasets are presented in Fig.3, demonstrating the variety of scenes encompassed. Key among these are two fundamental datasets, Scene-15 [13] and MIT Indoor Scene-67 [14], utilized as benchmarks to gauge our model's efficacy.

- Scene-15: This dataset encompasses 15 varied categories, with 13 originally compiled by Feifei [16]. It features 200 to 400 images per category, with an average resolution of 320 × 250 pixels. The images are primarily sourced from the COREL database, augmented by contributions from individual photographers and Google Images.
- Scene-67: Dedicated to indoor scenes, this expansive dataset covers a broad spectrum of interior spaces across 67 categories. It is meticulously compiled from Picasa and Altavista for a diverse range of indoor settings, specialized photography sharing sites, and the extensive LabelMe database, offering a detailed view of various indoor living and public spaces.

Moreover, our assessment extends to four more modern scenic image collections: ZJU aerial imagery [3], ILSVRC-2010 [34], SUN [39], and Places [10], each contributing a unique angle on scenic diversity. In parallel, we introduce a novel dataset tailored for sports education, termed the Large-scale Dancing Action (LSDA) dataset. This exclusive collection boasts around 2,300,000 images across 22 dancing categories, as detailed in Fig. 4. An overview of this extensive collection is depicted in Fig.4, with detailed statistics and categorizations provided in Table1, illustrating the dataset's scope and its significance in educational scenarios. Before proceeding with the evaluation of baseline models, we outline the settings of our empirical analysis, designed for a thorough and equitable assessment of all algorithms under consideration:

TABLE 1. Characteristics of our LSDA image collection.

Sport name	Training image #	Training image #
Ballet	83021	23121
Swing	74343	25330
Tap dance	73021	29843
Ballroom	83243	24394
Belly dance	65993	34220
Irish dance	68321	21203
Folk dance	73421	25436
Hip pop	74355	24453
Modern dance	85464	23112
Salsa	80334	21332
Aerial dance	78798	24365
Contemporary	78879	24344
Jazz dance	87545	25543
Latin dance	72335	33225
Interpretive	73334	25476
Rapper dance	77668	24436
Waltz	78779	27668
Animal dance	81224	25557
Azerbaijani	83043	298321
Bharatnatyam	79910	22231
Tango	76576	29332
Samba	68879	32311

1) Object Patches Configuration: By employing the BING algorithm [56], we standardize the extraction of scenic patches to 1000 per dataset across all six scenic image datasets. This uniformity ensures comprehensive object detection within the scenes.

2) Spatial Neighbors Setting: The count of spatial neighbors, indicated by L, is fixed at five for every object patch. This configuration mirrors the human visual system's preference for concentrating on a select group of significant areas within a scene.

3) Low-Level Feature Extraction: We integrate three specific low-level features to encapsulate the characteristics of each object patch: a 16-dimensional color moment [68], a 64-dimensional Histogram of Oriented Gradients (HOG) [69], and a 160-dimensional combination of edge and color histograms [9]. These features are chosen for their effectiveness in capturing essential visual attributes.

4) GSP Internal Regions Count: The number of internal regions within a Gaze Shift Path (GSP) is established at five, noted as K. This figure is reflective of the general tendency for human viewers to focus on a maximum of five key areas within a scene, grounding our model's structure in real-world observation patterns.

5) Patch-Level Deep Feature Dimensionality: The dimensionality of deep features at the patch level is defined to

