Journal of Information Systems Engineering and Management
2025, 10(17s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Novel Uncertainty-Aware Evidential Multimodal Deep Learning for RGB-D Household Object Recognition

Smita Gour [1*], Pushpa B Patil [2]

[1] *Department of Computer Science and Engineering, Basaveshwar Engineering College, Bagalkot-587102, Karnataka, India*
[2] *Department of Computer Science and Engineering (Data Science), BLDEA's V P Dr P G Halakatti College of Engineering & Technology, Vijayapur-586102, Karnataka, India*
[1] *smita.gour@gmail.com*; [2] pushpapatil2008@gmail.com;*
** corresponding author*

ARTICLE INFO

Received: 06 Dec 2024
Revised: 29 Jan 2025
Accepted: 12 Feb 2025

ABSTRACT

RGB-D house hold object recognition is essential for robotic perception, enabling accurate object identification by leveraging both visual (RGB) and depth information. However, traditional deep learning models struggle with sensor noise, occlusions, and overconfident misclassifications. To address this, we propose an Evidential Multimodal Deep Learning (EMDL) framework, integrating Evidential Deep Learning (EDL) with CNN (Convolutional Neural Network) and attention based feature fusion. Our model extracts features using CNNs for RGB and depth, and then fuses them through a cross-attention mechanism, allowing adaptive weighting of modalities based on uncertainty. Instead of softmax classifiers, Dirichlet-based evidential output layer has been used. It quantifies both classification confidence and epistemic uncertainty, improving robustness. Evaluations on the Washington RGB-D dataset demonstrate superior performance in classification accuracy, noise handling, and domain generalization compared to baseline models. Accuracy of 92.2% is reached with this novel approach considering 10-fold cross validation method. By enhancing uncertainty-aware decision-making, our approach ensures safer and more reliable robotic perception, making it suitable for real-world applications like grasping, manipulation, and autonomous navigation.

Keywords: Object Recognition, Uncertainty Aware, Evidential Deep Learning, Cross-Modal Attention based Fusion, RGB-D Dataset.

1. Introduction

In the domains of robotics and computer vision, one of the basic issues is object recognition. The majority of object recognition techniques that have been proposed up to this point are based on RGB (Red Green Blue) images. However, the RGB image can only reflect the scene's color, lighting, and texture information since the depth information of the image is lost during the optical projection process from the 3D (Three Dimensional) space to the 2D (Two Dimensional) environment. This makes it challenging to apply RGB image-based object recognition methods in real-world situations since they are susceptible to external factors like illumination and a complex background. [1–5].

Since the introduction of low-cost RGB-D (Red Green Blue-Depth) sensors such as Microsoft Kinect and Intel RealSense [6,7], the RGB-D sample has been used extensively in medical diagnostics, video surveillance, intelligence robotics, and scene analysis and understanding . Both color and depth images can be simultaneously captured by the RGB-D sensor. Information about color and appearance is contained in the RGB image, while information about the distance between the household item and the RGB-D sensor is contained in the depth image. The RGB-D image contains more useful information for household object recognition than an RGB image since it can reveal more details about the object's 3D geometry structure. The depth image is also resistant to changes in lighting and color. It has been demonstrated that the RGB-D image-based approach to household object recognition outperforms the RGB image-based approach. Thus, the study of the multi-modal object detection approach based on RGB-D images has gained increasing attention in recent years [8–10]. Existing RGB-D image-based object recognition techniques can be categorized into two groups based on the types of features: learnt feature-based techniques and hand-crafted feature-based techniques. Hand-crafted features such as spin images [13,14], scale-invariant feature transform (SIFT) [11], and speeded up robust features (SURF) [12] are used to describe the RGB and depth images for the first category. These features are then input into classifiers such SVMs

(Support Vector Machines) for classification. The chosen hand-crafted features have an impact on this type of method's performance. In addition to not being able to capture all of the valuable discriminative information of various object classes, the hand-crafted features frequently require human tuning for various scenarios. The classifiers are then used for the classification after the RGB and depth images are used to learn the features for the second category. Even while this method is more effective, it still does not make full use of the valuable information that RGB-D saples contain. For recognition, most existing algorithms usually learn the RGB and depth images separately and then simply combine the two feature types [15,16]. Therefore, how to properly leverage the link between the RGB data and the depth feature remains one of the primary challenges in the field of RGB-D object recognition.

Deep learning has gained a lot of popularity recently and has been effectively used in household RGB-D object detection. Socher et al. proposed a model based on a combination of Convolutional Neural Network (CNN) and Recursive Neural Networks (RNN) to learn features and categorize RGB-D images [17]. The RNN layer generates higher level features and the CNN layer extracts lower level characteristics. For RGB-D object recognition, Rahman et al. suggested deep neural network framework based on three cascaded multi-modal CNNs such as RGB, color and surface normal [18]. Tang et al. introduced multi-view convolutional neural networks based on canonical correlation analysis (CCA) for RGBD object recognition, which can successfully find the relationships between various viewpoints of the same shaped model [19]. Recent advancements in deep learning and multimodal fusion have significantly improved RGB-D object recognition. With the rise of Convolutional Neural Networks (CNNs) and Transformers, researchers developed deep learning architectures capable of extracting more robust and hierarchical representations from RGB and depth data. Two-stream CNNs [20] process each modality separately before fusion, whereas attention-based Transformer models [21] enable better cross-modal interactions. Despite these advances, deep learning models still struggle with uncertainty, often making overconfident and unreliable predictions when faced with noisy or out-of-distribution data.

