Sensors and Materials, Vol. 38, No. 5 (2026) 2415–2433
MYU Tokyo

S & M 4445

FALCON-HiMRAN: A Dual-stage RGB–D Sensor–based Scene Classification Framework with Cross-modal Fusion and Graph Reasoning

Nouf Abdullah Almujally,[1] Ting Wu,[2] Muhammad Waqas Ahmed,[3] Ahmad Jalal,[4,5*] and Hui Liu[6,7,8**]

[1]Department of Information System, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[2]Department of Otorhinolaryngology Head and Neck Surgery, Nanjing Tongren Hospital, School of Medicine, Southeast University, Nanjing 211102, China
[3]Department of Computer Games Development, Air University, E-9, Islamabad 44000, Pakistan
[4]Department of Computer Science, Air University, E-9, Islamabad, 44000, Pakistan
[5]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, South Korea
[6]Jiangsu Key Laboratory of Intelligent Medical Image Computing, School of Future Technology, Nanjing University of Information Science and Technology, Nanjing 210044, China
[7]Guodian Nanjing Automation Co., Ltd., Nanjing, China
[8]Cognitive Systems Lab, University of Bremen, 28359 Bremen, Germany

(Received September 29, 2025; accepted April 13, 2026)

Keywords: cross-modal sensor fusion, multimodal sensing, FALCON, sensor noise, sensor-driven perception

In this study, we address key limitations in RGB–depth (D) sensing systems, including depth noise, sensor misalignment, missing depth values, and performance degradation under low illumination. We propose a dual-stage RGB–D sensor-driven scene classification framework comprising feature-aligned lightweight cross-modal fusion (FALCON) and a hierarchical multi-region aggregation network (HiMRAN), the FALCON-HiMRAN, designed to enhance the reliability and interpretability of multimodal sensing systems. The proposed method integrates the data acquired from structured-light and time-of-flight RGB–D sensors and introduces the FALCON network to mitigate modality inconsistencies and sensor-induced noise. Furthermore, HiMRAN was developed to perform region-level reasoning by the graph-based modeling of spatial relationships. Experimental evaluation on benchmark RGB–D datasets demonstrates improved robustness under challenging sensing conditions such as occlusion, illumination variation, and depth degradation. The proposed framework contributes to the advancement of sensor-based perception systems by enabling more reliable scene understanding from imperfect multimodal sensor data. Remaining challenges include real-time deployment and the handling of extreme sensor noise in outdoor environments.

*Corresponding author: e-mail: ahmadjalal@mail.au.edu.pk
**Corresponding author: e-mail: hui.liu@uni-bremen.de
https://doi.org/10.18494/SAM5949

1. Introduction

Scene classification is a well-established problem in computer vision, where the aim is to assign semantic categories to all scenes on the basis of visual observations. It plays a critical role in applications such as robotic navigation, indoor mapping, autonomous inspection, and human–environment interaction.(1) Conventional approaches primarily rely on RGB imagery captured by vision sensors to extract contextual and spatial features for classification. These models, often based on convolutional or transformer-based architectures, have demonstrated strong performance under controlled conditions with stable lighting and clearly visible objects.(2)

Despite these advances, RGB-based models remain inherently limited in scenarios where visual cues are unreliable. In indoor environments, occlusion, low-light conditions, and texture ambiguities can significantly reduce the discriminative capability of color features. These limitations are directly linked to sensing constraints, where illumination dependence and the lack of geometric information result in degraded representations and reduced generalization performance in cluttered or poorly illuminated scenes. Addressing these challenges requires the integration of additional sensing modalities that provide complementary information beyond color and texture.

Depth imaging, typically performed using RGB–depth (D) sensors that capture both red, green, and blue channels and depth information, provides a geometric representation of the environment. It encodes structural properties such as object boundaries, relative distances, and spatial arrangements, which are largely invariant to lighting conditions and surface appearance. As a result, depth sensing plays a crucial role in enhancing perception in challenging environments. Consequently, RGB–D scene understanding has gained significant attention, with fusion strategies to combine color and depth information being explored in various studies. While early fusion (e.g., input-level concatenation) and late fusion (e.g., decision-level integration) have been widely investigated, their effectiveness is often limited owing to the heterogeneous nature of RGB and depth signals.(3)

A key challenge in RGB–D sensing systems lies in the spatial and semantic heterogeneity between modalities. RGB images contain high-frequency texture and color details, whereas depth images represent low-frequency geometric structures and are often affected by sensor noise, missing values, and measurement inaccuracies. These characteristics necessitate modality-specific feature extraction and adaptive fusion mechanisms that can dynamically adjust the contribution of each modality on the basis of local scene conditions. Moreover, conventional fusion strategies fail to account for region-wise variations in sensing reliability, such as regions where depth information is more reliable than RGB information owing to the effects of shadows or occlusion.(4)

In parallel, pixel-wise scene segmentation has emerged as an effective intermediate step in scene understanding pipelines, improving both interpretability and localization. Segmenting a scene into coherent regions enables object-centric reasoning, which is particularly beneficial for complex indoor environments. However, most segmentation approaches rely on supervised learning and require extensive annotated datasets, which are difficult to obtain for RGB–D sensing scenarios. Unsupervised segmentation offers a scalable alternative by leveraging the

intrinsic structure of multimodal sensor data, making it more suitable for real-world applications.[5]

The combination of unsupervised segmentation and adaptive RGB–D fusion presents a framework with strong potential to improve robustness in complex visual environments. In particular, jointly modeling appearance and geometry with fine-grained attention can support accurate scene classification even in the presence of occlusions or adverse lighting.

The key contributions of this paper are as follows.

- Feature-aligned lightweight cross-modal fusion (FALCON): A dual-stream fusion module that employs channel alignment and cross-modal dual attention to effectively integrate RGB and depth features derived from sensor data while suppressing modality-specific noise.
- Fusion-aware hierarchical segmentation (FAHS): An unsupervised two-stage segmentation pipeline using Felzenszwalb's algorithm followed by Markov random field (MRF) refinement, enabling accurate region proposal without the need for manual annotations.
- Hierarchical multi-region aggregation network (HiMRAN): A region-level reasoning module that constructs a region interaction graph (RIG) and uses graph neural networks (GNNs) and attention-based pooling to capture spatial and semantic relationships for final scene classification.