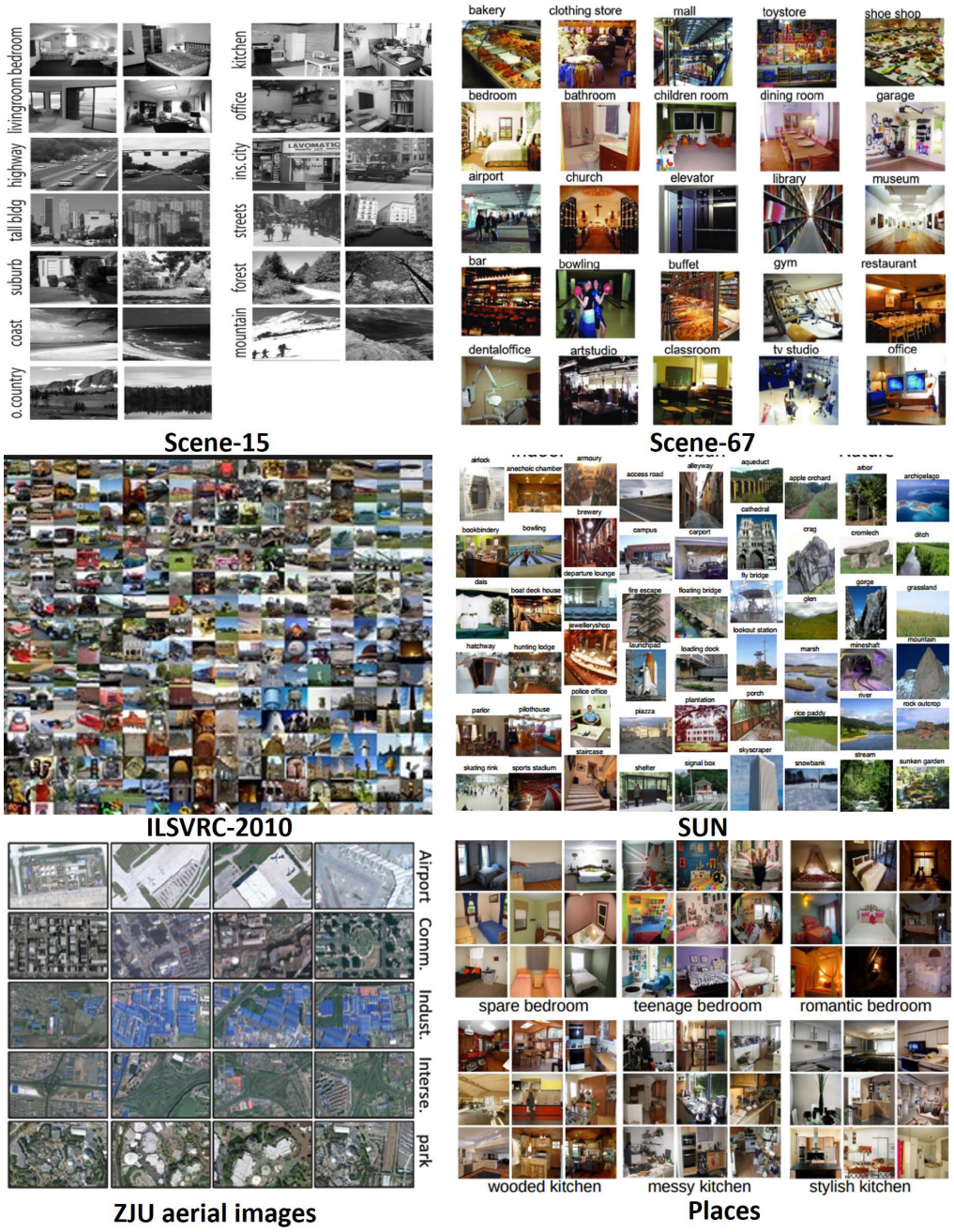

FIGURE 3. Sample images from the aforementioned six scene datasets.

be 212. This dimensionality is set to ensure uniformity and manageability of the feature sets for classification processes.

These configurations are carefully selected to emulate human visual perception and enhance the efficacy of our scene categorization model in various visual scenarios.

B. COMPARATIVE ANALYSIS WITH EXISTING MODELS

1) TASK OF SCENERY CLASSIFICATION

This segment of our study positions the efficacy of our perception-driven scene classification framework against four traditional shallow classification schemes: Fixed-Length Walk Kernel (FWK) and Tree Kernel (FTK) [20]: These kernels are adept at recognizing structural image patterns, with FTK broadening FWK's applicability to hierarchical data structures. Multi-Resolution Histogram (MRH) [27]: Employing multi-scale texture analysis, MRH provides a nuanced approach to texture-based scene classification. Spatial Pyramid Matching with Kernel Techniques (SPM): This includes three variants - Locality-constrained Linear Coding plus SPM (LLC-SPM) [21], Sparse Coding plus