Evidential Deep Learning (EDL) is a framework that integrates uncertainty estimation into deep learning models by leveraging evidence theory. Instead of treating model outputs as deterministic probabilities, EDL represents them as belief distributions (Dirichlet distributions), allowing the model to express confidence in its predictions. Evidential Deep Learning (EDL) has emerged as a promising approach [22, 23] to enhance RGB-D object recognition by incorporating uncertainty quantification into deep learning models. Traditional deep networks, particularly those based on CNNs and Transformers, rely on softmax probabilities for classification, which can lead to overconfident predictions even in ambiguous or noisy scenarios. EDL addresses this limitation by leveraging Dempster-Shafer evidence theory, which models both model uncertainty and data uncertainty through Dirichlet distributions. This is particularly beneficial for RGB-D object recognition, where sensor noise, occlusions, and modality imbalances often degrade performance. Recent works [24,25] have explored EDL-integrated multimodal fusion frameworks, where the model dynamically adjusts feature contributions from different streams based on their estimated uncertainty. Such uncertainty-aware learning not only improves robustness but also enables safer decision-making in robotic perception systems, ensuring that the model can defer predictions or request additional data when confidence is low. By integrating EDL with CNN-Transformer-based fusion mechanisms, modern RGB-D recognition systems can achieve more reliable, interpretable, and adaptive performance in real-world applications. This work introduces a novel Evidential Multimodal Deep Learning (EMDL) framework for RGB-D household object recognition, incorporating uncertainty-aware feature fusion and classification. Our contributions are twofold:

1. Uncertainty-Guided Feature Fusion: We propose a novel cross-modal attention based fusion strategy that dynamically adjusts RGB and depth contributions based on evidential uncertainty scores.
2. Robust and Interpretable Predictions: By incorporating EDL with Dirchlet Distribution-based uncertainty estimation, our model not only enhances recognition performance but also provides confidence scores that aid in decision-making, especially in robotic and autonomous applications.

By leveraging uncertainty-aware multimodal learning, our approach significantly improves robustness, adaptability, and interpretability in RGB-D object recognition, making it suitable for real-world robotic perception systems.

2. Proposed Methodology

The Evidential Deep Learning (EDL) framework shown in Fig. 1 for RGB-D household object recognition follows a structured pipeline that incorporates uncertainty estimation into the classification process. Prior to being rescaled

to their normalized size during the training phase, the RGB and depth images undergo preprocessing to minimize noise. Then, using the HHA encoding method [26–30], we calculate the depth image's three channels. The HHA code stands for the horizontal disparity, height above ground, and angle with gravity. Initially, feature extraction is performed separately on RGB and depth data of household object images using deep neural networks such as CNNs capturing both appearance-based (RGB) and geometric (depth) features. These extracted features are then combined using a multimodal fusion strategy, which is an attention-driven, ensuring that the network effectively leverages both modalities. The fused representation is processed by an evidential output layer, with a Dirichlet distribution-based evidence model. Instead of directly outputting class probabilities, the model predicts evidence scores, representing the level of belief assigned to each class. The uncertainty quantification is computed based on the total evidence, where lower evidence across all classes signifies higher uncertainty. Training is guided by a modified loss function, which optimizes the model to increase confidence for correct classifications while discouraging overconfident incorrect predictions. This uncertainty-aware approach makes EDL particularly useful in real-world RGB-D applications, where sensor noise, occlusions, and missing depth data can lead to unreliable predictions. The detailed discussion of each block is discussed in this section.

2.1 Preprocessing of RGB-D Image

In this study, the Washington RGB-D object dataset [31] is used to evaluate the suggested RGB-D object recognition algorithm. From Point Gray Research, a firewire camera and RGBD camera were used to gather the dataset, which consists of 300 home items arranged into 51 categories. Some of those household object categories from the Washington RGB-D object collection are shown in Fig. 2. To capture image of each object, the cameras were located at three different heights and in three distinct directions. Approximately 600 RGB-D samples per object make up the total of 207,920 images.

The RGB and depth images given as input to the system shown in Fig. 3a-c are initially scaled to 227 × 227 in order to satisfy the needs of the two CNNs. The two CNNs employ the fundamental AlexNet architecture. The simplest method is to directly resize the samples to the desired size. The direct approach, however, may alter the object's original geometric structure and ratio shown in Fig. 3d-e, which will affect the identification performance. Therefore, we applied the scaling processing technique suggested in [19].

