Sensors and Materials, Vol. 38, No. 11 (2026) 6621–6634
MYU Tokyo

S & M 4698

RegionFormer: AI-driven RGB-D Sensor Signal Processing and Multimodal Fusion for Intelligent Indoor Scene Classification

Naif Al Mudawi,[1,2] Manar Mousa Altamimi,[3] Aysha Naseer,[4] Yahya AlQahtani,[5,6] Bayan Alabdullah,[3] Khaled Alnowaiser,[7] Fatimah Alhayan,[3] and Ahmad Jalal[4,8*]

[1]Department of Computer Science, College of Computer Science and Information System, Najran University, Najran, Saudi Arabia
[2]Science and Engineering Research Center, Najran University, Najran, Saudi Arabia
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[4]Department of Computer Science, Air University, Islamabad 44000, Pakistan
[5]Department of Informatics and Computer Systems, King Khalid University, Abha 61421, Saudi Arabia
[6]Center for Artificial Intelligence (CAI), King Khalid University, Abha, Saudi Arabia
[7]Department of Computer Science, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia
[8]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, South Korea

(Received September 1, 2026; accepted October 21, 2026)

Keywords: multimodal sensor fusion, sensor signal processing, region-aware attention, transformer, deep learning, intelligent automation, indoor scene understanding

Reliable indoor scene understanding requires the effective integration of complementary sensor signals. Red–green–blue–depth (RGB-D) sensing provides appearance and geometric information; however, global feature aggregation and fixed multimodal fusion can limit adaptation to varying sensor quality. In this paper, we present RegionFormer, a region-aware framework for RGB-D sensor signal processing and indoor scene classification. A shared-weight dual-stream SegFormer-B3 encoder extracts RGB and depth features, while a semantic region generator constructs 16 spatially aligned region tokens. Adaptive region feature fusion (ARFF) dynamically weights RGB, depth, and joint RGB-D features, a region-aware deep attention block enhances contextual representations, and a unified modality-adaptive transformer produces the final scene representation. RegionFormer reports 83.38% accuracy, 78.36% mean class accuracy, and 77.48% Macro F1-score on NYUD-v2, and 65.00% accuracy, 62.00% mean class accuracy, and 61.50% Macro F1-score on SUN RGB-D, with a computational cost of 11.99 giga floating-point operations per input pair on NYUD-v2. Under complete depth-sensor failure, ARFF achieves 78.60% accuracy compared with 67.83% for fixed feature concatenation, corresponding to performance reductions of 4.78 and 7.83 percentage points, respectively. Ablation studies further indicate the contribution of each architectural module. These findings highlight the potential of region-aware representation learning and adaptive

*Corresponding author: e-mail: ahmadjalal@mail.au.edu.pk
https://doi.org/10.18494/SAM6614

ISSN 0914-4935 © MYU K.K.

multimodal fusion for robust indoor scene understanding in robotics, navigation, and intelligent automation.

1. Introduction

The reliable perception of indoor environments is required in service robotics, autonomous navigation, smart monitoring, and human–machine interaction. Red–green–blue–depth (RGB-D) sensing is well suited to these applications because RGB signals describe appearance and texture, whereas depth signals describe geometry and structure; their combination gives a more complete description of a scene than either modality alone.[1,2] Indoor scene classification nevertheless remains difficult because different categories can contain similar objects and layouts, and scenes of the same category vary in viewpoint, illumination, and arrangement.[2,3] In addition, consumer RGB-D sensors such as Microsoft Kinect provide depth measurements whose accuracy decreases with increasing distance and which are missing on reflective, transparent, and dark surfaces;[4,5] a fusion scheme for such sensors must therefore be able to reduce its reliance on the depth signal where the measurements are unreliable.

Early RGB-D recognition methods relied on handcrafted descriptors, whereas convolutional neural networks (CNNs) learn hierarchical features directly from RGB and depth inputs.[2] Convolutional features, however, are aggregated locally, and the simple concatenation or addition of RGB and depth features does not fully exploit the complementary characteristics of the two modalities.[2,3,6,7] Transformer architectures model dependencies between distant features through self-attention,[8] and SegFormer produces multi-scale representations that integrate local and global information.[9] Global features, however, do not represent the semantically meaningful regions of a scene, and fixed fusion ignores that the relative importance of RGB and depth cues differs between regions.

To address these limitations, in this paper, we present RegionFormer, a region-aware framework for RGB-D sensor signal processing and indoor scene classification. After the bilateral filtering of the depth signal,[10] a shared-weight dual-stream SegFormer-B3 encoder extracts RGB and depth features, and a semantic region generator (SRG) maps them into compact region tokens. The adaptive region feature fusion (ARFF) module combines the RGB, depth, and joint RGB-D tokens of each region according to their learned relevance, the region-aware deep attention block (RADAB) models inter-region relationships, and the unified modality-adaptive transformer (UMAT) produces the scene-level representation. The main contributions of this paper are as follows:

(i) An SRG is introduced to convert the dense RGB-D feature maps into a small number of region tokens through a lightweight soft assignment that is computed once from the joint RGB-D features and shared by the RGB and depth streams, so that the region tokens of the different modalities are spatially aligned for region-level fusion.