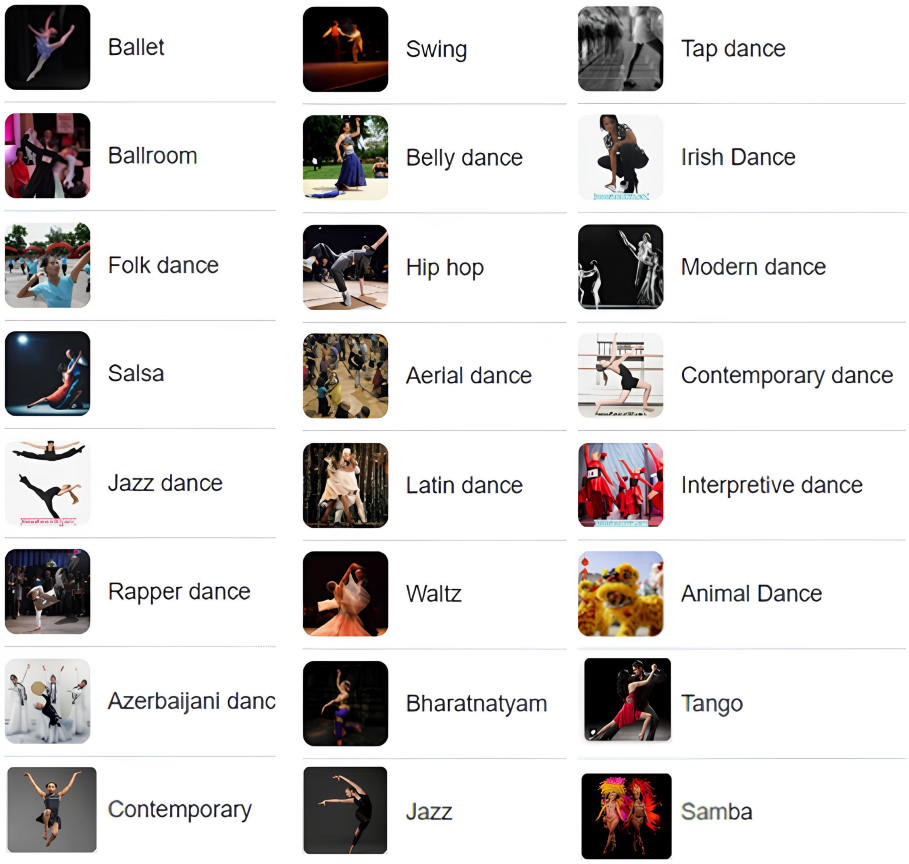

FIGURE 4. Sample images from our large-scale dancing action (LSDA) image collection.

SPM (SC-SPM) [22], and Object Bank plus SPM (OB-SPM) [23], each enhancing SPM through distinct feature coding strategies to better represent scenes. Super Vector Coding (SVC) and Supervised Image Coding (SSC) [24], [25]: These models advance image categorization through sophisticated vector quantization and supervised learning methods, respectively.

For a fair comparison, the configurations for each classification framework were precisely standardized. Parameters for both FWK and FTK are optimized across a spectrum from two to ten to ensure maximum performance. The MRH technique is applied with RBF smoothing at twelve grayscale levels for refined texture analysis. The SPM approach and its variations analyze training images through SIFT descriptors arranged on a 16 $\times$ 16 pixel grid, followed by the creation of a 400-term codebook via k-means clustering to assemble a rich feature set specifically tailored for scene classification challenges.