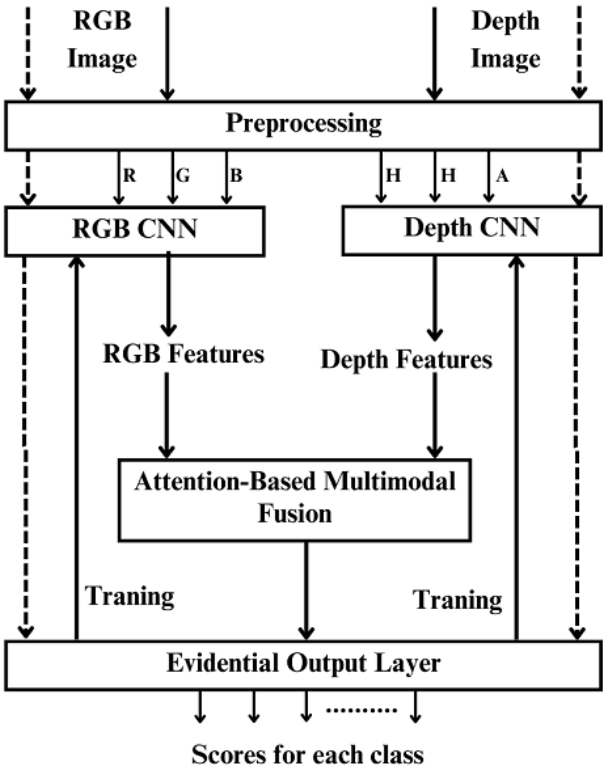

Figure 1. Proposed Methodology

Initially, we adjusted the original image to make its long side 227 pixels long. To create a square image, we then enlarged the resized picture along its short side. The resized image should be situated in the center of the expansion-scaled image, with both sides of the image expansion being equal. Black pixels are added to the samples to make them larger. The scaled pictures are displayed in Fig. 3g–i. Figure 3 shows that the scaled photographs successfully maintain the objects' shape information when compared to the resized ones.

The image from R, G and B channels obtained from scaled RGB image are the three inputs to RGB CNN First, we use the median filters to decrease noise and fill in the holes in the scaled depth image. A HHA encoding method is used to get H, H, and A channels which are the three input images for the depth. Several RGB-D image-based efforts have successfully exploited the HHA representation, which may store the geocentric pose features that highlight complementing discontinuities in the image [26–30].