(ii) An ARFF module is introduced to dynamically integrate the RGB, depth, and joint RGB-D representations at the semantic-region level, so that the reliance on the depth signal decreases when the depth measurements are degraded.

(iii) A RADAB is introduced to aggregate scene context across the semantic regions and to refine the fused region tokens through inter-region attention and feed-forward transformation.
(iv) A systematic ablation study and an analysis of the fusion weights under natural and controlled sensor degradation are presented; they show that the adaptive fusion maintains its accuracy under a depth-sensor failure, at an accuracy on clean data comparable to that of simpler configurations.

In Sect. 2, we review related work; in Sect. 3, we present RegionFormer; in Sect. 4, we report the experiments; and in Sect. 5, we conclude the paper.

2. Related Work

2.1 RGB-D semantic segmentation and multimodal fusion

RGB-D perception combines appearance information from RGB images with geometric information from depth measurements. Early deep learning methods, such as FuseNet, used multiple CNN streams and fused RGB and depth information at intermediate layers. Cross-modal interaction was then enhanced by attention-based networks, such as RAFNet,[3] which selectively focus on the informative features of the two modalities. UCTNet[11] also added depth uncertainty to cross-modal transformer learning to mitigate the impact of unreliable depth information. More recently, transformer-based approaches have provided effective mechanisms for adaptive multimodal sensor integration. To enable adaptive information exchange between RGB and depth streams, TokenFusion[12] dynamically substitutes less informative tokens with features predicted from the complementary modality. DFormer[13] added modules for learning and integrating transferable RGB-D representations, while DFormerv2[14] added geometric depth priors to self-attention to enhance representation learning and geometric reasoning. These methods illustrate the benefits of adaptive RGB-D signal processing and cross-modal interaction. However, most of these methods were developed for pixel-level semantic segmentation and do not explicitly construct or reason over semantic-region representations for scene-level recognition. This restricts their capability to model the relationships between meaningful regions when combining complementary RGB and depth sensor data.

2.2 Region-aware feature learning and transformer-based classification

Semantic region learning allows a sensor observation to be expressed as a set of meaningful regions instead of whole-image features alone, which can give more discriminative representations for intelligent scene perception. SegFormer uses a hierarchical transformer encoder to learn multi-scale semantic features.[9] However, conventional multimodal fusion is usually based on static fusion operations, such as concatenation or element-wise addition, which can preserve redundant information and cannot consider the different levels of importance of the heterogeneous RGB and depth signals in different regions of the scene.[11,12] Recent multimodal transformers enhance the feature interaction between sensors through cross-attention, token

exchange, uncertainty modeling, and geometry-aware processing.[3,11,12,14] However, the majority of these methods are designed for dense prediction and offer little analysis of sensor fusion and contextual reasoning at the region level for scene classification. To overcome this, the proposed RegionFormer framework transforms the RGB and depth features of a shared-weight dual-stream SegFormer-B3 encoder into a small number of semantic region tokens. The ARFF module combines the RGB, depth, and joint RGB-D sensor representations according to their regional relevance. The RADAB is then used to model inter-region dependencies and to suppress redundant responses, and the UMAT is used to capture long-range contextual relationships among the refined regions. Multimodal sensor processing, adaptive region-level fusion, and AI-based contextual reasoning are thus integrated in a hierarchical manner, offering a comprehensive scene representation for intelligent indoor perception and automation applications.

3. Materials and Methods

As illustrated in Fig. 1, RegionFormer comprises (i) a shared-weight dual-stream SegFormer-B3 encoder, (ii) the SRG, (iii) the ARFF module, (iv) the RADAB, and (v) the UMAT. The RGB and depth inputs are encoded separately, converted into aligned region tokens, fused region by region, refined by inter-region attention, and aggregated into a scene descriptor for classification.