With the rapid progression of multi-layer recognition technologies, a comparative study with the latest deep learning-based scene recognition frameworks was undertaken. This review encompasses notable models such as ImageNet CNN (IN-CNN) [11], Region-based CNN (R-CNN) [36], Meta Object CNN (M-CNN) [38], Deep Mining CNN (DM-CNN) [28], and Spatial Pyramid Pooling CNN (SPP-CNN) [29]. Apart from M-CNN [38], all models are accessible for direct comparison without parameter adjustments. For M-CNN [38], our process involves selecting 192 to 384 region proposals per image set via Multiscale Combinatorial Grouping (MCG) [30] and utilizing a 4096-dimensional feature vector from the FC7 layer of a comprehensive CNN [10]. Additionally, 400 superpixels per scene are generated using the SLIC algorithm [2], further processed by either a preset linear Discriminant Analysis (LDA) method (SP-LDA) or through selecting 120 visually significant patches identified by the GBV algorithm (SP-GBV). Our RDAL technique augments this process by identifying semantically and visually crucial superpixels, or Gaze Shift Paths (GSPs), from an array of low-level features, subsequently leveraging these GSPs to establish a graph-based superpixel framework that forms the basis of our scene classification kernel machine. The superiority of using BING-derived rectangular patches over superpixels is highlighted in Tables 2 and 3, underscoring the enhanced descriptiveness of our method. Additionally, our results are compared against recent developments in scene classification from Mesnil et al. [31], Xiao et al. [32], and Cong et al. [33], further affirming the comprehensive adaptability and robustness of our approach.

Besides the aforementioned 18 shallow/deep recognition models for comparison, it is necessary to compare our method with SOTA dancing movement recognition algorithms. Herein, five recently published dancing action recognition methods are employed [51], [52], [53], [54], [55]. We use the default empirical setups as in their publications. As the comparative results shown in Table 4, our method still performs the best, that is, the average categorization accuracy exceeds the second best one by 3.4%. This result again demonstrates the superiority of our method.

In examining the information presented in Tables 2 and 3, we undertook a comprehensive quantitative evaluation to contrast our model's performance with that of both contemporary deep learning-based and conventional visual recognition frameworks. This evaluation entailed executing each experiment 20 times to ensure the reliability of outcomes, with standard deviations noted to gauge the consistency of results. The findings unequivocally illustrate that our approach surpasses competing methodologies in classification accuracy and stability. Particularly noteworthy is our model's performance on the exclusive LSDA dataset, where our Robust Deep Active Learning (RDAL) strategy markedly outperforms the nearest competitor by an excess of 8% in classification accuracy. This distinction underscores the superior capability of our methodology, especially in specialized or niche datasets that demand intricate and discerning recognition capabilities.

In addition to the generic visual recognition algorithms compared above, we further compare our method with five dance movement recognition models published after 2022. Li et al. [59] explored using attitude estimation techniques to enhance dance motion recognition in dance videos, presenting innovative methods to understand complex dance movements. Bera et al. [60] focused on deep learning's role in identifying fine-grained movements in sports, yoga, and dance, offering an extensive benchmark analysis for these applications. Cheng et al. [61] improved action recognition efficiency in dance by optimizing frame feature restoration, significantly reducing the data requirement for effective learning processes. Yu et al. [62] introduced a novel synchronization framework for speech and dancing features, employing observably and unobservably learned features to enhance audio-visual representations in dance videos. Lastly, Pang and Niu [63] investigated the use of state-of-the-art AI techniques in dance motions classification, aiming to elevate the precision and analytical capabilities in interpreting dance movements. Reference [64] formulated a sophisticated model designed for dancing action categorization, leveraging hierarchical fusion techniques and adaptive graph transformers to analyze and interpret complex dance movements effectively. As shown in Table 5, our designed method performs better than the five competitors significantly on our LSDA image set, which is specifically compiled to evaluate dance action recognition. For the other generic image sets, our method's accuracies are close to its competitors.

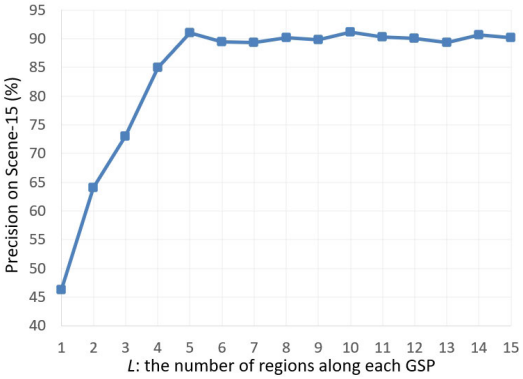

FIGURE 5. Categorization precision by adjusting L.