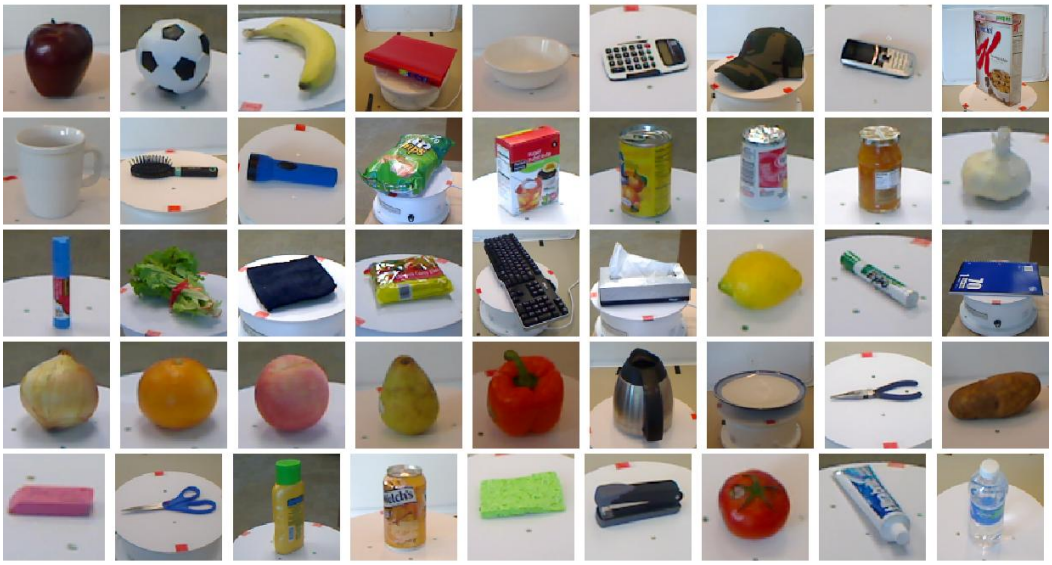

Figure 2: Samples from different categories in Washington RGBD dataset

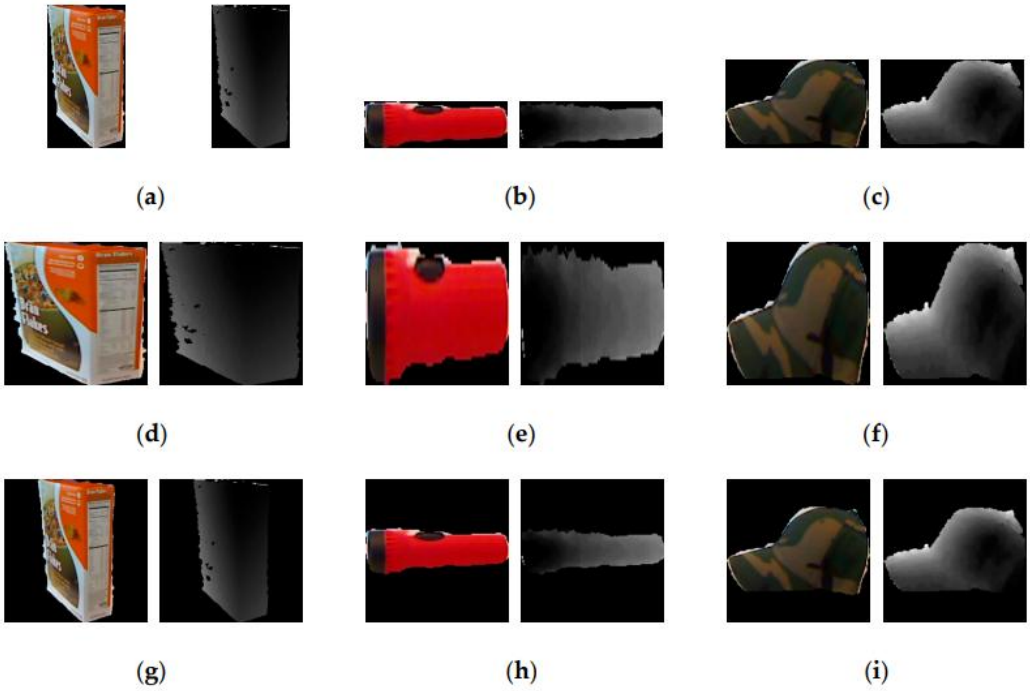

Figure 3: samples of RGB and corresponding Depth images from Washington RGBD dataset (a-c), The direct Resized saples (d-f), Rescaled images (g-i)

2.2 The architecture of multimodal CNN

The multi-modal convolution network that has been proposed for the task of household object recognition is shown in Fig. 4 aims to extract RGB and depth feature attributes from the household objects. On it, there are two branches. Each branch is a CNN with the same architecture as AlexNet [33]. The inputs of the first branch are the three channels of the RGB images, and the inputs of the second branch are the HHA encoding results of the depth images. The AlexNet consists of three completely linked layers, five convolutional layers, and a final 1000-way softmax. It has over 60 million features and 650,000 neurons. The first, second, and fifth convolutional layers are followed, successively, by max-pooling layers. For all convolutional and fully-connected layers, the activation function is the rectified linear unit (ReLU). The training of the proposed network is divided into two stages. In the first stage, the RGB and depth characteristics are learned separately using the relevant CNNs. In the second stage, the multi-modal network is fine-tuned using the RGB and depth pictures. The optimization method considers both the discriminative information of each modality and the correlation information between two modalities. The RGB and Depth CNN features obtained from this modal will fused using cross-modal attention based fusion method which is discussed in detail in next section.

2.3 Cross-Modal Attention based fusion

An expansion of the conventional attention mechanism, cross-modal (or multi-modal) attention allows for the interaction of many data modalities, such as RGB and depth data, or images and text, inside a single framework. By concentrating on the most instructive aspects of each modality when processing and combining the data, this technique enables a model to discover the complementary relationships between modalities.

In attention mechanisms (especially in Transformer architectures), each feature or token is transformed into three components: Queries (Q), Keys (K), Values (V). In cross-modal attention, these components can come from different modalities. For example: Use the RGB features to form queries. Use the depth features to form keys and values.
The attention weights are computed by taking the dot product between the queries and keys, scaling by the square root of the key dimension, and then applying the softmax function. Mathematically, for each query qi and for each kj , the attention weight aij is given by equation (1). In matrix form equation (1) can be given by equation (2)

$$w_{ij} = \frac{exp^{\frac{q_i \cdot k_j}{\sqrt{d_k}}}}{\sum_{j=1}^{N_B} exp^{\left(\frac{q_i \cdot k_j}{\sqrt{d_k}}\right)}} \quad (1)$$

$$M = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \; with\; M \in R^{N_A \times N_B} \quad (2)$$

Here, the queries from one modality interact with keys from another, generating weights that dictate how much each depth feature (value) should contribute to the final fused representation for each RGB feature.

Once the attention weights are computed, the model aggregates the information from the second modality (depth) and integrates it with the first modality (RGB). This can be done by either Concatenation: Combining the attended features with the original features or Addition: Adding the attended features to the original features. Once you have the attention weights M, you compute the cross-modal fused representation by weighting the values V.

$$Z = MV \; with\; Z \in R^{N_A \times d_v} \quad (3)$$

Here, each row zi in Z is a weighted sum of the values from modality B, where the weights w_{ij} indicate the relevance of each value vj to the query qi. These fused features are sent to Evidential output layer from which the training process begins.