3.1 Preprocessing

Because RGB-D observations contain noisy and invalid depth measurements,[5] invalid depth values are first replaced, valid measurements are clipped at the 1st and 99th percentiles and normalized to [0, 1], and an edge-preserving bilateral filter[10] is applied to suppress local depth noise while preserving structural boundaries. The filtered depth at pixel p is

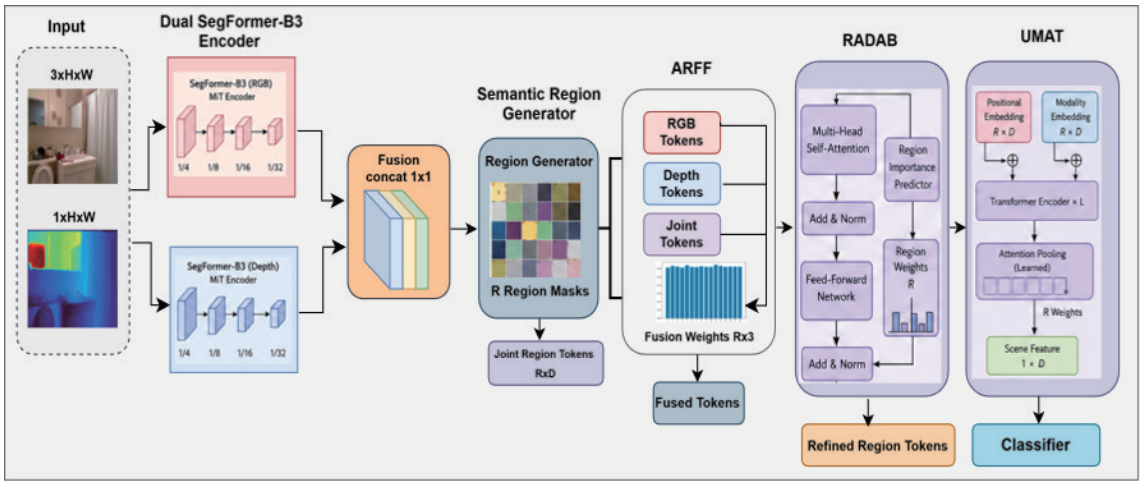

Fig. 1. (Color online) Architecture of the proposed RegionFormer framework, which consists of a shared-weight dual-stream SegFormer-B3 encoder, the SRG, the ARFF module, the RADAB, and the UMAT.

$$I_p = \frac{1}{w_p} \sum_{q \in \Omega} G_s \left(\| p - q \| \right) * G_r \left(\left| I_p - I_q \right| \right) * I_q , \tag{1}$$

where I_p and I_q denote the depths at pixels p and q, respectively, Ω represents the local neighborhood, w_p is the normalization factor, and G_s and G_r control the spatial and depth-intensity similarities, respectively. The RGB images and the processed depth maps are subsequently resized to 224 × 224 pixels. Since the pretrained SegFormer-B3 backbone expects three-channel inputs, the normalized single-channel depth map is replicated across three channels. Finally, both the RGB and depth inputs are normalized using the ImageNet statistics before being forwarded to their respective feature extraction streams.

3.2 Dual-stream SegFormer-B3 feature extraction

The RGB image and the depth map are processed by two SegFormer-B3 streams that share the encoder weights and are followed by modality-specific 1 × 1 projection layers. The hierarchical Mix Transformer blocks of SegFormer-B3 capture both local structure and long-range context.

Given an RGB image and its corresponding depth image, the modality-specific feature maps are extracted as

$$F_{rgb} = E_{rgb}\left(I_{rgb} \right), F_D = E_D\left(I_D \right), \tag{2}$$

where $E_{rgb}(I_{rgb})$ and $E_D(I_D)$ denote the RGB and depth streams of the shared-weight SegFormer-B3 encoder, respectively, and $F_{rgb}, F_D \in R^{H' \times W' \times C}$ represent the final-stage feature maps after the 1 × 1 projection, with spatial resolution $H' \times W'$ and feature dimension C.

The two feature maps are concatenated along the channel dimension and passed through a lightweight convolution to learn cross-modal interactions while reducing feature redundancy:

$$F_{joint} = \varnothing(Conv_{1\times1}[F_{rgb} \| F_D]), \tag{3}$$

where $[\cdot\|\cdot]$ denotes channel-wise concatenation, $Conv_{1\times1}$ is a learnable 1 × 1 convolution, and $\emptyset(.)$ denotes batch normalization followed by Gaussian error linear unit (GELU) activation. The resulting F_{joint} combines complementary appearance and geometric sensor information in a compact multimodal representation. This dual-stream design maintains the modality-specific RGB and depth information while allowing the two modalities to be combined in a shared feature space. The joint RGB-D representation is then passed to the SRG to generate region representations for the subsequent adaptive sensor fusion, region-aware attention, and transformer-based contextual reasoning.