C. OPTIMIZATION OF PARAMETERS FOR SUPERIOR PERFORMANCE

To augment the performance of our dancing scene recognition model, which is grounded in deep learning, we undertake the fine-tuning of several pivotal parameters, aiming to elevate its efficiency in dancing scene categorization. This endeavor entails a precise adjustment and examination of these parameters to discover the most efficacious configuration. The focal points of our parameter optimization include: 1) L - Adjacency of Object Patches: The extent of neighboring patches involved in reconstructing an object patch is varied, to gauge its influence on the precision of scene recognition by the model. 2) K - Composition of GSP Object Patches: We scrutinize the impact of altering the count of object patches comprised within a Gaze Shift Path (GSP) on the model operational effectiveness. 3) Regularization Coefficients α, β, γ: These coefficients are integral to the model's regularization mechanism, delicately adjusting the weightage of diverse model facets to avert overfitting and boost the model generalizability. 4) The Scene-15 dataset [13] is primarily employed in our analysis due to its compact size, which facilitates manageability and circumvents the high computational load associated with more voluminous datasets.

Through this targeted strategy, our goal is to ascertain the optimal parameter settings that significantly enhance the model accuracy and resilience in dancing scene categorization performance. Modifying the parameter L, which represents the count of adjacent patches involved in reconstructing a specific dancing scenic patch, plays a crucial role in the refinement of our model. Preserving the local coherence of object patches is fundamental for the success of our feature integration approach. We conducted systematic evaluations of L values from one to fifteen to assess their effect on the model scene recognition capability. The findings, illustrated in Fig. 6, delineate a discernible trend: as L increases, scene recognition accuracy initially rises, reaching an optimum when L is set between three and five, and subsequently decreases with larger values.

This observation implies that a range of three to five neighboring patches is most conducive for accurately

TABLE 2. Mean classification performance of the evaluated models across the mentioned datasets.

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%
LSDA	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%
LSDA	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				
Scene-15	82.8%	86.6%	90.2%	91.2%	**93.4%**				
Scene-67	71.3%	72.1%	71.5%	74.7%	**75.6%**				
ZJU Aerial	81.1%	80.1%	78.6%	81.2%	**84.3%**				
ILSVRC-2010	40.5%	41.1%	40.8%	41.1%	**44.2%**				
SUN397	50.4%	51.2%	52.2%	52.0%	**56.3%**				
Places205	49.3%	48.2%	48.3%	49.3%	**52.1%**				
LSDA	59.7%	61.5%	62.5%	64.7%	**73.6%**				

TABLE 3. Analysis of model performances across the specified datasets.

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
LSDA	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	0.014
Data set	Xiao	Cong	Fast R-CNN	Faster R-CNN	Ours				
Scene-15	0.012	0.014	0.013	0.014	0.009				
Scene-67	0.017	0.012	0.013	0.013	0.007				
ZJU Aerial	0.014	0.013	0.014	0.012	0.008				
ILSVRC-2010	0.013	0.013	0.014	0.011	0.009				
SUN397	0.012	0.013	0.014	0.013	0.009				
Places205	0.013	0.012	0.014	0.012	0.008				
LSDA	0.014	0.011	0.015	0.014	0.006				

TABLE 4. Comparative results of our method and five SOTA dancing action recognizers on our LSDA.

Method	[51]	[52]	[53]	[54]	[55]	Ours
Accuracy	69.4%	68.9%	66.5%	70.2%	68.7%	**73.6%**

reconstructing scenes. Our detailed examination, especially using the Scene-15 dataset, indicates that scenic patches generally coexist with approximately three to five adjacent patches, affirming the appropriateness of this interval for preserving locality in scene reconstruction. Furthermore, Fig. 6 demonstrates that including too many neighboring patches might introduce extraneous noise and irrelevancy, thereby negatively impacting the model's accuracy and efficiency. Our examination further extends to the influence of the regularization coefficients α, β, and γ on the process of scene categorization. Commencing with an initial value of 0.1 for each parameter, we proceed to methodically adjust them to ascertain the most effective equilibrium. Specifically, we explore the range of 0 to 0.95 for α and assess its impact on the accuracy of scene classification. The results, as depicted in Table 6, indicate a gradual enhancement in accuracy, reaching an apex at $\alpha = 0.25$. Beyond this threshold, a notable decline in performance is observed, implying that while a modest elevation in α aids in counteracting overfitting, an excessive focus on this parameter negatively impacts the model's capacity for sparsity management and its semantic analysis of scenic patches. Thus, we identify $\alpha = 0.25$ as the ideal parameter setting.