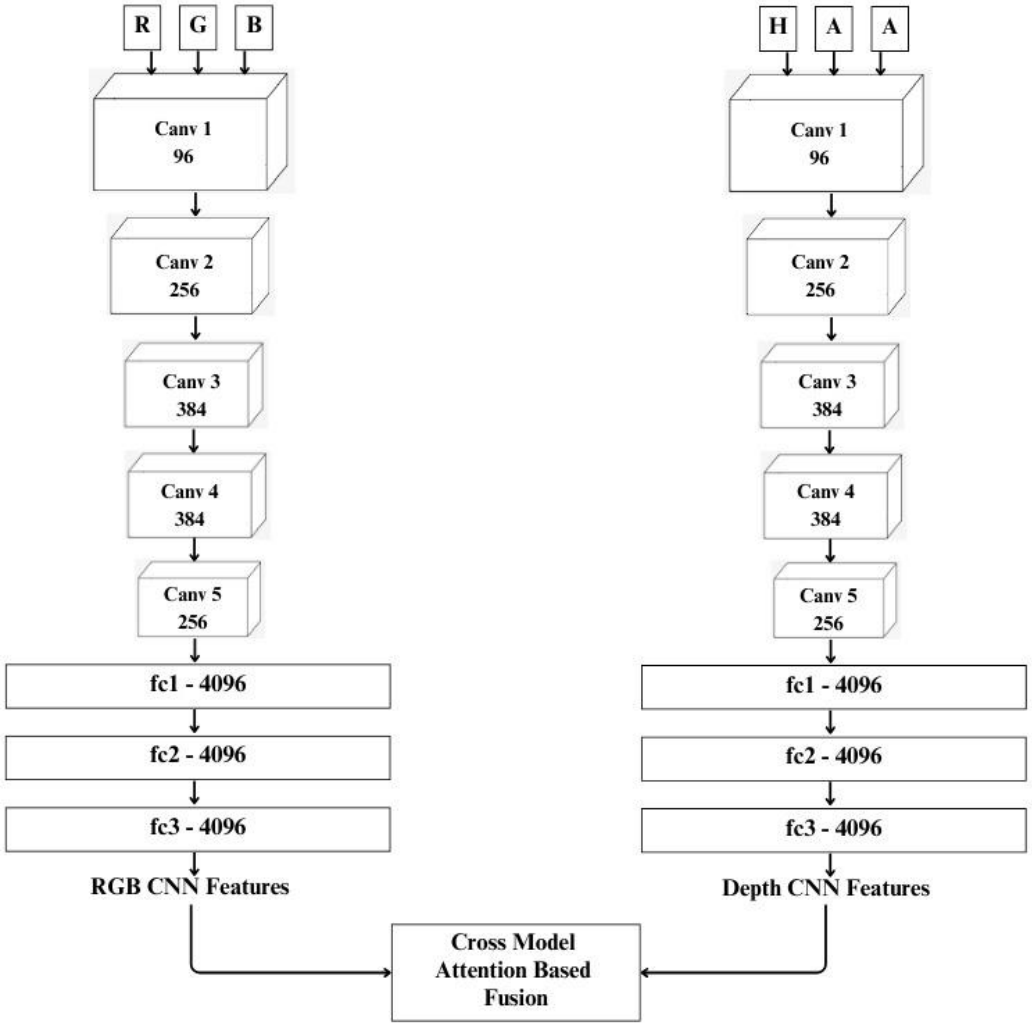

Figure 4: Architecture of Multi-modal CNN

2.4 Evidential output layer

An evidential output layer is designed to produce not only class predictions but also an associated measure of uncertainty by outputting parameters of a probability distribution—typically a Dirichlet distribution in classification tasks. This approach, often referred to as "evidential deep learning," moves beyond point estimates (like softmax probabilities) and instead characterizes uncertainty directly in the Traditional neural network classifiers typically output a probability vector using a softmax layer, which represents the network's confidence in each class. However, these probabilities can be overconfident, even when the input is ambiguous or out-of-distribution. Evidential deep learning addresses this by:

Modeling Uncertainty Explicitly: Instead of outputting a single probability, the network outputs parameters that define a Dirichlet distribution over class probabilities.

Quantifying Evidence: The network learns to predict "evidence" for each class. The total amount of evidence influences both the predicted probabilities and the degree of uncertainty.

2.4.1 Dirichlet Distribution

The Dirichlet distribution, a probability distribution over a set of probabilities, is widely used in multi-class classification problems and uncertainty modeling. Its ability to generalize the Beta distribution to multiple categories makes it useful in Evidential Deep Learning (EDL). To help measure classification uncertainty, the model predicts a range of probabilities instead of assigning a single probability to each class.

For the classification task with K classes, Dirichlet Distribution is defined as in equation (4).

$$D(p \mid \alpha) = \frac{1}{B(\alpha)} \prod_{K=1}^{K} p_k^{a_{k-1}} \quad (4)$$

Where,

$P=(p_1,p_2, p_3.....,p_k)$ is a probability vector ($p_k \geq 0$ and $\sum_{k=1}^{K} p_k = 1$)

$\alpha = (\alpha_1, \alpha_2, \ldots . \alpha_K)$ are concentration parameters.

$B(\alpha)$ is the multivariate beta function ensuring proper normalization.

High value of α_K indicates strong evidence of the class k. Low total evidence (i.e. $\sum_{k=1}^{K} a_k$) implies high uncertainty while high total evidence indicates confident prediction.

2.4.2 Loss Functions

Training an evidential output layer typically involves specialized loss functions that encourage both correct predictions and appropriate uncertainty estimation. The Evidential Loss Function for household object recognition is derived from the log-likelihood of the Dirichlet distribution computed using equation (4) and consists of two terms:

Data Fit Term: Encourages the expected probabilities to match the true labels. The Data Fit Term is a Mean Squared Error (MSE) l**oss** between the predicted belief p_k and the one-hot encoded true label y_k.

