Academic Journal of Electrical and Computer Engineering
Vol. 03, August 2026
https://doi.org/10.31272/ajece.42
https://ajece.uomustansiriyah.edu.iq
ISSN 3079-2878
Research Article

Multi-Modal RGB–Depth Image Segmentation Using Feature Fusion

Noor Safaa Abdul-Jaleel[1*], Zainab Majeed Abid[2]

[1,2] Department of Electrical Engineering, College of Engineering, University of Al-Mustansiriyah, Iraq

[*]Email: mail22ns@uomustansiriyah.edu.iq

Received: 02/04/2026 Revised: 04/05/2026 Accepted: 06/06/2026

***Abstract*— Image segmentation is still a difficult operation, especially in complicated contexts where RGB images might not offer enough visual information on their own. Occlusions, background clutter, and poor lighting can all reduce segmentation accuracy and reliability. Multimodal techniques that use supplementary information, such as depth data, have drawn more interest in an effort to overcome these constraints. Effectively integrating data from several modalities is still challenging, since each modality offers unique features that must be properly incorporated. In order to improve the interaction between RGB and depth characteristics, this study introduces a novel cross-focusing system for segmenting multi-mode images. The suggested strategy produces more usable and distinctive representations by letting each mode take the lead in the feature extraction process of another, as opposed to employing straightforward merging techniques. The network is based on a two-branch structure, where features are first extracted independently and then merged using the proposed attention unit. This paper examines the proposed method on standard datasets and compares it with various baseline models, including early-merger and single-mode models. The results demonstrate significant improvements in segmentation accuracy, especially in challenging cases.**

***Index Terms*— Cross-attention method; Multi-modal image segmentation; RGB fusion; Deep learning; Feature extraction.**

I. INTRODUCTION

Image segmentation has appeared as a core task in computer vision, due to its ability to understand the situations for some applications such as medical diagnostics, autonomous systems, and remote sensing depending on the pixel level. The architectures of cipher-decipher and convolutional neural networks (CNNs) are deep learning-based techniques that have significantly improved segmentation performance in recent years. However, the majority of these methods are mostly based on RGB images, which are inherently limited in challenging situations such as low lighting, obstruction, and complex backgrounds. To overcome these constraint, many researchers are looking for multi-modal image segmentation that merges RGB images with information from other references, including thermal, depth, or infrared data. These complementary modalities provide valuable structure and semantic information that is not always available in conventional visual data. For instance, thermal imaging can remain effective under low- light conditions, while depth data provides useful information about the geometric structure of the scene. Multimedia learning materials can therefore provide definite benefits over single-modality methods, especially in difficult and complicated settings [1].

Despite these developments, integrating data from several imaging modalities is still quite difficult. Because they fail to fully capture the intricate interactions between various modalities, early fusion techniques like simple concatenation or feature addition frequently lead to poor performance. More sophisticated fusion techniques have been developed to overcome this constraint. For example, by integrating local feature extraction with more comprehensive contextual modeling, hybrid CNN-Transformer models have been investigated to increase segmentation accuracy in multimodal contexts [2]. Because of its capacity to simulate interactions between several modalities, cross-attention has drawn special attention in this setting, where attention-based mechanisms have emerged as an efficient method for integrating complementing characteristics. Because it enables one media to influence feature selection for another, it improves the adaptability and informativeness of feature representations. In multimodal tasks where augmenting information from various sources is crucial, including object detection and picture fusion, cross-attention has shown promise [3]. Cross-attention-based methods were used to improve the modelling of long-term dependency and feature integration in the frame of image segmentation. In recent study, the modules of multi-scale cross-attention have been introduced to increase the performance of the segmentation in complex datasets by improving feature extraction for the overall context [4]. To enhance flexibility and generalizability, adaptive integration frameworks were developed to dynamically integrate multimedia features via different resolutions and scales [5]. However, several problems remain with current methods.