TABLE 5. Mean classification accuracies of the evaluated models across the aforementioned datasets.

Data set	Li	Bera	Cheng	Yu	Pang & Niu	Jia	Ours
Scene-15	89.1%	89.9%	88.6%	90.1%	90.5%	89.6%	**93.4%**
Scene-67	73.2%	73.8%	72.4%	74.1%	74.8%	75.3%	**75.6%**
ZJU Aerial	82.8%	83.1%	81.5%	83.5%	83.8%	81.8%	**84.3%**
ILSVRC-2010	42.1%	42.7%	41.4%	43.1%	43.8%	42.0%	**44.2%**
SUN397	54.3%	54.6%	53.2%	55.3%	55.8%	53.1%	**56.3%**
Places205	50.1%	50.2%	49.6%	51.5%	51.8%	50.7%	**52.1%**
LSDA	58.5%	58.7%	59.3%	63.0%	65.4%	68.9%	**73.6%**

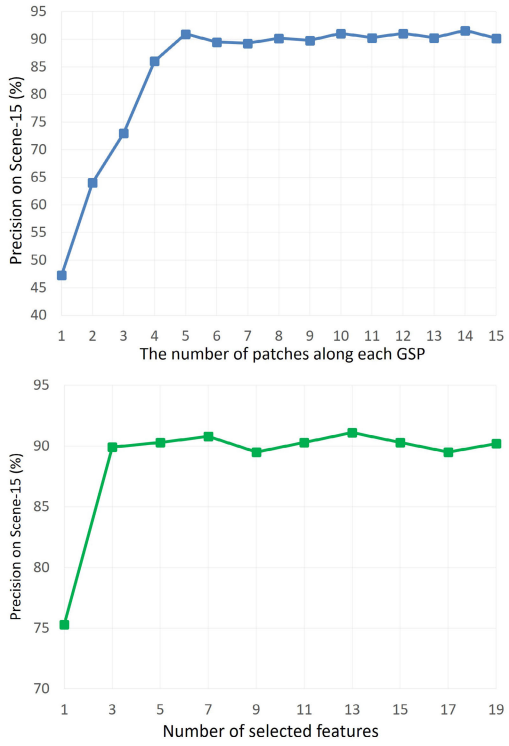

FIGURE 6. Categorization precision by adjusting M.

TABLE 6. Categorization precision by adjusting α.

α	Accuracy	α	Accuracy
0	73.53%	0.55	71.24%
0.05	80.23%	0.6	69.43%
0.1	81.76%	0.65	66.76%
0.15	85.44%	0.7	63.44%
0.2	87.79%	0.75	59.23%
0.25	90.21%	0.8	55.54%
0.3	86.43%	0.85	54.65%
0.35	85.89%	0.9	48.70%
0.4	85.12%	0.95	46.12%
0.45	84.43%	1	42.04%
0.5	78.12%		

TABLE 7. Categorization precision by adjusting β.

β	Accuracy	β	Accuracy
0	73.33%	0.55	78.43%
0.05	82.413%	0.6	76.12%
0.1	84.54%	0.65	75.06%
0.15	85.42%	0.7	73.21%
0.2	87.11%	0.75	73.02%
0.25	88.76%	0.8	72.12%
0.3	90.05%	0.85	71.32%
0.35	87.54%	0.9	70.12%
0.4	83.43%	0.95	66.76%
0.45	82.24%	1	65.32%
0.5	80.13%		

TABLE 8. Categorization precision by adjusting γ.

γ	Accuracy	γ	Accuracy
0	78.43%	0.55	77.54%
0.05	81.55%	0.6	76.65%
0.1	86.43%	0.65	74.22%
0.15	88.54%	0.7	73.76%
0.2	90.75%	0.75	72.76%
0.25	87.12%	0.8	72.11%
0.3	85.55%	0.85	71.65%
0.35	84.03%	0.9	72.65%
0.4	83.43%	0.95	71.24%
0.45	82.32%	1	69.21%
0.5	81.85%		

The subsequent phase of our analysis involves the fine-tuning of β and γ, with their respective influences on scene categorization efficacy recorded in Tables 7 and 8. Employing a similar investigative approach as with α, the optimal values for β and γ are determined to be 0.3 and 0.2, respectively. This precise calibration of parameters ensures that our model attains an optimal blend of accuracy and adaptability, rendering it highly effective for the intricate task of scene categorization.