Uncertainty Regularization: Penalizes overconfident predictions when there is insufficient evidence. A common choice is to include a Kullback–Leibler (KL) divergence term between the predicted Dirichlet distribution $D(\alpha)$ and uniform Dirichlet D(1). By Combining these two terms we define our Evidential Loss function as shown in equation (5). The computed loss is going to be used for training a CNN models by Stochastic Gradient Descent (SGD) with back propagation method discussed in next section.

$$L(\alpha, y) = \sum_{k=1}^{K} (y_k - p_k)^2 + \lambda KL\big(D(\alpha) \parallel D(1)\big) \quad (5)$$

Where,

$\sum_{k=1}^{K} (y_k - p_k)^2$ indicate data-fit,

y is the one-hot encoded ground truth,

λ is the hyperparameter balancing the terms

2.5 Training

After obtaining fused featured of household objects and loss, next is to train the RGB CNN and the depth CNN, respectively, using the Stochastic Gradient Descent (SGD) method with back-propagation. It combines two key ideas:

Stochastic Gradient Descent: An iterative optimization method that updates the model parameters using the gradient of the loss function computed using equation (5) on a small (often random) subset of the training data (called a mini-batch). Rather than computing gradient over the entire dataset, SGD uses mini-batch of m samples at each iteration. Let B $\subset$ {1,2.....N} be the indices of samples in mini-batch. The mini-batch loss is given by equation (6). SGD updates its parameter using equation (7).

$$L_B(\theta) = \frac{1}{m} \sum_{i \in B} L\big(f\big(x^{(i)}; \theta\big), y^{(i)}\big) \quad (6)$$

$$\theta \leftarrow \theta - \eta \nabla_\theta L_B(\theta) \quad (7)$$

Where:

η is learning rate

$\nabla_\theta L_B(\theta)$ is the gradient of mini-batch loss with respect to θ

Back-propagation: It is a method for efficiently computing gradients of the loss with respect to the network's parameters using the chain rule of calculus. Suppose you have a neural network with parameter θ and a training dataset $_{i=1}^{N}\{(x^{(i)}, y^{(i)})\}$. The goal is to minimize the $L_B(\theta)$ that measures the discrepancy between the network's predictions and the true labels. The overall objective is depicted in equation (8).

$$\theta^* = argmin \frac{1}{N} \sum_{i=1}^{N} L(f(x^i; \theta), y^{(i)}) \quad (8)$$

Where $f(x^i; \theta)$ denotes the network's output given input x (i)

3. Experimentation

We trained the suggested multi-modal network before implementing it. The three steps of the training stage were as follows: (1) rescale the RGB and depth images to 227 × 227; (2) train the RGB and depth CNNs, respectively; (3) train the multi-modal network; The trained AlexNet on the ImageNet dataset was used to initialize the RGB CNN and the depth CNN. The pretrained network was used to initialize the CNNs' weights. With the performance improvement, the learning rate was adjusted from its initial setting of 0.01 to 0.001. 128 was chosen as the batch size, N.

When testing a neural network with an evidential output layer, the goal is not only to obtain class predictions but also to quantify the associated uncertainty. During testing, each input sample is processed by the network just as in training in three stages.

Feature Extraction: The trained CNN features (both RGB and Depth) are extracted from the input household image (e.g., an image or multimodal data).

Evidence Generation: The evidential output layer takes the feature representation and computes evidence for each class. This is usually done via a linear transformation followed by a non-negative activation (e.g., ReLU or softplus):

Constructing Dirichlet Parameters: Using the Dirichlet parameters, we derive both the predictive probabilities and an uncertainty measure: where a larger value indicates higher uncertainty (i.e., less total evidence). Alternatively, you might compute the variance of the Dirichlet distribution or other uncertainty metrics that capture how peaked or flat the predicted distribution is.

Ten recognition accuracies of our suggested approach employing ten cross-validation splits are shown in Table 1. Table 1 shows that the variance of the 10 recognition accuracies is very low and that they are comparable to their mean value. Therefore, our suggested approach is reliable for many splits. We directly provide the mean and standard deviation values of ten recognition findings in the experimental data that follow in this research.

Table 1. Result of 10-fold cross validation

1	2	3	4	5	6	7	8	9	10	Mean	Var
90.5	92.7	91.8	93.1	89.8	90.2	92.2	91.2	91.8	90.8	91.45	1.38

3.1 Comparison with Different Baselines

We experimented with RGB and depth images using the following distinct baselines as [32] in order to verify the efficacy of the proposed Evidential Multimodal Deep Learning (EMDL) framework

1. RGB CNN + Softmax: Added a softmax layer to the network's end for categorization after using the CNN to learn RGB characteristics.
2. Depth CNN+ Softmax: Added a softmax layer to the end of the network for classification after using the CNN to learn depth features.

3. RGB-D CNNs+Softmax: Initially, the RGB images are used to train the RGB CNN, and then the depth CNN. Then, transmitted the connected RGB and depth features to the softmax layer for object recognition.

4. RGBD CNNs+Attention-Based-Fusion+Softmax: To fuse RGB and Depth CNN features cross-modal attension based technique is used and transmitted to softmax layer.
5. RGBD CNNs + attention –based fusion + EDL-Dirichlet (Proposed): Instead of Softmax, evidential output layer with Dirchelet Distribution is used.

The Washington RGB-D object dataset is used to evaluate all the first three baseline methods and a proposed method. The recognition accuracy is displayed in Table 2, and the best method's score is bolded. The third baseline method RGB-D image-based technique outperform single modality-based (first 2) techniques by a wide margin. This is due to a certain complementary between the identity information found in the depth image and the RGB image.