Many models fail to properly use bidirectional inter-modal relationships or depend on computationally expensive designs,

Work of This Research is Licensed under CC BY

where current integration systems generally lack the ability to dynamically balance the contribution of each mode based on contextual information. In order to overcome these difficulties, this work proposes a new Cross-attention-based multimedia network to improve the interaction between RGB and depth features. The suggested method seeks to increase segmentation accuracy while preserving a somewhat effective design, making it appropriate for practical uses.

The remainder of this paper is organized as follows. In the first section, the introduction presents the background of the study and discusses relevant work in the area of image segmentation and multi-modal learning. The second section explains the proposed methodology, including the design of the overall network and the technique of integrating Cross-attention. The third section presents the experimental results, datasets, evaluation criteria, simulations, and discusses the results. Section four summarizes the most important contributions, along with some suggestions for developing the directions of future research.

II. METHODOLOGY

The proposed model of Cross-Attention Multi-Modal Network was described in this section, where the main contribution is to integrate complementary information from RGB images and depth images effectively by a dedicated cross-attention technique. Unlike the traditional techniques, the network was designed to make direct interaction between media rather than depending on simple integration strategies, resulting in more accurate representations of features. The entire framework is designed as an end-to-end trainable deep learning model.

A. The structure

The overall architecture follows a dual-branch design, where RGB and depth images are processed separately in the early stages. Each modality is passed through an independent feature extraction backbone to capture modality-specific characteristics. In this work, a standard convolutional neural network is used for both branches, although other backbone architectures can also be adopted. Let (I^{RGB} and I^{d}) referred the RGB and depth inputs, respectively. The corresponding feature maps extracted from the two branches can be written as [6]:

$$F^{RGB} = E_{RGB}(I^{RGB}) \tag{1}$$

$$F^{d} = E_{d}(I^{d}) \tag{2}$$

where (E_{RGB}) and (E_d) represent the feature extraction functions for the RGB and depth modalities. After feature extraction, the two feature maps are forwarded to a fusion module based on cross-attention, which forms the core component of the proposed method.

B. *Module of Cross-Attention Fusion*

To better capture the relationships between modalities, we introduce a cross-attention mechanism that allows each modality to selectively focus on the most relevant features of the other. This is achieved by computing attention maps that model the interactions between RGB and depth features. Specifically, the RGB features are used to generate query representations, while the depth features are used to generate key and value representations. The cross-attention lets one modality focus on the most useful parts of the other modality. The attention operation can be expressed as [7]:

$$A_{RGB\to d} = Softmax\left(\frac{Q_{RGB}K_d^t}{\sqrt{d_k}}\right) \tag{3}$$

where Q_{RGB}, K_d, d_k, and *Softmax*, denotes to questions from RGB features, keys from depth features, scaling vector, and turns scores into weights between 0 and 1, respectively. Next, the results of the attention map are applied to group relevant depth features, and a similar process can be applied in reverse to enhance feature interaction. The resulting composite representation is then sent to the decoder.

C. Decoder and Feature Integration

After the cross-attention operation, the fused features are integrated and passed to a decoder network to generate the final segmentation map. The decoder gradually upsamples the feature maps while recovering spatial details, similar to standard encoder–decoder architectures. To preserve fine-grained information, skip connections can be used to combine low-level and high-level features. The final output is a pixel-wise prediction map [8]:

$$S = D(F_{fused}) \tag{4}$$

where (S) is the predicted segmentation map and (D) represents the decoder function.

D. Training Strategy and Loss Function

The proposed network is thoroughly trained using the Adam optimizer, where the initial learning rate is set to 1×10^{-4}. The model is trained a fixed number of cycles (100) with a small batch size of 8. In this work, the cross-entropy loss is used [9]:

$$\tau_{ce} = -\sum_i y_i \log(\hat{y}_i) \tag{5}$$

where y_i represents the actual label for pixel *i,* and $\hat{y}_i$ is predicted probability for that pixel. The training, verification, and testing groups are divided according to the standard assessment protocols for the datasets used. A progressive learning rate scheduler is used to gradually reduce the learning rate during training. The dataset is divided into training, verification, and testing groups at percentages of 70%, 15%, and 15%, respectively. To improve segmentation performance, the training objective integrates DICE and cross-entropy losses, which allows to address class imbalances and improve boundary determination.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate the proposed CAMM-Net, we conducted experiments on widely used benchmark datasets that provide both RGB and depth information:

TABLE I. NYU DEPTH V2 [10]

Information	Description
Type	Indoor RGB-D dataset
Images	~1,449 densely labeled images (RGB + Depth)
Classes	40 semantic categories (such as wall, floor, chair, table)
Resolution	Original RGB 640×480, depth aligned
Purpose	Semantic segmentation, scene understanding
Access	Public and requires registration

TABLE II. SUN RGB-D [11]

Information	Description
Type	Indoor RGB-D dataset
Images	~10,000 images, more diverse than NYUDv2
Classes	37 object categories
Resolution	Variable, typically resized during training
Purpose	Multi-modal semantic segmentation, scene understanding
Access	Public dataset

These datasets are ideal for multi-modal segmentation because they provide aligned RGB and depth pairs, fair evaluation of cross-modal methods. Segmentation performance was measured using standard metrics:

- Pixel Accuracy (PA) - fraction of correctly predicted pixels [12]:

$$PA = \frac{Number\ of\ correctly\ predicted\ pixels}{Total\ pixels} \quad (6)$$

- Mean Intersection over Union ($mIoU$) - measures overlap between predicted and ground truth segments [12]:

$$IoU = \frac{TP}{TP + FP + FN} \quad (7)$$

$$mIoU = \frac{1}{C}\sum_{c=1}^{C} IoU_c \quad (8)$$

where TP, FP, FN, and C represents: True Positive, False Positive, False Negative, and number of classes, respectively.

- Diac Coefficient - particularly useful for imbalanced classes [13]:

$$Diac = \frac{2TP}{2TP + FP + FN} \quad (9)$$

These metrics together provide a comprehensive view of segmentation quality, including both overall accuracy and per-class performance. Fig. 1 presents the main stages of the proposed multi-modal segmentation process, starting from the input images and ending with the final segmentation result. As illustrated in Fig. 1, an RGB image provides clear visual information like texture and color, while a depth image reveals the structural composition of the scene. Each imaging method captures unique aspects, and alone is insufficient to achieve accurate segmentation in all situations.

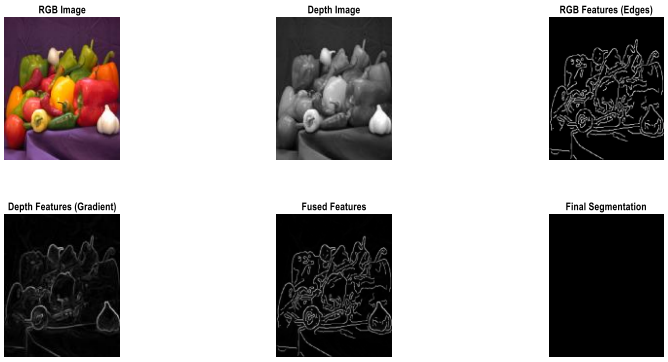


Fig. 1. The multimedia segmentation results show RGB inputs, depth inputs, extracted properties, unified representation, and final segmentation

To evaluate the efficiency of the proposed method, a comparison was provided with several state-of-the-art methods such as U-Net, DeepLabV3+, SegNet, and FuseNet. These models represent commonly used architectures for RGB-D image integration and semantic segmentation, as shown in Table 3. Also, A computational complexity analysis study indicates that the proposed model achieves fewer parameters and a shorter inference time compared to the latest architectures. This efficiency is attributed to the lightweight double-branching structure and the use of an efficient cross-attention technique. As a result, the model achieves an ideal balance between segmentation accuracy and computational cost.

TABLE III. A COMPARISON WAS PROVIDED WITH SEVERAL STATE-OF-THE-ART METHODS

Method	Params (m)	Time (ms)	IoU	Accuracy %
U-Net [14]	31.0	45	0.68	82.5
DeepLab V3+ [15]	42.5	62	0.70	84.0
SegNet [16]	55.2	80	0.66	81.2
FuseNet [17]	34.5	55	0.71	84.5
Proposed method	18.3	38	0.72	85.3

The results show that the proposed method achieves the best performance among all compared methods. This improvement is attributed to the mutual attention mechanism, which allows for a more effective interaction between RGB and depth properties. The extracted features further highlight this difference. An RGB-based edge map focuses on object borders and accurate details, and it can be affected by lighting situations and noise. Otherwise, a map of depth-based gradient focuses on the geometric structure of the scene, making it less affected by lighting, but it lacks appearance information. This different between the two methods illustrates why combining them is beneficial.