D. GMM-BASED DANCING ACTION OPTIMIZATION

With the deep features identified for each Gaze Shift Path (GSP), we can accurately describe each scenic image by its human perceptual characteristics. Subsequently, we construct a probabilistic framework that captures the distribution of these deep GSP features, acquired during the training phase, to effectively retarget future dancing scenic images.

Interpreting different dancing scenic images is a subjective task, as individuals may have varied perceptions of the same scene. To address this diversity, our optimization scenario integrates insights from the visual perception of experienced photographers. A Gaussian Mixture Model (GMM) is utilized to represent the GSP features that have been refined during training, enabling a nuanced optimizing process based on human perceptual attributes.

$$prob(\mu|\Theta) = \sum\nolimits_i \theta_i * k_i(\nu|\beta_i, \Sigma_i), \tag{12}$$

In this model, θ_i signifies the relevance of the i-th component in the Gaussian Mixture Model (GMM); ν represents the feature associated with the Gaze Shift Path (GSP); while β_i and Σ_i are the mean and variance of the GMM, respectively.

The similarity between chosen GSP features is assessed using the Euclidean distance.

Based on this, give a testing dancing action picture, we use (12) to calculate its probability to the training well-aesthetic dancing actions. If the probability score is larger, we consider the testing dancing action is better. In our experiment, we test such optimization on 1342 dancing action pictures, wherein 1301 dancing actions are accurately optimized.

V. CONCLUDING REMARKS

The capability to accurately classify dancing scenes into specific categories is crucial for a wide range of applications in artificial intelligence (AI). In this study, we present an innovative approach known as Robust Deep Active Learning (RDAL), designed to craft a detailed image kernel that encapsulates human gaze dynamics effectively. Our methodology begins with a comprehensive set of scene images, applying a technique that accurately depict each scene's unique characteristics. Through the RDAL process, we pinpoint regions within the scenery that are both visually compelling and semantically rich, thereby establishing a GSP that underlies the scene's deep feature representation. These deep GSP features are subsequently selected using the MRFS algorithm, based on which a linear classifier is trained for the recognition of diverse sceneries. The efficacy of our approach, inspired by biological perception mechanisms, is validated by extensive testing, demonstrating its robustness and accuracy in dancing scene categorization tasks. Moreover, our learned GMM can well optimize dancing actions in practice.

However, a limitation arises from the potential discrepancy between the GSPs produced by our model and the gaze patterns observed in natural human vision. To address this, future efforts will involve conducting an exhaustive user study aimed at contrasting our generated GSPs with authentic human gaze trajectories. The goal is to refine the RDAL algorithm to more closely mimic the intricacies of the human visual system, thereby enhancing the fidelity of our dancing optimization outcomes. Moreover, we observe that the training time cost of our RDAL might be intolerable in practice. In the future, we plan to implement it on the Nvidia Compute Unified Devices Architecture (CUDA) platform, wherein the training can be noticeably accelerated.

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] 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.
[9] D. K. Park, Y. S. Jeon, and C. S. Won, "Efficient use of local edge histogram descriptor," in *Proc. ACM Multimedia Workshop*, 2000, pp. 51–54.
[10] 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. 44–48.
[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. NIPS*, 2012, pp. 1–14.
[12] Z. Xu, I. King, M. R. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Trans. Neural Netw.*, vol. 21, no. 7, pp. 1033–1047, Jul. 2010.
[13] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, 2006, pp. 117–132.
[14] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 413–420.
[15] U. von Luxburg, "A tutorial on spectral clustering," Max Planck Inst. Biol. Cybern., Germany, Tech. Rep. TR-149, 2006.
[16] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jul. 2005, pp. 524–531.
[17] 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.
[18] 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.
[19] 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.
[20] Z. Harchaoui and F. Bach, "Image classification with segmentation graph kernels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007.
[21] 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.
[22] 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.
[23] 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.
[24] 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.
[25] J. Yang, K. Yu, and T. Huang, "Supervised translation-invariant sparse coding," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] 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.