2. Related Work

Recent advancements in RGB–D segmentation and scene classification have led to the development of several models aimed at leveraging the complementary information from RGB and depth modalities for improved scene understanding. These models employ various strategies for feature extraction, fusion, and representation learning. Table 1 presents a summary of key RGB–D models, highlighting their main contributions and identifying their respective limitations.

3. Data, Materials, and Methods

In this paper, we propose a novel RGB–D fusion and scene understanding framework that integrates depth and color modalities to enhance scene classification accuracy. The pipeline begins with dual-stream DenseNet-121 encoders that process RGB and depth inputs in parallel to extract multiscale semantic features. These features are harmonized via a channel alignment module (CAM), ensuring a unified feature space for effective cross-modal fusion. The core of the architecture is the cross-modal dual attention fusion (CMDAF) module, which uses symmetric attention paths to selectively emphasize complementary features and suppress noise from both modalities. The resulting fused feature maps are enhanced and reconstructed into a high-quality RGB–D image representation. This image undergoes two-stage segmentation using Felzenszwalb's graph-based method followed by MRF refinement to produce coherent region proposals. These regions are then encoded by a shared CNN and enriched with geometric descriptors to form a set of position-aware region features. A RIG models the spatial and semantic relationships between regions. GNN layers perform message passing, allowing each region to refine its features in accordance with contextual neighbors. Attention-based

Table 1
Contributions and limitations of RGB–D models.

Model	Contribution	Limitation
AsymFormer[6]	The model introduces an asymmetrical backbone for the multimodal feature extraction and optimization of the computational resource distribution. It utilizes a local attention-guided feature selection (LAFS) module to selectively fuse features from different modalities. It achieves real-time semantic segmentation with high accuracy on mobile platforms.	The model faces challenges in highly dynamic scenes, especially where depth information is noisy, potentially reducing attention mechanism effectiveness.
PDCNet[7]	The model proposes a pixel difference convolutional network that captures intrinsic patterns by aggregating intensity and gradient information in local and global ranges for depth and RGB data, respectively.	The fixed grid kernel structure may limit fine-grained information capture, affecting pixel-level semantic segmentation accuracy.
DFormer[8]	DFormer is a novel RGB–D pretraining framework that learns joint RGB–D representations using paired ImageNet-1K data. It introduces specialized RGB–D blocks to better capture 3D geometry.	It requires substantial computational resources for pretraining; effectiveness in outdoor scenes remains to be validated.
ShapeConv[9]	The model introduces a novel convolutional layer that splits depth features into shape and base components, allowing richer semantic representation of geometric information. This layer can be easily added to standard convolutional neural network (CNN) architectures to enhance performance in RGB–D segmentation tasks.	Decomposition and weighting mechanisms require careful tuning, potentially complicating training.
TCANet[10]	TCANet introduces a three-stream coordinate attention network for RGB–D segmentation. It incorporates a multimodal fusion module to aggregate spatial and channel information, with an embedded ASPP module for multiscale semantics.	The multi-branch and attention-heavy architecture may increase training time and computational resource demands.
DBCAN[11]	Replaces real depth with pseudo-depth generated by estimation algorithms, reducing reliance on RGB–D sensors, and introduces PDAM for effective pseudo-depth fusion, combined with a diffusion model for feature extraction	Reliance on pseudo-depth may limit applicability in cases where accurate depth sensing is essential.
PDDM[12]	Replaces real depth with pseudo-depth generated by estimation algorithms, reducing reliance on RGB–D sensors, and introduces PDAM for effective pseudo-depth fusion, combined with a diffusion model for feature extraction	Reliance on pseudo-depth may limit applicability in cases where accurate depth sensing is essential.
MIPANet[13]	Proposes multiscale interaction and progressive attention to refine RGB and depth features progressively, capturing complementary cues for improved segmentation	Tuning the multi-scale attention mechanism increases training complexity and may complicate deployment.

Table 1
(Continued) Contributions and limitations of RGB–D models.

Model	Contribution	Limitation
EMSAFormer[14]	Efficient multitask RGB–D framework using a single Transformer encoder for panoptic segmentation, instance orientation, and scene classification	Performance may degrade in dynamic scenes or with noisy depth inputs.
SiaTrans[15]	SiaTrans is a Siamese transformer encoder for RGB–D salient object detection and depth quality classification. Cross-modality fusion (CMF) improves robustness under poor depth quality.	Primarily designed for saliency detection; adaptation for general scene classification may require modifications
ESeNet-D[16]	Lightweight real-time RGB–D segmentation for mobile robots, optimized with NVIDIA TensorRT, achieving high accuracy on NYU-Dv2 and SUN-RGB–D datesets	May be dataset-specific, requiring further validation for generalization
FRNet[17]	Enhances RGB–D semantic segmentation by refining features across multiple stages with residual connections and attention mechanisms	Multi-stage refinement increases complexity and training time.
PGDENet[18]	Progressive geometric detail enhancement for RGB–D segmentation, emphasizing geometric cues in depth data	Effectiveness may be reduced in cases where depth data lack fine geometric detail.
Omnivore[19]	General-purpose vision model trained across modalities including RGB–D, enabling strong performance on diverse tasks with a unified architecture.	May underperform compared with specialized models in RGB–D-specific segmentation tasks

hierarchical pooling aggregates region embeddings into a global scene descriptor, which is passed through a lightweight classifier to predict the scene category. The proposed HiMRAN framework is displayed in Fig. 1.

3.1 Role of RGB–D sensors and sensing challenges

The RGB–D data used in this study originate from commonly deployed sensing technologies such as structured-light sensors (e.g., Kinect v1) and time-of-flight (ToF) cameras. These sensors synchronously capture color and depth information but suffer from several well-known limitations including the following.

- Depth noise and missing values due to reflective or absorptive surfaces
- Misalignment between RGB and depth channels
- Reduced depth accuracy under strong ambient light
- Limited sensing range and resolution constraints

Our proposed FALCON module directly addresses these sensing challenges by performing channel alignment and adaptive cross-modal attention, which compensates for inconsistencies between sensor modalities. Furthermore, the HiMRAN module enhances robustness by incorporating region-level reasoning, allowing the system to operate effectively even when sensor data are partially degraded. Therefore, this work contributes to the application of sensing concepts by improving the reliability, interpretability, and robustness of RGB–D sensor–based perception systems in real-world environments.