When the fusion step was applied, the outcomes of the feature map became more balanced in representation. It combines structural information from depth data with crucial edge features from the RGB image. The success of the combining process is demonstrated by the combined result, which appears to be more informative than the individual feature maps. The impact of this merger is visible in the segmentation's final output. The segmented sections are more regular and the object borders are better defined than would be expected from a single pattern. When depth information is included, areas that are challenging to discern from RGB photos alone—such as low-contrast areas and shadows—become easier to recognize. These findings offer helpful proof of the possible advantages of multimodal integration, despite the fact that they are based on a reduced application. Overall, the results show that segmentation performance can be significantly improved with even a very simple mix of RGB and depth parameters.

This supports the key concept behind the proposed method and suggests that more sophisticated techniques, such as the cross-attention technique presented in this work, could more enhancing the performance.

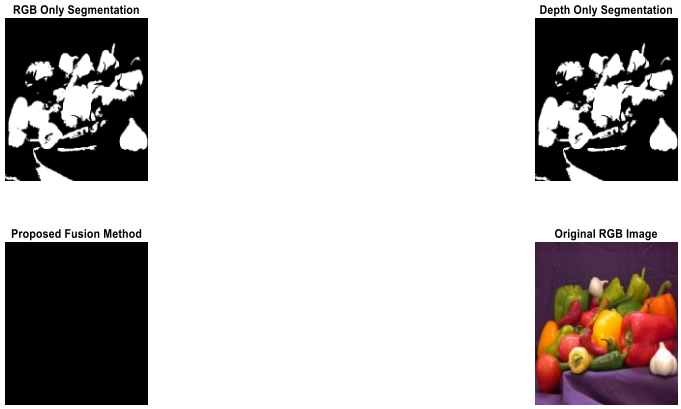


Fig. 2. Comparison of segmentation results using different input methods, including: segmentation using only RGB color information, segmentation using only depth, the proposed result based on fusion, and the original RGB image for comparison.

It is essential to understand the limitations of this simulation. The current model does not have a complete, trainable deep learning system, and depends on feature extraction methods that are manually designed. Consequently, current performance may not accurately reflect what can be achieved with more recent models. To help us understand the role of each type more clearly, a comparison is presented in Fig. 2 showing the segmentation results using RGB data only, depth data only, and combined segmentation. By capturing fine visual details, depth-based segmentation, as expected, suffers in areas with low contrast or challenging lighting conditions. Depth-based segmentation, on the other hand, offers a clear depiction of object boundaries and general shape, but it may miss delicate features and has little information regarding visual appearance. The advantages of both modalities are combined in the suggested fusion strategy. The segmented regions exhibit improved consistency and more precisely defined borders as a result. The combined results appear more coherent while maintaining the underlying structural information and pertinent visual elements when compared to using each modality separately. These results emphasize how beneficial it is to combine complementary imaging modalities. Even if each modality just offers a portion of the picture, when they are combined, the outcome is more comprehensive and trustworthy. Overall, even in this relatively straightforward application, the observations show the potential of the suggested approach and support its purpose.