[31] 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.
[32] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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. 45–56.
[37] 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. 54–87.
[38] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep CNN features for scene classification," in *Proc. ICCV*, 2015, pp. 885–894.
[39] 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.
[40] 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.
[41] 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.
[42] 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. 65–76.
[43] 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.
[44] 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.
[45] 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.
[46] 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.
[47] 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.
[48] 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.
[49] 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.
[50] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, pp. 5–5, Mar. 2009.
[51] X. Guo, Y. Zhao, and J. Li, "DanceIt: Music-inspired dancing video synthesis," *IEEE Trans. Image Process.*, vol. 30, pp. 5559–5572, 2021.
[52] H. Matsuyama, S. Aoki, T. Yonezawa, K. Hiroi, K. Kaji, and N. Kawaguchi, "Deep learning for ballroom dance recognition: A temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing," *IEEE Sensors J.*, vol. 21, no. 22, pp. 25437–25448, Nov. 2021.
[53] X. Hu and N. Ahuja, "Unsupervised 3D pose estimation for hierarchical dance video recognition," in *Proc. ICCV*, 2021, pp. 10995–11004.
[54] D. Arpitha, M. Balasubrahmanyam, and D. A. Kumar, "Depth based Indian classical dance Mudra's recognition using support vector machine," in *Proc. 4th Int. Conf. Smart Syst. Inventive Technol. (ICSSIT)*, Jan. 2022, pp. 885–888.
[55] H. Bhuyan, P. P. Das, J. K. Dash, and J. Killi, "An automated method for identification of key frames in bharatanatyam dance videos," *IEEE Access*, vol. 9, pp. 72670–72680, 2021.
[56] 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.
[57] B. Cheng, B. Ni, S. Yan, and Q. Tian, "Learning to photograph," in *Proc. ACM MM*, 2010, pp. 89–95.
[58] 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.
[59] N. Li and S. Boers, "Human motion recognition in dance video images based on attitude estimation," *Wireless Commun. Mobile Comput.*, vol. 2023, pp. 1–11, May 2023.
[60] A. Bera, M. Nasipuri, O. Krejcar, and D. Bhattacharjee, "Fine-grained sports, yoga, and dance postures recognition: A benchmark analysis," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–13, 2023.
[61] H. Cheng, Y. Guo, L. Nie, Z. Cheng, and M. Kankanhalli, "Sample less, learn more: Efficient action recognition via frame feature restoration," in *Proc. ACM MM*, Oct. 2023, pp. 7101–7110.
[62] J. Yu, J. Pu, Y. Cheng, R. Feng, and Y. Shan, "Learning music-dance representations through explicit-implicit rhythm synchronization," *IEEE Trans. Multimedia*, vol. 26, pp. 8454–8463, 2024.
[63] Y. Pang and Y. Niu, "Dance video motion recognition based on computer vision and image processing," *Appl. Artif. Intell.*, vol. 37, May 2023, Art. no. 2226962.
[64] R. Jia, L. Zhao, R. Yang, H. Yang, X. Wu, Y. Zhang, P. Li, and Y. Su, "HFA-GTNet: Hierarchical fusion adaptive graph transformer network for dance action recognition," *J. Vis. Commun. Image Represent.*, vol. 98, Feb. 2024, Art. no. 104038.
[65] 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.
[66] D. Costea and M. Leordeanu, "Aerial image geolocalization from recognition and matching of roads and intersections," 2016, *arXiv:1605.08323*.
[67] 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.
[68] M. Stricker and M. Orengo, "Similarity of color images," in *Storage and Retrieval of Image and Video Databases*. Portland, OR, USA: IEEE Press, 1995.
[69] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jul. 2005, pp. 886–893.

LINGJUN XIANG is a Researcher with the Department of Physical Education and Research, Hunan Institute of Technology, Hengyang, China. His research interests include computer vision, multimedia, and image processing.

XIANG GAO is a Researcher with the Sports Training College, Guangzhou Sport University, Guangzhou, China. His research interests include AI, NLP, and image processing.