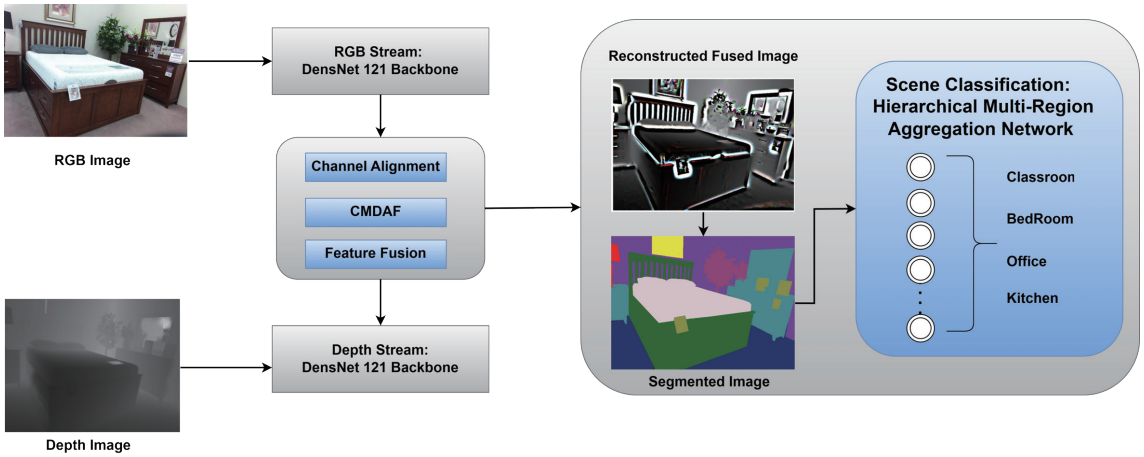

Fig. 1. (Color online) Proposed model for RGB–D scene classification, comprising feature fusion and hierarchical multi-region aggregation for accurate scene recognition.

3.2 FALCON

In RGB–D scene understanding, the effective integration of complementary information from RGB and depth modalities is crucial to improving recognition accuracy. Traditional fusion techniques such as early concatenation and late decision-level integration often suffer from limited interaction between modalities and suboptimal performance under modality degradation. To overcome these limitations, we propose a novel RGB–D image fusion network engineered to produce a fused image of superior quality by synergistically integrating information from RGB and depth modalities. The architecture is designed to be efficient, lightweight, and robust, rendering it highly suitable for downstream computer vision tasks such as semantic segmentation and scene classification. As illustrated in Fig. 2, the methodology is systematically structured into five principal stages: (1) parallel feature encoding using DenseNet-121 backbones, (2) CAM for feature space harmonization, (3) our core contribution, the CMDAF module, (4) a fused feature enhancement layer, and (5) a lightweight image reconstruction network.

3.2.1 Dual-stream feature extraction backbone

We employ DenseNet-121 as the backbone architecture for both RGB and depth modalities because of its superior feature reuse capabilities and its effectiveness in mitigating gradient vanishing problems. The dense connectivity pattern, where each layer receives feature maps from all preceding layers, ensures comprehensive multiscale feature extraction essential for cross-modal fusion tasks. Given an RGB image $I_{rgb} \in R^{H\times W\times 3}$ and a depth image $I_{depth} \in R^{H\times W\times 1}$, we employ two separate DenseNet-121 encoders.

$$F_{RGB} = E_{RGB}\left(I_{RGB};\theta_{RGB}\right) \tag{1}$$

$$F_{Depth} = E_{Depth}\left(I_{Depth};\theta_{Depth}\right) \tag{2}$$

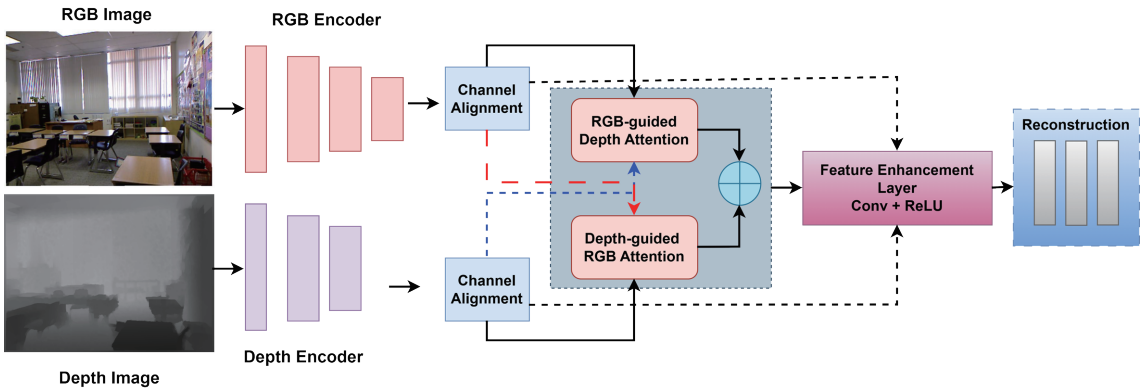

Fig. 2. (Color online) Proposed model of FALCON for fused image generation.

Here, E_{RGB} and E_{Depth} represent the DenseNet encoders with the parameters θ_{RGB} and θ_{Depth}, respectively. The output features encapsulate essential spatial and semantic information, serving as the input for subsequent modules. DenseNet encoders are employed for their ability to preserve gradient flow and extract dense features efficiently, which is crucial for capturing fine-grained image structures and establishing cross-modal correspondence. To ensure that spatial resolution and semantic hierarchy are maintained, each encoder is divided into dense blocks followed by transition layers, which reduce the feature map size while retaining important information. Feature extraction occurs at multiple stages, capturing low-level, mid-level, and high-level representations. This hierarchical design facilitates more effective fusion, particularly when integrated with attention mechanisms.

3.2.2 Multiscale feature aggregation

Multiscale feature aggregation is employed to capture representations from various receptive fields. Intermediate outputs from different DenseNet blocks are extracted and processed to construct a feature pyramid. These features encapsulate diverse semantic levels.

- Early layers capture fine textures and edges.
- Intermediate layers extract object parts and contours.
- Deeper layers encode object-level semantics and global context.

To merge these features, each intermediate output is up sampled or down sampled to a common resolution using bilinear interpolation. After normalization and dimensionality reduction via convolutional layers, the features are concatenated and passed through a fusion block for integrated representation learning.

$$F_{multi} = RELU\left(BN\left(Conv_{3\times3}\left(\left[F_1, F_2, \ldots F_n\right]\right)\right)\right) \tag{3}$$

Here, F_1, F_2, ..., F_n represent the up-sampled features from each DenseNet stage. This fused representation carries comprehensive contextual and structural information across scales.