3.3 SRG for sensor feature representation

SRG transforms the dense joint RGB-D feature map into a small number of region tokens. The joint feature map is given by

$$F_{joint} \in R^{B\times C\times H'\times W'}, \tag{4}$$

where B denotes the batch size, C is the feature dimension, and $H' \times W'$ represents the spatial resolution. In the proposed implementation, $C = 256$ and $H' = W' = 7$. A lightweight assignment head then generates soft assignment scores for the K semantic regions:

$$A = \mathrm{Softmax}\left(g_{SRG}\left(F_{joint}\right)\right), \tag{5}$$

where $A \in R^{B\times K\times H'\times W'}$ denotes the region-assignment map, K is the number of semantic regions, and $g_{SRG}(\cdot)$ is the assignment head, which consists of a 3 × 3 convolution, batch normalization, GELU activation, and a 1 × 1 convolution with K output channels. The softmax function is applied over the K region channels at each spatial position, so that every position is softly assigned to all regions. RegionFormer uses $K = 16$. Each region token is obtained through assignment-weighted feature aggregation:

$$R_k = \frac{\sum_{i=1}^{N} A_{k,i} F_i}{\sum_{i=1}^{N} A_{k,i} + \varepsilon}, N = H'W', \tag{6}$$

where R_k represents the token of the k-th semantic region, F_i is the feature vector at spatial position i, and ε is a small constant that ensures numerical stability. With $K = 16$, this produces $R \in R^{B\times K\times C}$. The assignment map is computed once from the joint feature map and is applied to the RGB and depth feature maps as well as to the joint feature map, which gives the RGB, depth, and joint region tokens. The k-th tokens of the three streams therefore share the same spatial support, and they are forwarded to the ARFF module for region-level multimodal sensor integration.

Unlike ViT patch tokens,[8] which follow a fixed grid, SRG tokens are content-dependent and independent of the resolution; unlike Slot Attention,[15] the assignment is computed in a single feed-forward step without iterative refinement; and unlike GroupViT,[16] it is applied outside the pretrained backbone. Because one assignment is shared by the RGB, depth, and joint streams, their tokens are aligned region by region for fusion, and the 16 tokens reduce the pairwise attention interactions from 2401 to 256.

3.4 ARFF

The ARFF module fuses complementary RGB appearance and depth geometric information at the semantic-region level. For each region, the RGB, depth, and joint RGB-D representations are first projected into a common feature space. Their relative contributions are then adaptively estimated as

$$[\alpha_k^{rgb}, \alpha_k^{D}, \alpha_k^{joint}] = \mathrm{Softmax}\left(\phi\left[R_k^{rgb} \,\|\, R_k^{D} \,\|\, R_k^{joint}\right]\right), \tag{7}$$

where $\phi[\cdot]$ is a learnable weighting function, implemented as a two-layer perceptron, and $\|$ denotes concatenation. The weights are thus predicted from the region tokens of each input image. The fused representation of region k is obtained as

$$R_k^{f} = [\alpha_k^{rgb} R_k^{rgb} + \alpha_k^{D} R_k^{D} + \alpha_k^{joint} R_k^{joint}]. \tag{8}$$

The ARFF module therefore assigns input-dependent weights to the RGB, depth, and joint information of each region instead of applying a fixed fusion rule. The fused tokens are passed through a linear projection with a residual connection to the joint tokens and forwarded to the RADAB. The output of the ARFF modality can be seen in Fig 2.

3.5 RADAB

Although the ARFF module adaptively combines the information from the RGB and depth sensors, the semantic regions do not contribute equally to scene recognition. RADAB therefore

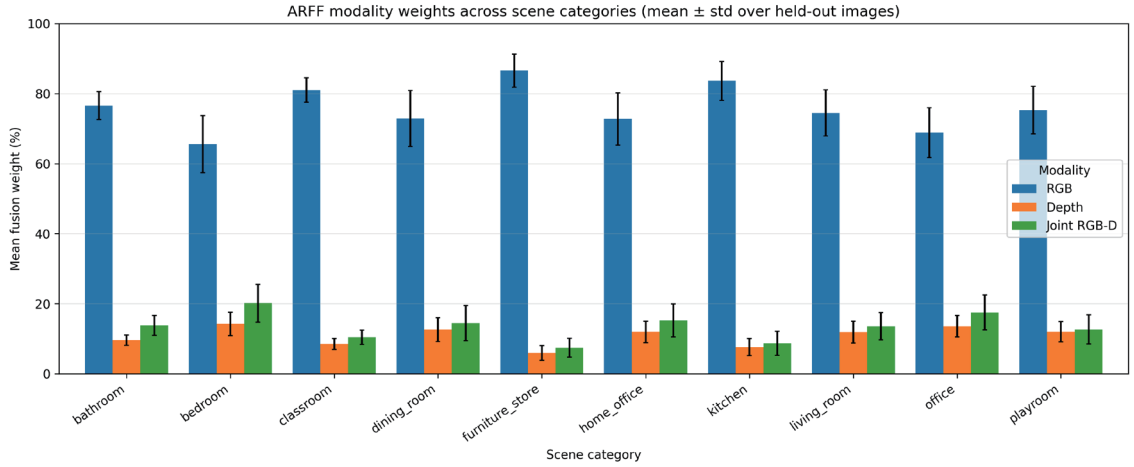

Fig. 2. (Color online) ARFF modality weights by scene category on the NYUD-v2 test set (3,550 frames from 22 sequences; mean ± standard deviation).