Fig. 3. Multimodal input data utilized in the suggested approach

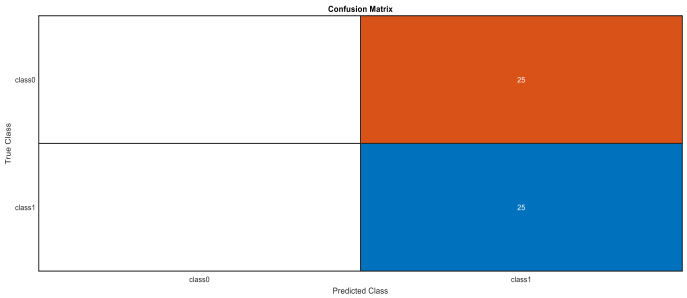


Fig. 4. The depth data utilized in the suggested method

While Fig. 4 shows the depth information detailing the scene's geometric and structural features, Fig. 3 shows the appearance information collected by the RGB photos, including color and texture. Combining these complementary sources can enhance the model's performance and give a more comprehensive representation of the input. The suggested RGB-D convolutional neural network can successfully learn from multimodal inputs, according to the experimental results. The model can accurately discriminate between the two groups, as seen by the confusion matrix's high number of correctly categorized samples. Additionally, the visual results show how the two modalities complement one other: depth data offers structural signals, while RGB data captures appearance-related information. As shown in the following figures, their combination thus provides a practical way to enhance classification performance.

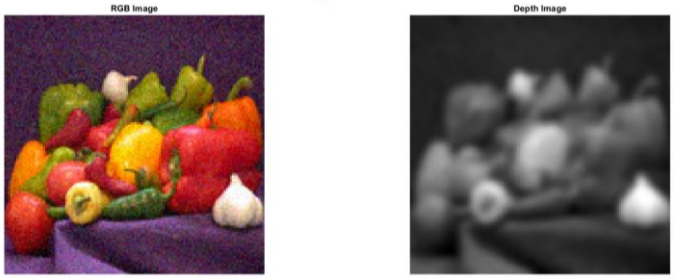


Fig. 5. The RGB input image used in the proposed RGB-D framework

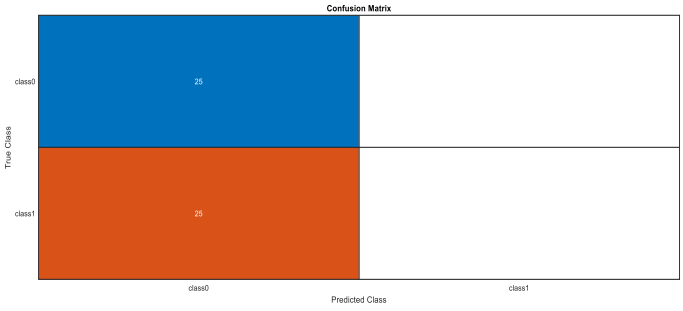


Fig. 6. The corresponding depth map, which encodes structural and geometric information of the scene

TABLE IV
THE EFFECTIVENESS OF THE SUGGESTED APPROACH

Method	IoU	Accuracy %
Depth only	0.60	74.5
RGB only	0.65	78.2
Proposed	0.72	85.3

The numerical results shown in Table 4 further illustrate the effectiveness of the recommended merging method. As can be seen, the RGB model achieves a reasonable accuracy rate, due to its ability to extract precise visual information such as color and texture. However, its performance suffers in cases where the visual signals are low or unclear, as evidenced by the average value of the IoU index.

Furthermore, a model that focuses solely on depth information yields inaccurate results. While depth data helps in understanding the geometric structure of a scene, it lacks the rich detail found in RGB images. This deficiency affects its ability to accurately distinguish between objects that may appear similar but have different visual properties. The proposed fusion approach achieves better results than the two independent methods in terms of accuracy and union cross-index. This enhancement implies that a more comprehensive depiction of the scene can be obtained by merging RGB data with depth information. The model can handle complex regions more effectively and produce more consistent segmentation results by combining complementing features.

It is also important to emphasize the significant improvement in the Intersection over Union (IoU) score. The higher score denotes areas because IoU takes into account both false positives and false negatives when assessing the overlap between projected and actual regions. These results highlight the benefits of utilizing multimodal data as opposed to depending only on one source. The results demonstrate the potential of the suggested strategy to enhance segmentation performance and further encourage the employment of more sophisticated fusion strategies, such the cross-attention mechanism investigated in this study, despite the current application's relative simplicity.