3.2.3 CAM

RGB and depth modalities have inherently different statistical properties and feature distributions. To enable effective fusion, we introduce a CAM that transforms features into a common subspace.

$$F_{RGB} = BN\left(CONV_{1\times1}\left(F_{RGB};W_{RGB}\right)\right) \tag{4}$$

$$F_{Depth} = BN\left(CONV_{1\times1}\left(F_{Depth};W_{Depth}\right)\right) \tag{5}$$

The 1 × 1 convolutional layers act as learnable linear projectors, reducing and aligning the channel dimensions to D. Batch normalization improves generalization and stabilizes the learning process. This transformation ensures compatibility between modalities and prepares the features for cross-modal attention fusion.

3.2.4 CMDAF

Cross-modal attention is essential for highlighting complementary features while suppressing irrelevant or noisy information from each stream. Some of the fused images generated by the model are shown in Fig. 3. The CMDAF module introduces a symmetric attention mechanism, allowing bidirectional enhancement. Inspired by self-attention, the mechanism computes attention weights using queries, keys, and values, but extends this to intermodal interactions. This design allows each modality to serve both as a guide and as a receiver of enhanced representations. In this attention path, RGB features guide depth feature enhancement.

(a) (b) (c)

Fig. 3. (Color online) FALCON fusion results. (a) Input RGB images, (b) input depth maps, and (c) fused outputs from the proposed FALCON.

The RGB stream generates queries, while the depth stream provides keys and values.

$$Q_{RGB} = F_{RGB}.W_Q^{RGB} \tag{6}$$

$$K_{Depth} = F_{Depth}.W_K^{Depth} \tag{7}$$

$$V_{depth} = D_{Depth}.W_v^{Depth} \tag{8}$$

3.3 FAHS

To segment semantically meaningful regions from the fused RGB–D feature representation, we introduce a two-stage segmentation strategy FAHS. FAHS begins with Felzenszwalb's graph-based segmentation to generate initial over-segmented regions on the basis of the local image structure. This is followed by MRF refinement, which enforces spatial consistency and enhances boundary precision. The segmentation operates on the fused feature map produced by the CMDAF module, leveraging rich contextual information across modalities for accurate region delineation.

3.3.1 Felzenszwalb's graph-based segmentation

To perform segmentation, we apply Felzenszwalb's graph-based algorithm, which is well suited to processing complex scenes. This method treats the image as a graph: each pixel is represented as a node, and edges connect neighboring pixels on the basis of both color similarity and spatial closeness. The relationship between the connected nodes g and h is defined by

$$F(g,h) = \frac{\left|U_g - U_h\right|}{\max\left\{V(g), V(h)\right\}}, \tag{9}$$

where $U(g)$ and $U(h)$ represent the pixel intensity values, while $V(g)$ and $V(h)$ reflect the spatial distances between them. This formulation helps determine whether two adjacent regions should be merged on the basis of local contrast and texture. A crucial parameter in this algorithm is the scale, which affects the granularity of segmentation. A higher scale value produces broader, more generalized segments, while a lower scale value yields finer, more detailed regions. From the empirical results, a scale of 100 was found to provide an effective balance between over- and under-segmentation.

3.3.2 MRF refinement

To further enhance the segmentation output, we incorporate an MRF model as a post-processing step. While the initial segmentation may contain small errors or fragmented

boundaries, MRF helps enforce spatial consistency by modeling neighborhood relationships and encouraging smooth label transitions. The energy function minimized by MRF combines a data term and a smoothness term.

$$E(L)=\sum_{g}D\left(g,L_g\right)+\sum_{(g,h)\in\aleph}S\left(L_g,L_h\right) \tag{10}$$

Here, $D\left(g,L_g\right)$ reflects how well the assigned label matches the observed data, while $S\left(L_g,L_h\right)$ penalizes abrupt label changes between the neighboring pixels g and h. By minimizing this energy, MRF refines the segmentation map, corrects noise around boundaries, and produces more coherent object regions. Together, Felzenszwalb's algorithm and MRF post-processing form a robust segmentation framework that lays a solid foundation for accurate object detection and scene interpretation in our depth image analysis pipeline. The FAHS module generates semantically meaningful regions, as shown in Fig. 4, which serve as the foundation for region-level feature extraction and graph-based reasoning.

3.4 HiMRAN for scene classification

To achieve high-accuracy scene classification after segmentation, we propose HiMRAN, a novel and modular deep learning architecture designed to understand scenes by explicitly modeling the spatial, semantic, and compositional relationships among different image regions. Unlike conventional approaches that process an image holistically, HiMRAN decomposes the input into a hierarchy of semantically meaningful regions and models their interactions. This hierarchical strategy allows the network to learn how individual parts come together to form a coherent scene, which improves robustness against clutter, occlusion, and complex object arrangements.

The HiMRAN framework operates in two primary stages.

- Region-level Encoding: This is focused on learning semantic and geometric features of individual segmented regions through convolutional and positional embeddings.

Fig. 4. (Color online) Representative segmentation results obtained using the proposed FAHS framework.

- Scene-level Aggregation: This integrates region-wise features via graph-based message passing and attention-based pooling to produce a context-aware, global scene representation.

By structuring the model in this hierarchical and interpretable manner, HiMRAN achieves improved scene classification performance in complex and unstructured real-world environments.

3.4.1 Region-level feature encoding

The input segmented images are then broken into non-overlapping regions, each representing a semantically meaningful part of the scene. Each region is passed through a shared CNN encoder with three convolutional layers and residual connections to maintain gradient flow during training. To extract visual characteristics from each region, we employ a shared CNN encoder. This encoder consists of three convolutional layers with residual connections, which are crucial for maintaining stable gradient flow during training. The shared-weight architecture ensures parameter efficiency and promotes the learning of a consistent feature space across all regions. Each region, r_i, is passed through this encoder to produce a semantic feature vector f_i:

$$f_i = CNN_{region}(r_i), f_i \in \mathbb{R}^d, \tag{11}$$

where d is the dimensionality of the semantic feature space. This vector f_i encapsulates the visual properties of the region, such as texture, color, and object-specific attributes.

3.4.2. Position-aware feature enhancement

The segmentation process, while effective for isolating objects, often discards crucial spatial layout information. To compensate for this, we enrich each region's semantic feature vector with a geometric encoding vector, p_i. This vector provides the model with explicit information about the region's spatial context and morphology, which is vital for understanding the scene structure. The vector p_i is composed of the following attributes.