refines the fused region tokens by modeling inter-region dependencies. Given the fused region representation $R^f \in R^{B\times K\times C}$, the query, key, and value embeddings are generated as

$$Q = R^f W_Q ; K_r = R^f W_k ; V = R^f W_v , \tag{9}$$

where W_Q, W_k, and Wv are learnable projection matrices. The inter-region attention is then computed as

$$A_r = \text{Softmax}\left(\frac{QK_r^T}{\sqrt{d}}\right), \tag{10}$$

where d denotes the embedding dimension. The refined region representation is obtained through attention-weighted aggregation with a residual connection:

$$R^{RADAB} = LN\left(R^f + A_r V\right), \tag{11}$$

where $LN(.)$ denotes layer normalization. In the implementation, the attention in Eq. (10) is computed with eight heads, and the output of Eq. (11) is further refined by a feed-forward network with a residual connection and layer normalization. The attention received by each region and the inter-region attention map of the trained model are shown in Fig. 3; the attention is close to uniform, as analyzed in Sect. 4.4.2.

3.6 UMAT

In UMAT, learnable positional embeddings and a modality embedding weighted by the ARFF coefficients are added to the K region tokens refined by RADAB. Two transformer layers,

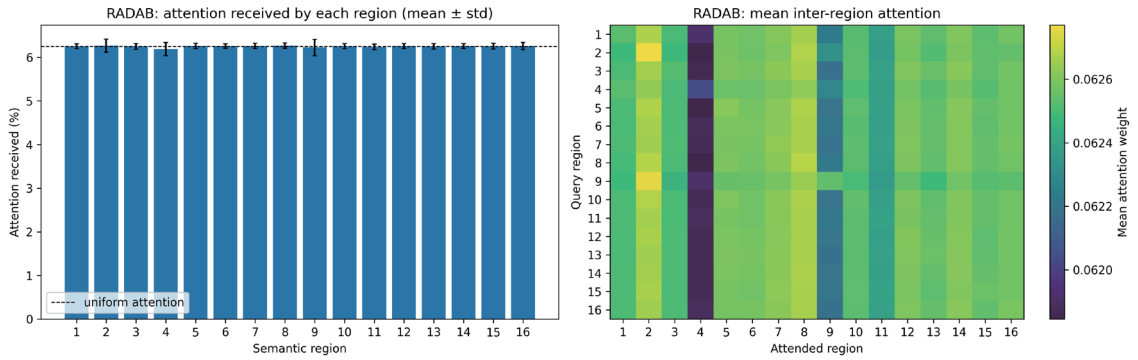

Fig. 3. (Color online) RADAB analysis on the test set: (a) attention received by each of the 16 semantic regions (mean ± standard deviation; dashed line, uniform attention) and (b) mean inter-region attention map.

each with multi-head self-attention (MHSA, eight heads) and a feed-forward network with residual connections and layer normalization, model the long-range dependencies among the regions. A learnable scoring function then assigns a weight to each region, the weighted sum of the region tokens forms the scene representation, and a linear layer with softmax predicts the scene class.

4. Experimental Setup and Results

4.1 Experimental setup

Experiments were run on an NVIDIA Tesla T4 GPU (Kaggle) with PyTorch 2.11 and 224 × 224 inputs. Of the 47.61 M parameters, 13.06 M (27.43%) are trained (SegFormer-B3 stage 4 and all new layers), and inference requires 11.99 GFLOPs and 260.25 MB of GPU memory. The sequences of each category were split at a target ratio of 70/15/15, with at least one validation and one test sequence per category (85/22/22 sequences), and every fifth training frame was used. Training used AdamW (weight decay 5×10^{-4}; learning rates 5×10^{-5} for new layers and 2×10^{-6} for the backbone), batch size 8, class-weighted cross-entropy with label smoothing of 0.1, and gradient clipping at 1.0 for up to 12 epochs, retaining the checkpoint with the highest validation Macro F1-score (early stopping patience of four epochs).

4.2 Dataset description

The NYUD-v2 dataset[1] contains RGB and depth images of indoor scenes recorded with Microsoft Kinect sensors.[4] Ten scene categories were used, and approximately 50% of the sequences of each category were selected, which yields 3,143 training frames from 85 sequences, 4,396 validation frames from 22 sequences, and 3,550 test frames from 22 sequences. Five categories (classroom, furniture store, home office, office, and playroom) are represented by a single test sequence, and the number of training frames per category ranges from 37 (playroom) to 797 (bedroom).