Still the recognition performance can be enhanced by applying some feature fusion techniques on multi-modal data. Cross-modal attention based fusion technique is used and fused features are transmitted to softmax layer for prediction. The accuracy is improved than those without fusion techniques. At last the limitation of softmax was identified that is softmax probabilities for classification can lead to overconfident predictions even in ambiguous or noisy scenarios. Hence the accuracy still can be improved by replacing this with evidential based output layer with Dirchlet Distribution.

Therefore, we may say that our Evidential Multimodal Deep Learning with cross-modal attention based fusion approach works better than other baseline approach. Our suggested multi-modal learning approach allows us to extract more effective discriminative features from the RGB-D images. Nonetheless, several classes are frequently misclassified. The primary cause of the misclassifications is that samples from many classes share a similar color and form.

Table 2. Comparision of Different Baselines on the Washington RGBD Object Dataset

Method	Accuracy
RGB-CNN+Softmax	85.7±2.3
Depth-CNN+Softmax	81.3±2.2
RGBD-CNNs + Softmax	90.2± 1.8
RGBD CNNs + attention –based fusion+Softmax	88.9 ±1.9
RGBD CNNs + attention –based fusion + EDL-Dirichlet (Proposed Approach)	**92.2 ±1.3**

3.2 Comparasion with State-of-the-Art Methods

We contrasted our suggested method's recognition accuracy with that of the 9 most advanced techniques listed below: (1) Linear SVM [31]: Texton histograms and color histograms are utilized for RGB feature extraction, while spin images and SIFT descriptors are employed for depth feature extraction. For classification, the linear support vector machine is employed. (2) Nonlinear SVM [31]: The Gaussian kernel SVM is used for classification, and the selected features are same to those used in the "Linear SVM" method. (3) HKDES [34]: RGB and depth features are extracted using a mixture of hierarchical kernel descriptors, and classification is done using a linear SVM. (4) Kernel Descriptor [35]: Linear SVM is utilized for classification, and a collection of kernel descriptors is employed for feature extraction. (5) CNN-RNN [36]: To learn features and categorize RGB-D pictures, a model comprising CNN and RNN is utilized. (6) The VGGNet+3D CNN+ VGG3D [34] approach, which employs the 16-layer VGGNet to learn features from RGB images, obtained the highest recognition accuracy for RGB image-based object recognition. The deep network's scale is greater than that of the CNN that we employed in our suggested method, and its recognition performance is marginally superior. (7) and (8) RGB CNN-SVM and Depth CNN-SVM [32]: It used a CNN to learn RGB and depth properties properties respectively, and then added a softmax layer at the end of the network for categorization. (9) Multimodal CNNs+DS [32]: Here first they used multi-modal learning for feature extraction. Second, in order to successfully fuse the classification outcomes of the two SVMs, they created the DS evidence theory-based decision fusion scheme. Our suggested approach outperformed the majority of the state-of-the-art techniques for RGB-D object recognition in general.

Table 2. Comparison with state-of-the-art methods on the Washington RGBD object dataset

Method	Accuracy		
	RGB	**Depth**	**RGBD**
Linear SVM [21]	74.3±3.3	53.1±1.7	81.9±2.8
kSVM [21]	74.5±3.1	64.7±2.2	83.8±3.5
HKDES [34]	76.1±2.2	75.7±2.6	84.1±2.2
Kernel Descriptor [22]	77.7±1.9	78.8±2.7	86.2±2.1
CNN-RNN [30]	80.8±4.2	78.9±3.8	86.8±3.3
VGGnet+3DCNN+VGG3D [37]	88.9±2.1	78.4±2.4	91.8±1.4
RGB-CNN+SVM [32]	87.5±2.1	-	-
Depth-CNN+SVM [32]	-	84.8± 2.0	-
Multimodal CNNs+DS [32]	87.5±2.1	84.8±2.0	91.8±1.4
Proposed	**88.9±1.1**	**84.9±1.0**	**92.2±1.3**

3.3 Experimentation with Real Time Household Object Images

Some of the real time household objects like Bottle, Cup, Knife, Spoon, Apple, Cell phone, Clock, Scissor, Tooth brush, Fork are considered to evaluate the system performance. In this evaluation step, 50 samples of each are tested, and the confusion matrix for the same is depicted in Fig. 5. An average accuracy of 90% is obtained.