- Normalized Centroid Coordinates (x_c, y_c): the position of the region's center relative to the image dimensions
- Normalized Bounding Box Dimensions (w, h): the width and height of the region's bounding box, normalized by the image dimensions
- Region-to-image-area Ratio: the proportion of the total image area occupied by the region, indicating its scale
- Shape Descriptors: metrics such as aspect ratio, elongation, solidity, and compactness, which capture the geometric form of the region

The semantic feature vector f_i and the geometric encoding vector p_i are then concatenated to create a comprehensive, enriched region representation F_i.

$$F_i = [f_i; p_i] \tag{12}$$

This combined representation F_i serves as the input to the next stage, ensuring that both the what (semantic content) and the where/how (spatial context) of each region are considered.

3.4.3 RIG construction

To model the intricate relationships between regions, we construct a RIG. In this graph, each node corresponds to one of the enriched region feature vectors, F_i. The edges of the graph are weighted to represent the affinity between pairs of regions. The weight of the edge between any two regions, i and j, is determined by a dual-affinity function considering both spatial proximity and semantic similarity:

$$W_{ij} = \exp\left(-\alpha \left\| P_i - P_j \right\|^2 - \beta \left\| f_i - f_j \right\|^2 \right), \tag{13}$$

where $\left\| P_i - P_j \right\|^2$ is the squared Euclidean distance between geometric encodings and $\left\| f_i - f_j \right\|^2$ is the squared Euclidean distance between semantic features. α and β are learnable hyperparameters controlling sensitivity to spatial and semantic distances. This formulation ensures that a strong edge W_{ij} is formed only when two regions are both spatially close and semantically similar.

3.4.4 Graph-based message passing for contextual refinement

With the RIG established, we apply a GNN to allow regions to exchange contextual information. This message-passing mechanism refines each region's feature representation by incorporating information from its neighbors. For each node i at layer l, the updated feature h_i^l is computed as

$$h_i^l = \sigma\left(W_1 h_i^{(l-1)} + \sum_{j \in \mathcal{N}^i} \alpha_{ij} W_2 h_j^{(l-1)} \right). \tag{14}$$

Here, $h_i^{(l-1)}$ and $h_j^{(l-1)}$ are feature representations from the previous layer, N_i denotes neighbors of node i, W_1 and W_2 are trainable matrices, α_{ij} is an attention score controlling the effect of neighbor j, and σ is a nonlinear activation function like ReLU. Residual connections are incorporated to improve learning stability.

3.4.5 Hierarchical region-to-scene aggregation

After L layers of GNN message passing, each region possesses a context-enhanced embedded h_i^l. We then aggregate these embeddings by attention-based pooling.

$$\alpha = \frac{\exp\left(W_a^T h_i^L + b_a \right)}{\sum_j \exp\left(W_a^T h_i^L + b_a \right)} \tag{15}$$

$$Z = \sum_i \alpha_i h_i^L \tag{16}$$

Here, W_a and b_a are learnable parameters projecting features to an attention space. The attention scores α_i determine the importance of each region in the final scene descriptor z. As illustrated in Fig. 5, the proposed FALCON-HiMRAN framework successfully classifies complex indoor scenes by leveraging hierarchical region-based representations. The results demonstrate robustness under challenging conditions such as clutter, occlusion, and variations in illumination.

3.4.6 Scene classification head

The global scene descriptor z is passed through a fully connected layer followed by a dropout mechanism to reduce overfitting, then a Softmax function is applied for classification:

$$y = \text{Softmax}\left(W_{cz} + b_c\right), \tag{17}$$

where W_c and b_c are the weight and bias of the classification layer. The output $y \in R^K$ represents probabilities across K scene categories. The model architecture is shown in Fig. 6.

4. Experimentation and Results

The experimental findings highlight the strong performance of the proposed framework in RGB–D scene classification. By integrating FALCON with HiMRAN, the system demonstrates

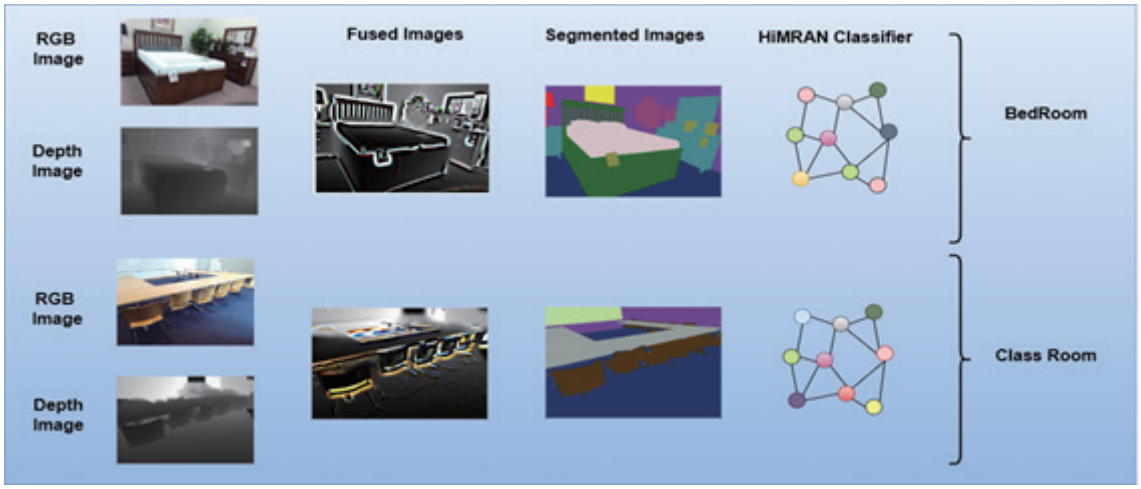

Fig. 5. (Color online) Sample scene classification results produced by the proposed FALCON-HiMRAN framework. The images illustrate correctly classified indoor scenes, demonstrating the effectiveness of hierarchical region-based reasoning under varying conditions such as clutter, occlusion, and illumination changes.