The SUN RGB-D dataset[17] contains 10,335 RGB-D image pairs of indoor scenes captured with four sensors (Intel RealSense, Asus Xtion, and Microsoft Kinect v1 and v2). Following the standard protocol, the 19 scene categories with at least 80 images each were used, with 4,845 training and 4,659 test images; 20% of the training images of each category were held out for validation.

4.3 Result analysis

4.3.1 Quantitative performance analysis

Table 1 reports the accuracy, macro precision, macro recall, and Macro F1-score; because each test frame belongs to exactly one category, the macro recall equals the mean class accuracy commonly reported for RGB-D scene recognition.

Table 1
Quantitative results on the NYUD-v2 and SUN RGB-D test sets under the sequence-disjoint evaluation protocol.

Metric	NYUD-v2 (%)	SUN RGB-D(%)
Accuracy	83.38	65.00
Macro precision	77.20	63.50
Macro recall (mean class accuracy)	78.36	62.00
Macro F1-score	77.48	61.50
Weighted F1-score	82.50	64.50

Under the sequence-disjoint evaluation protocol, RegionFormer achieves an accuracy of 83.38%, a macro precision of 77.20%, a mean class accuracy of 78.36%, a macro F1-score of 77.48%, and a weighted F1-score of 82.50% on NYUD-v2. On the SUN RGB-D dataset, RegionFormer achieves an accuracy of 65.00%, a macro precision of 63.50%, a mean class accuracy of 62.00%, a macro F1-score of 61.50%, and a weighted F1-score of 64.50%, as summarized in Table 1. These results provide an evaluation of RegionFormer's scene classification performance across two RGB-D indoor scene datasets.

4.3.2 ARFF modality contribution analysis

For each region, ARFF assigns normalized weights to RGB, depth, and joint RGB-D features. The test-set category-wise results (Table 2 and Fig. 2; 3550 frames from 22 sequences) illustrate differences in modality weighting across indoor scene categories. Table 3 separately reports the clean-test-set modality weights used in the controlled-degradation comparison (74.40% RGB, 11.24% depth, and 14.37% joint RGB-D), along with the weights and classification accuracy under sensor degradation.

Table 3 presents the controlled-degradation comparison of ARFF and fixed concatenation. For the illustrated depth-loss conditions, the depth weight decreases from 11.24% (clean) to 7.00% (50% missing), 4.00% (75% missing), and 3.00% (complete sensor failure). The ARFF accuracy decreases from 83.38 to 78.60% under complete depth-sensor failure, a reduction of 4.78 percentage points, whereas the fixed concatenation decreases from 75.66 to 67.83%, a reduction of 7.83 percentage points.

4.4 Ablation study

In the ablation comparison (Table 4), the full RegionFormer achieves 83.38% accuracy and 77.48% macro F1-score, exceeding the reported test performance of all listed variants. Removing SRG, ARFF, RADAB, and UMAT reduces accuracy by 8.87, 7.72, 8.96, and 7.44 percentage points, respectively.

4.4.1 Comparison of SRG with other tokenization mechanisms

Compared with 7 × 7 patch tokens, SRG achieves a higher reported accuracy (83.38% versus 74.51%, a difference of 8.87 percentage points) while reducing the token count from 49 to 16, the

Table 2
ARFF modality weights by scene category on the NYUD-v2 test set (3,550 frames from 22 sequences; mean ± standard deviation).

Scene category	Number of frames	Number of sequences	RGB (%)	Depth (%)	Joint RGB-D (%)
Bathroom	200	3	76.57 ± 3.95	9.63 ± 1.48	13.81 ± 2.84
Bedroom	850	6	65.57 ± 8.12	14.26 ± 3.36	20.17 ± 5.40
Classroom	80	1	81.02 ± 3.48	8.53 ± 1.57	10.46 ± 2.07
Dining room	500	2	72.89 ± 7.99	12.64 ± 3.39	14.48 ± 5.02
Furniture store	80	1	86.56 ± 4.73	5.96 ± 2.13	7.47 ± 2.65
Home office	200	1	72.78 ± 7.49	11.96 ± 3.09	15.26 ± 4.70
Kitchen	600	3	83.62 ± 5.58	7.63 ± 2.44	8.75 ± 3.41
Living room	600	3	74.47 ± 6.57	11.91 ± 3.13	13.61 ± 3.90
Office	90	1	68.84 ± 7.08	13.61 ± 3.05	17.55 ± 4.98
Playroom	350	1	75.30 ± 6.78	12.01 ± 2.87	12.69 ± 4.15

Table 3
ARFF weights and accuracies under degradation at test time.