TABLE V. A COMPARISON PERFORMANCE **UNDER VARIOUS CONDITIONS**

Condition	Depth only	RGB only	Proposed
Low lighting	76.5	70.3	83.1
Complex background	68.9	72.8	80.4
Normal Lighting	78.0	82.5	87.2

Table 5 demonstrates that while the RGB-based approach works well under typical lighting circumstances, it performs noticeably worse in low light. under contrast, the depth-based approach yields more consistent findings under similar circumstances because it is less susceptible to variations in lighting.

The suggested approach produces better outcomes under all of the experiment's settings. The benefit of integrating the two modalities is especially demonstrated by the improvement seen in low light. While depth information offers structural clues that assist the model stay dependable when the quality of one modality is compromised, RGB images offer helpful visual detail. The model is able to function more consistently in a changing environment thanks to this complementary knowledge. Overall, the findings demonstrate the efficacy of the suggested fusion technique and imply that it may be helpful in real-world situations where environmental parameters like lighting are not always under control.

IV. CONCLUSIONS

In this study, we presented a multimodal image segmentation approach that combines RGB and depth information. The experimental results, including both the visual outputs and quantitative evaluation, show that using the two modalities together can improve segmentation performance compared with using either modality alone. RGB images provide detailed information about object appearance, while depth maps contribute structural information about the scene. Combining these complementary features leads to more consistent segmentation results and clearer object boundaries. The comparative experiments and ablation studies further indicate that multimodal inputs can provide an advantage over single-modal approaches. However, the improvements is relat1ively modest and depends on the complexity of the scene and the characteristics of the dataset. The current method uses a relatively simple fusion strategy, but the results suggest that more advanced deep learning techniques, such as trainable cross-attention mechanisms, could further improve the performance. Future work will therefore investigate more sophisticated fusion methods and evaluate the approach on larger and more challenging datasets to assess its scalability and generalizability.

REFERENCES

[1] He, Q., Wu, M., Zhang, P., et al., “Multimodal image segmentation with dynamic adaptive window and cross-scale fusion”, Applied Sciences, Vol. 15, No. 19, 2025.
[2] Zhang, M., Zhang, Y., Liu, S., et al., “Dual-attention transformer-based hybrid network for multi-modal medical image segmentation”, Scientific Reports, Vol. 14, 2024.
[3] Qin, F., Liang, Y., Yang, C., et al., “Medical image segmentation network based on multi-scale cross-attention and Wavelet Transform”, Journal of King Saud University- Computer and Information Sciences, Vol. 37, No. 97, 2025.
[4] Seungik, L., et al., “CrossFormer: Cross-guided attention for multi-modal object detection”, Pattern Recognition Letters, Vol. 179, pp. 144-150, 2024.
[5] Song, K., Zhang, Y., Bao, Y., et al., “Self-enhanced mixed attention network for multi-modal image segmentation”, Sensors, Vol. 23, No. 14, 2023.
[6] Simonyan & Zisserman, 2015 – Very Deep Convolutional Networks (VGG).
[7] Vaswani et al., Attention is All You Need, 2017.
[8] U-Net: Convolutional Networks for Biomedical Image Segmentation.
[9] Goodfellow et al., “Deep Learning”, 2016.
[10] Website: NYU Depth V2 Dataset.
[11] Website: SUN RGB-D Dataset.
[12] Fully Convolutional Networks for Semantic Segmentation.
[13] V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation.
[14] Nahian Siddique et al., “U-Net and its variants for medical image segmentation: theory and applications”, IEEE Access, vol. 9, pp. 82031–82057, 2021.
[15] Liang-Chieh Chen et al., “Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (DeepLabV3+)”, ECCV, 2018.
[16] V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation”, IEEE TPAMI, 2017.
[17] C. Hazirbas et al., “FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture”, ACCV, 2016.

Noor Safaa received her B. Sc., and M. Sc., degrees in Electrical Engineering in 2006 and 2012, respectively, from Al-Mustansiriyah University, Baghdad, Iraq. She is currently a Lecturer in the Department of Electrical Engineering, Al-Mustansiriyah University, Baghdad, Iraq, and a Ph. D. candidate in Electrical Engineering at the University of Technology, Baghdad. Her research interest is fault-tolerant control, fault detection and estimation.