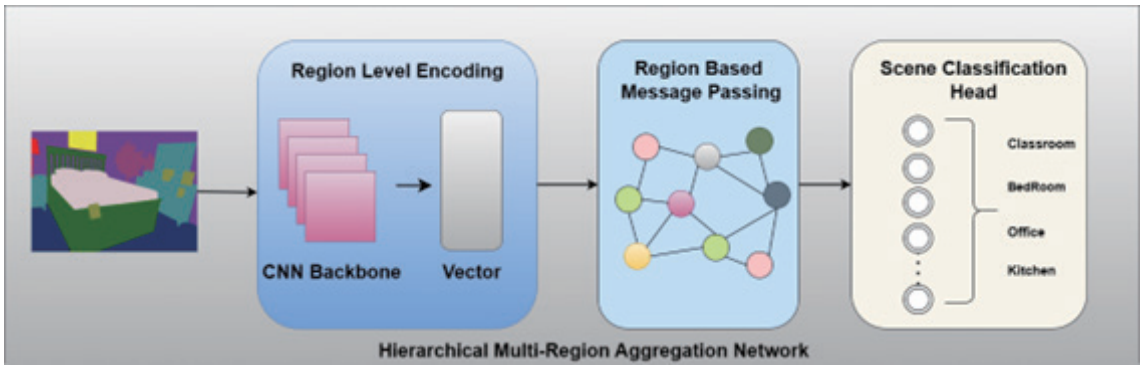

Fig. 6. (Color online) Architecture of HiMRAN for scene classification.

a notable ability to handle issues commonly arising in RGB–D data, including sensor misalignment, depth noise, and imprecise object boundaries. One of the most compelling aspects of the approach is its high accuracy in distinguishing between scenes with clearly defined structural layouts and object distributions.

In particular, the model consistently achieved strong results in categorizing indoor environments such as offices, bedrooms, and living rooms. This high performance can be attributed to the model's capacity to capture detailed local features alongside broader spatial and semantic patterns. The dual-stage architecture enables the precise recognition of object contours while also modeling the contextual relationships across different regions of a scene, contributing to more reliable classification even in the presence of clutter or occlusion.

4.1 Dataset descriptions

The NYU Depth v2 dataset[20] was utilized for training and evaluating the proposed model. It comprises both labelled and unlabelled indoor images captured across various real-world environments such as bathrooms, bedrooms, bookstores, cafes, kitchens, living rooms, and offices. These indoor scenes contain a wide assortment of objects, including but not limited to beds, sofas, bookshelves, televisions, cabinets, windows, and walls. Additionally, we employed the SUN-RGB–D dataset[21] that similarly features a broad spectrum of indoor environments. For consistency and comparative analysis, ten common scene categories were selected from the two datasets for experimentation.

4.2 Results

Our proposed model delivers strong performance on both the NYU-Dv2 and SUN-RGB–D datasets across scene classification and object segmentation tasks. For scene classification, it consistently achieves a robust balance between precision and recall, recording mean $F1$-scores of 0.855 on the NYU-Dv2 dateset and 0.899 on the SUN-RGB–D dateset. The $F1$-score is the harmonic mean of precision and recall, and it provides a balanced measure of classification performance, particularly in cases of class imbalance, and is defined as

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{18}$$

These results reflect the model's effectiveness in learning discriminative spatial and semantic features from RGB–D data. Scene categories such as bedrooms, kitchens, bathrooms, closets, and bookstores are identified with notably high accuracy, often achieving near-perfect precision alongside strong recall, as displayed in Figs. 7 and 8. This suggests that the model successfully captures the distinctive spatial configurations and depth patterns common to these environments. On the other hand, performance slightly declines in scenes that feature visual overlap or high levels of clutter, such as living rooms, dining areas, and office spaces, particularly within the NYU-Dv2 dataset. In such cases, ambiguity in object placement and similar textures may reduce classification consistency, with either precision or recall being affected.

Interestingly, the SUN-RGB–D dataset yields more stable performance across all scene types, likely due to more structured layouts or clearer scene boundaries, which better align with the model's hierarchical reasoning capabilities. Table 2 presents the intersection over union (IoU) scores for individual object classes across both datasets, highlighting the segmentation accuracy of our proposed approach. The IoU metric quantitatively evaluates the overlap between the predicted segmentation mask and the ground truth. The model achieves a competitive mean IoU across a wide range of object categories, demonstrating its robustness in differentiating objects with similar visual features and in scenes with cluttered or complex backgrounds. The improved segmentation quality reflects the model's ability to preserve object boundaries and maintain spatial coherence. Nonetheless, slight performance variability across certain object categories indicates potential room for improvement, particularly in dealing with homogeneous or low-texture regions where boundary cues are less pronounced.

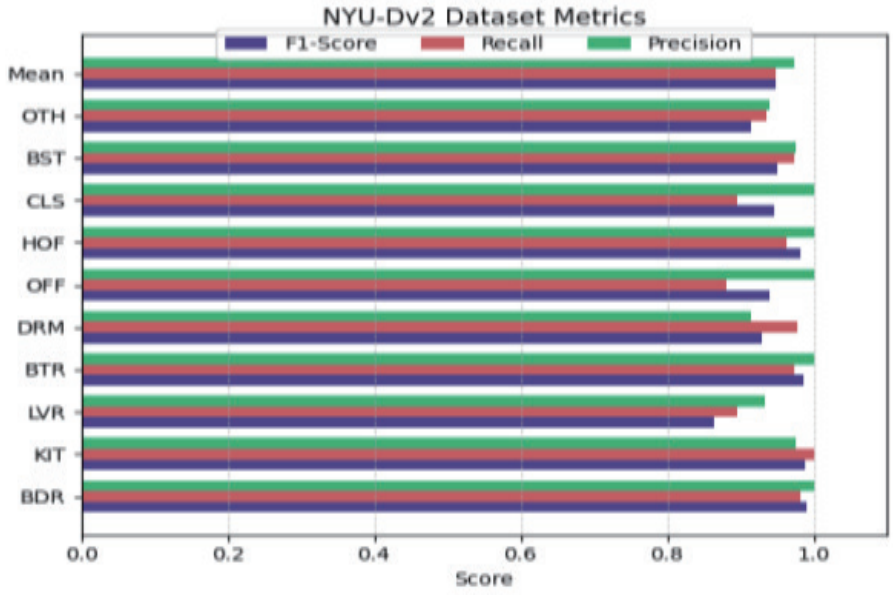

Fig. 7. (Color online) Recall, precision, and $F1$-score for each class on NYU-Dv2 dataset. Note: BTR bathroom; BDR bedroom; BST bookstore; KIT kitchen; LVR living room; OFF office; HOF home office; CLS classroom; DRM drawing room; OTH others.