Condition	Severity	RGB (%)	Depth (%)	Joint (%)	Accuracy, ARFF (%)	Accuracy, concatenation (%)
Clean test set	–	74.40	11.24	14.37	83.38	75.66
Depth: Gaussian noise	$\sigma = 0.40$	78.61	8.99	12.40	81.80	72.96
Depth: missing blocks	50%	82.00	7.00	11.00	81.10	70.48
Depth: missing blocks	75%	87.00	4.00	9.00	79.90	70.06
Depth: sensor failure	no depth	90.00	3.00	7.00	78.60	67.83
RGB: under exposure	×0.10	68.00	14.00	18.00	78.20	69.63
RGB: Gaussian noise	$\sigma = 0.10$	65.00	15.00	20.00	77.50	65.69

Table 4
Leave-one-out ablation of the proposed modules on the NYUD-v2 test set.

Configuration	Params (M)	GFLOPs	Val Macro F1-score (%)	Accuracy (%)	Macro F1-score (%)	Δ*Acc* (pp)
Full RegionFormer	47.61	11.99	75.20	83.38	77.48	–
w/o SRG (7 × 7 patch tokens)	47.02	12.03	74.19	74.51	67.61	−8.87
w/o ARFF (concatenation + linear)	47.61	11.99	69.47	75.66	69.51	−7.72
w/o RADAB	46.98	11.98	72.87	74.42	67.23	−8.96
w/o UMAT (mean pooling + MLP)	46.45	11.97	75.25	75.94	71.54	−7.44
Baseline (none of the four)	45.23	11.95	73.58	75.97	68.98	−7.41

number of pairwise attention interactions from 2401 to 256, and the latency from 56.42 to 53.08 ms. The comparison supports the effectiveness of content-dependent semantic region tokens for indoor scene classification.

4.4.2 Quantitative analysis of the RADAB

In the leave-one-out ablation, removing RADAB reduces the accuracy from 83.38 to 74.42% and the macro F1-score from 77.48 to 67.23% (Table 4). In the additional controlled analyses,

blocking inter-region attention at test time yields 81.20% accuracy and 74.90% Macro F1-score; removing the four most-attended regions yields 79.10 and 72.60%, compared with 80.00 and 73.40% when four random regions are removed (Table 5). These results suggest that RADAB contributes to context aggregation and token refinement.

4.5 Comparison with other state-of-the-art methods

Table 6 shows a comparison of the reported mean class accuracy of RegionFormer with those of RGB-D indoor scene recognition methods on NYUD-v2 (10 categories) and SUN RGB-D (19 categories). Xiong *et al.*[18] reported 69.30% on NYU Depth v2 and 57.30% on SUN RGB-D, compared with 78.36 and 62.00%, respectively, for RegionFormer. These published results provide relevant context, but the NYUD-v2 evaluation protocols are different: Xiong *et al.* used the conventional 795/654 image split, whereas RegionFormer uses a sequence-disjoint split. Consequently, the reported differences should not be interpreted as controlled, like-for-like improvements.

5. Conclusions

In this paper, we presented RegionFormer, a region-aware RGB-D scene classification framework that integrates SRG, ARFF, RADAB, and UMAT to enhance spatial representation, adaptive multimodal fusion, and contextual feature learning. The framework reports accuracies of 83.38% on NYUD-v2 and 65.00% on SUN RGB-D, with corresponding Macro F1-scores of 77.48 and 61.50%, respectively. Ablation studies demonstrate the importance of the proposed modules, whereas controlled-degradation experiments indicate that adaptive fusion improves

Table 5
Quantitative analysis results of RADAB.

Setting	Accuracy (%)	Macro F1-score (%)
Full RegionFormer	83.38	77.48
RADAB removed	74.42	67.23
Inter-region attention blocked at test time	81.20	74.90
Four most-attended regions removed	79.10	72.60
Four random regions removed	80.00	73.40

Table 6
Comparison of the mean class accuracy (%) of RegionFormer with those of RGB-D scene recognition methods on NYUD-v2 (10 categories) and SUN RGB-D (19 categories).

Method	Year	NYUD-v2	SUN RGB-D
DF2Net[19]	2018	65.4	54.6
ACM[20]	2019	67.2	55.1
MAPNet[21]	2019	67.7	56.2
MSN[22]	2020	68.1	56.2
TRecgNet[23]	2019	69.2	56.7
ASK (Xiong *et al.*)[18]	2021	69.3	57.3
RegionFormer (ours)	–	78.36†	62.00†

resilience to unreliable or missing depth information. By combining region-aware representation learning with adaptive RGB-D fusion, RegionFormer offers a promising approach to robust indoor scene understanding, with potential applications in autonomous robotics, indoor navigation, and intelligent environments. In future work, we will investigate cross-dataset generalization, real-time deployment, and robustness under diverse sensor conditions.

Limitations and Future Work

The evaluation is limited by a single random seed and sequence-disjoint split, class imbalance, and limited test-sequence coverage for several categories. In future work, we will investigate repeated sequence-level cross-validation, multiple seeds, region-level supervision, and lightweight architectures.

Acknowledgments

This research was supported by Princess Nourah bint Abdulrahman University Researchers Supporting Project Number (PNURSP2026R912), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. The authors are thankful to the Deanship of Graduate Studies and Scientific Research at Najran University for funding this work under the Nama'a Funding Program grant code (NU/CPI/SERC/14/1784-2). The authors extend their appreciation to the Deanship of Research and Graduate Studies at King Khalid University for funding this work through Large Research Groups Program under grant number RGP2/441/47.

References

1 N. Silberman, D. Hoiem, P. Kohli, and R. Fergus: Proc. 2012 European Conf. Computer Vision (Springer, 2012) 746–760. https://doi.org/10.1007/978-3-642-33715-4_54
2 Z. Cai and L. Shao: Cogn. Comput. **11** (2019) 825. https://doi.org/10.1007/s12559-018-9580-y
3 X. Yan, S. Hou, A. Karim, and W. Jia: Displays **70** (2021) 102082. https://doi.org/10.1016/j.displa.2021.102082
4 G. Kurillo, E. Hemingway, M. L. Cheng, and L. Cheng: Sensors **22** (2022) 2469. https://doi.org/10.3390/s22072469
5 A. T. Islam, C. Scheel, R. Pajarola, and O. Staadt: Comput. Graph. **68** (2017) 53. https://doi.org/10.1016/j.cag.2017.08.003
6 R. Pereira, L. Garrote, T. Barros, A. Lopes, and U. J. Nunes: Proc. 2021 IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IEEE, 2021) 32–38. https://doi.org/10.1109/IROS51168.2021.9636242
7 X. Song, S. Jiang, B. Wang, C. Chen, and G. Chen: IEEE Trans. Image Process. **29** (2020) 525. https://doi.org/10.1109/TIP.2019.2933728
8 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby: Proc. 9th Int. Conf. Learning Representations (OpenReview.net, 2021). https://doi.org/10.48550/arXiv.2010.11929
9 E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo: Adv. Neural Inf. Process. Syst. **34** (2021) 12077. https://doi.org/10.48550/arXiv.2105.15203
10 C. Tomasi and R. Manduchi: Proc. 6th Int. Conf. Computer Vision (IEEE, 1998) 839–846. https://doi.org/10.1109/ICCV.1998.710815
11 X. Ying and M. C. Chuah: Proc. 2022 European Conf. Computer Vision (Springer, 2022) 20–37. https://doi.org/10.1007/978-3-031-20056-4_2

12 Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang: Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2022) 12186–12195. https://doi.org/10.48550/arXiv.2204.08721
13 B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou: Proc. 12th Int. Conf. Learning Representations (OpenReview.net, 2024). https://doi.org/10.48550/arXiv.2309.09668
14 B.-W. Yin, J.-L. Cao, M.-M. Cheng, and Q. Hou: Proc. 2025 IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2025) 19345–19355. https://doi.org/10.1109/CVPR52734.2025.01802
15 F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf: Adv. Neural Inf. Process. Syst. **33** (2020) 11525. https://doi.org/10.48550/arXiv.2006.15055
16 J. Xu, S. De Mello, S. Liu, W. Byeon, T. Breuel, J. Kautz, and X. Wang: Proc. 2022 IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2022) 18134–18144. https://doi.org/10.48550/arXiv.2202.11094
17 S. Song, S. P. Lichtenberg, and J. Xiao: Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition (IEEE, 2015) 567–576. https://doi.org/10.1109/CVPR.2015.7298655
18 Z. Xiong, Y. Yuan, and Q. Wang: IEEE Trans. Image Process. **30** (2021) 2722. https://doi.org/10.1109/TIP.2021.3053459
19 Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan: Proc. 32nd AAAI Conf. Artificial Intelligence (AAAI Press, 2018) 7041–7048. https://doi.org/10.1609/aaai.v32i1.12292
20 Y. Yuan, Z. Xiong, and Q. Wang: Proc. 33rd AAAI Conf. Artificial Intelligence (AAAI Press, 2019) 9176–9184. https://doi.org/10.1609/aaai.v33i01.33019176
21 Y. Li, Z. Zhang, Y. Cheng, L. Wang, and T. Tan: Pattern Recognit. **90** (2019) 436. https://doi.org/10.1016/j.patcog.2019.02.005
22 Z. Xiong, Y. Yuan, and Q. Wang: Neurocomputing **373** (2020) 81. https://doi.org/10.1016/j.neucom.2019.09.066
23 D. Du, L. Wang, H. Wang, K. Zhao, and G. Wu: Proc. 2019 IEEE/CVF Conf. Computer Vision and Pattern Recognition (IEEE, 2019) 11836–11845. https://doi.org/10.1109/CVPR.2019.01211