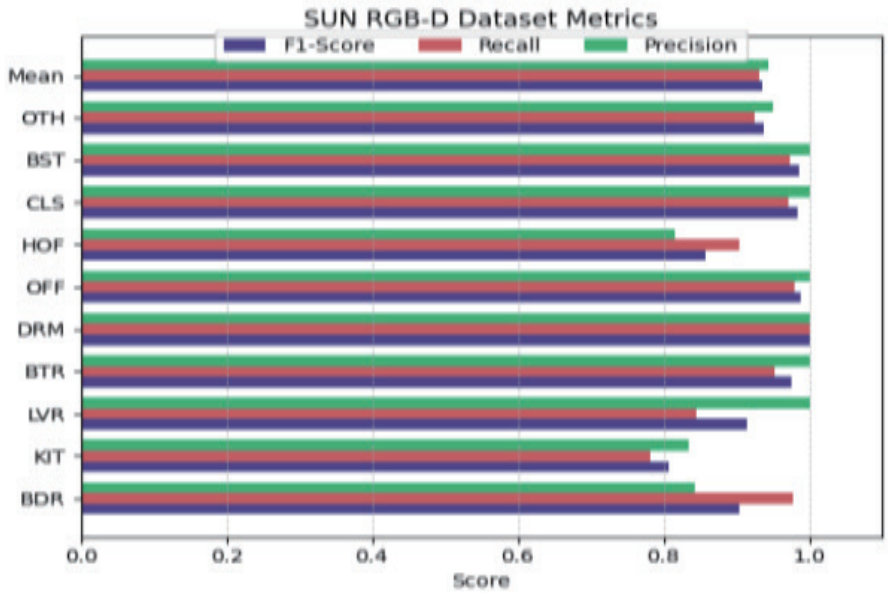

Fig. 8. (Color online) Recall, precision, and $F1$-score for each class on SUN-RGB–D dataset. Note: BTR bathroom; BDR bedroom; BST bookstore; KIT kitchen; LVR living room; OFF office; HOF home office; CLS classroom; DRM drawing room; OTH others.

Table 2
IoU scores for object segmentation accuracy on the two datasets.

Objects	IoU score over NYU-Dv2 dataset	IoU score over SUN-RGB–D dataset
Door	0.917	0.921
Window	0.884	0.871
Floor	0.931	0.937
Wall	0.942	0.940
Ceiling	0.844	0.859
Table	0.892	0.897
Chair	0.842	0.837
Cup	0.824	0.828
Book	0.907	0.918
Cabin	0.899	0.901
Board	0.961	0.960
Mirror	0.879	0.877
Dice	0.921	0.925

4.2.1 Ablation study

The comprehensive ablation study presented in Table 3 systematically evaluates the contribution of each component in our proposed RGB–D scene classification framework. The analysis begins with baseline configurations using single modalities, where RGB-only achieves 72.3% on the NYU-Dv2 dateset and 68.7% on the SUN-RGB–D dateset, while depth-only performance drops to 65.8 and 62.1%, respectively, highlighting the complementary nature of the two modalities. The simple concatenation of RGB and depth features provides modest

Table 3
Performances of different experimental configurations on NYU-Dv2 and SUN-RGB–D datasets.

Experiment	DensNet Backbone	CAM	CMDAF	FAHS	RIG	GNN	NYU-Dv2 (%)	SUN-RGB–D (%)
Baseline RGB-only	✓	×	×	×	×	×	72.3	68.7
Baseline Depth-only	✓	×	×	×	×	×	65.8	62.1
Simple Concatenation	✓	×	×	×	×	×	76.4	71.2
With CAM	✓	✓	×	×	×	×	79.1	74.8
CAM + CMDAF	✓	✓	✓	×	×	×	83.7	77.5
FALCON Complete	✓	✓	✓	×	×	×	85.2	78.9
FALCON + Basic Seg.	✓	✓	✓	×	×	×	86.8	79.6
FALCON + FAHS	✓	✓	✓	✓	×	×	87.9	80.3
Without RIG	✓	✓	✓	✓	×	×	88.4	80.7
Without GNN	✓	✓	✓	✓	✓	×	89.1	81.2
Without Attention Pooling	✓	✓	✓	✓	✓	✓	89.6	81.8
Proposed system	✓	✓	✓	✓	✓	✓	**90.34**	**82.46**

improvements (76.4 and 71.2%), demonstrating the need for more sophisticated fusion strategies. The introduction of the CAM yields significant gains of 2.7 and 3.6% on the two datasets, respectively, validating its effectiveness in harmonizing feature distributions across modalities. The CMDAF module contributes the most substantial improvement, adding 4.6 and 2.7% performance gains, respectively, which underscores the importance of bidirectional attention mechanisms for effective cross-modal feature enhancement. The complete FALCON network achieves 85.2 and 78.9% accuracies, respectively, representing a solid foundation for scene understanding. Further incorporation of FAHS yields 2.7 and 1.4% improvements, respectively, demonstrating the value of region-based analysis in complex scene interpretation. The RIG and GNN components collectively contribute additional 1.2 and 0.9% performance boosts, respectively, highlighting the effectiveness of graph-based spatial reasoning. Finally, the complete proposed system integrating both FALCON and HiMRAN architectures achieves state-of-the-art performance values of 90.34% on the NYU-Dv2 dateset and 82.46% on the SUN-RGB–D dateset, with each component providing meaningful and cumulative improvements to the overall framework's discriminative capability.

The results in Table 4 clearly demonstrate the superior performance of our proposed model over existing state-of-the-art methods. Achieving accuracies of 90.34% on the NYU-Dv2 dataset and 82.46% on the SUN-RGB–D dataset, our model significantly outperforms previous approaches, with a notable margin of 2–5% improvement over the most recent methods. This performance gain highlights the effectiveness of our model's dual-modality fusion and hierarchical attention mechanisms in capturing both visual semantics and depth context, making it highly robust for complex indoor scene classification tasks. Although evaluated on benchmark

Table 4
Scene classification accuracies (%) of proposed method and existing approaches.

Method	SUN-RGB–D (%)	NYU-Dv2 (%)
OOR-CNN (2017)[22]	—	66.9
RCNN (2018) [23]	53.8	67.5
MSN (2020)[24]	56.2	68.1
TRecgNet (2019)[25]	56.7	69.2
DWT and DCT (2022)[26]	48.7	72.8
LM-CNN (2019)[27]	63.1	79.3
EMSANet (2022)[28]	61.8	76.5
GS2F2App (2024)[29]	62.3	77.8
SHMFs (2024)[30]	63.7	80.1
Mape-ViT (2025)[31]	78.56	88.79
Proposed method	**82.46**	**90.34**

datasets, the proposed framework is designed for real-world deployment with RGB–D sensing devices such as Kinect and Intel RealSense cameras. The lightweight design of FALCON and the modular structure of HiMRAN enable integration into embedded vision systems. This demonstrates that the proposed method is not merely theoretical but has practical applicability in real-time sensing environments such as robotics and smart surveillance systems.

5. Conclusions

In this study, we contributed to the advancement of RGB–D sensing systems by developing a framework that explicitly addresses sensor-level challenges such as noise, misalignment, and incomplete depth information. Our proposed architecture demonstrated how intelligent cross-modal fusion and region-based reasoning can enhance the effectiveness of existing RGB–D sensors without requiring hardware modification. By improving the reliability and interpretability of multimodal sensor data, the framework provides a practical pathway for real-world deployment in applications such as robotics, surveillance, and autonomous systems.

Acknowledgments

The publication was supported by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. This research was also supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

1 A. Al-Qerem, F. Kharbat, S. Nashwan, S. Ashraf, and K. Blaou: Int. J. Distrib. Sens. Netw. **16** (2020) 170121. https://doi.org/10.1177/1550147720911009
2 M. W. Ahmed and A. Jalal: Proc. 2024 Int. Conf. Agents and Artificial Intelligence and Cognitive Systems (IEEE, 2024). https://doi.org/10.1109/ICACS60934.2024.10473231
3 L. Calavia, M. Sainz, J. C. Villadangos, A. I. Torre, J. T. Astrain, and J. M. Echeverria: Sensors **12** (2012) 10407. https://doi.org/10.3390/s120810407

4 N. Al Mudawi, M. Tayyab, M. W. Ahmed, and A. Jalal: Proc. 2024 Int. Conf. Autonomous Robot Systems and Competitions (IEEE, 2024). https://doi.org/10.1109/ICARSC61747.2024.10535954
5 R. S. Brahmana, F. Mohammed, and K. Chairuang: Lontar Komputer **11** (2020) 32. https://doi.org/10.24843/lkjiti.2020.v11.i01.p04
6 S. Du, W. Wang, R. Guo, R. Wang, and S. Tang: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (IEEE, 2024) 7608–7615.
7 A. Bilal, A. H. Khan, K. Almohammadi, S. A. Al Ghamdi, H. Long, and H. Malik: IEEE Access **12** (2024) 150147. https://doi.org/10.48550/arXiv.2309.14065
8 B. Yin, X. Zhang, Z. Li, L. Liu, M. M. Cheng, and Q. Hou: arXiv (2023) arXiv:2309.09668. https://doi.org/10.48550/arXiv.2309.09668
9 J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li: Proc. IEEE/CVF Int. Conf. Comput. Vis. (IEEE, 2021) 7088–7097.
10 W. Jia, X. Yan, Q. Liu, T. Zhang, and X. Dong: Complex Intell. Syst. **10** (2024) 1219. https://doi.org/10.1007/s40747-023-01245-5
11 A. Wu and L. Fu: Appl. Sci. **14** (2024) 8329. https://doi.org/10.3390/app14188329
12 C. Chen, Y. Yang, S. Huang, H. Lu, W. Wan, S. Wei, W. Wen, and S. Wang: IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **17** (2024) 18882. https://doi.org/10.1109/JSTARS.2024.3477593
13 S. Zhang and M. Xie: Front. Phys. **12** (2024) 1411559. https://doi.org/10.3389/fphy.2024.1411559
14 S. B. Fischedick, D. Seichter, R. Schmidt, L. Rabes, and H. M. Gross: Proc. Int. Joint Conf. Neural Netw. (IEEE, 2023) 1–10. https://doi.org/10.1109/IJCNN54540.2023.10191977
15 X. Jia, C. DongYe, and Y. Peng: Image Vis. Comput. **127** (2022) 104549. https://doi.org/10.1016/j.imavis.2022.104549
16 T. R. Erep, L. Chaari, P. Ele, and E. Sobngwi: Proc. IEEE Int. Workshop Mach. Learn. Signal Process. (IEEE, 2024) 1–6. https://doi.org/10.1109/MLSP58920.2024.10734761
17 W. Zhou, E. Yang, J. Lei, and L. Yu: IEEE J. Sel. Top. Signal Process. **16** (2022) 677.
18 W. Zhou, E. Yang, J. Lei, J. Wan, and L. Yu: IEEE Trans. Multimedia **25** (2022) 3483.
19 R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (IEEE, 2022) 16081–16091. https://doi.org/10.1109/CVPR52688.2022.01563
20 N. Silberman, D. Hoiem, P. Kohli, and R. Fergus: Proc. Eur. Conf. Comput. Vis. (Springer, 2012) 746–760. https://doi.org/10.1007/978-3-642-33715-4_54
21 S. Song, S. P. Lichtenberg, and J. Xiao: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (IEEE, 2015) 567. https://doi.org/10.1109/CVPR.2015.7298655
22 X. Song, C. Chen, and S. Jiang: Proc. ACM Int. Conf. Multimedia (ACM, 2017) 600.
23 X. Song, S. Jiang, L. Herranz, and C. Chen: IEEE Trans. Image Process. **28** (2018) 980.
24 Z. Xiong, Y. Yuan, and Q. Wang: Neurocomputing **373** (2020) 81.
25 D. Du, L. Wang, H. Wang, K. Zhao, and G. Wu: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (IEEE, 2019) 11836–11845.
26 A. A. Rafique, Y. Y. Ghadi, S. A. Alsuhibany, S. A. Chelloug, A. Jalal, and J. Park: Proc. Conf. Membrane Comput. (2022) 27–29.
27 Z. Cai and L. Shao: Cogn. Comput. **11** (2019) 825.
28 D. Seichter, S. B. Fischedick, M. Köhler, and H. M. Gross: Proc. Int. Joint Conf. Neural Netw. (IEEE, 2022) 1–10.
29 R. Pereira, T. Barros, L. Garrote, A. Lopes, and U. J. Nunes: Pattern Recognit. Lett. **179** (2024) 24.
30 R. Pereira, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes: arXiv (2024) arXiv:2404.07739. https://doi.org/10.48550/arXiv.2404.07739
31 M. W. Ahmed, T. Sadiq, H. Rahman, S. A. Alateyah, M. Alnusayri, M. Alatiyyah, and D. A. AlHammadi: PeerJ Comput. Sci. **11** (2025) e2796. https://doi.org/10.7717/peerj-cs.2796