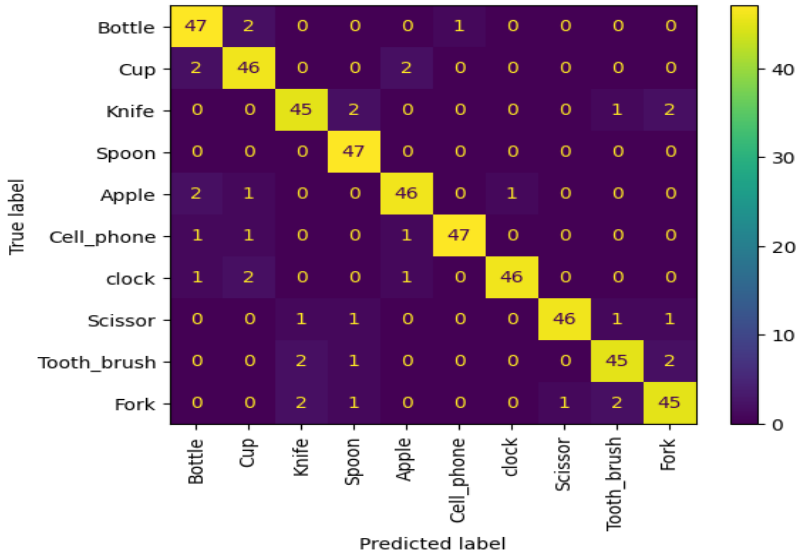

Figure 5: Confusion Matrix

4. Conclusion

In this work, a cross-modal attention-based fusion strategy is included into a unique uncertainty-aware Evidential Multimodal Deep Learning (EMDL) framework for RGB-D household object detection. The suggested model improves classification accuracy and offers a trustworthy measure of uncertainty by utilizing Evidential Deep Learning (EDL) and the Dirichlet distribution. This is crucial for real-world applications that demand sound decision-making. A distribution across categorical probabilities is modeled by the Dirichlet distribution, which enables the model to capture differing degrees of belief in various classes. The network predicts Dirichlet parameters rather than a single probability value for each class, allowing for uncertainty-aware classification and avoiding overconfident wrong predictions. Additionally, by dynamically highlighting informative characteristics

while suppressing less pertinent ones, the cross-modal attention mechanism guarantees the successful merger of RGB and Depth modalities. Experiments demonstrate that the proposed approach works better in terms of classification accuracy (92.2%) and uncertainty estimates than conventional multimodal strategies. The ability to quantify uncertainty allows the model to recognize ambiguous or out-of-distribution occurrences, enhancing its robustness in assistive AI and real-world robotic vision applications. Future research could look into extending this approach to open-set recognition, few-shot learning, and continual learning in order to further improve adaption and generalization.

References

[1] Wong, S.C.; Stamatescu, V.; Gatt, A.; Kearney, D.; Lee, I.; McDonnell, M.D. Track Everything: Limiting Prior Knowledge in Online Multi-Object Recognition. IEEE Trans. Image Process. 2017, 26, 4669–4683. [CrossRef] [PubMed]

[2] Aldoma, A.; Tombari, F.; Stefano, L.D.; Vincze, M. A Global Hypothesis Verification Framework for 3D Object Recognition in Clutter. IEEE Trans. Pattern Anal. Mach. Intell. 2016, 38, 1383–1396. [CrossRef] [PubMed]

[3] Oliveira, F.F.; Souza, A.A.F.; Fernandes, M.A.C.; Gomes, R.B.; Goncalves, L.M.G. Efficient 3D Objects Recognition Using Multifoveated Point Clouds. Sensors 2018, 18, 2302. [CrossRef] [PubMed]

[4] Chuang, M.C.; Hwang, J.N.; Williams, K. A Feature Learning and Object Recognition Framework for Underwater Fish Images. IEEE Trans. Image Process. 2016, 25, 1862–1872. [CrossRef] [PubMed]

[5] Gandarias, J.M.; Gómez-de-Gabriel, J.M.; García-Cerezo, A.J. Enhancing Perception with Tactile Object Recognition in Adaptive Grippers for Human–Robot Interaction. Sensors 2018, 18, 692. [CrossRef] [PubMed]

[6] Sanchez-Riera, J.; Hua, K.L.; Hsiao, Y.S.; Lim, T.; Hidayati, S.C.; Cheng, W.H. A comparative study of data fusion for RGB-D based visual recognition. Pattern Recognit. Lett. 2016, 73, 1–16. [CrossRef]

[7] Ren, L.; Lu, J.; Feng, J.; Zhou, J. Multi-modal uniform deep learning for RGB-D person re-identification. Pettern Recognit. 2017, 72, 446–457. [CrossRef]

[8] Xu, X.; Li, Y.; Wu, G.; Luo, J. Multi-modal deep feature learning for RGB-D object detection. Pattern Recognit. 2017, 72, 300–313. [CrossRef]

[9] Bai, J.; Wu, Y.; Zhang, J.; Chen, F. Subset based deep learning for RGB-D object recognition. Neurocomputing 2015, 165, 280–292. [CrossRef]

[10] Li, X.; Fang, M.; Zhang, J.J.; Wu, J. Learning coupled classifiers with RGB images for RGB-D object recognition. Pattern Recognit. 2017, 61, 433–446.

[11] Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vis. 2004, 60, 91–110.

[12] Bay, H.; Tuytelaars, T.; Gool, L.V. SURF: Speeded up Robust Features. In Proceedings of the European Conference on Computer Vision, Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; Volume 3951, pp. 404–417.

[13] Johnson, A.E.; Hebert, M. Surface matching for object recognition in complex three-dimensional scenes. Image Vis. Comput. 1998, 16, 635–651. [CrossRef]

[14] Johnson, A.E.; Hebert, M. Using Spin Images for Efficient Object Recognition in Cluttered 3D Scenes. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 21, 433–449. [CrossRef]

[15] Schwarz, M.; Schulz, H.; Behnke, S. RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features. In Proceedings of the IEEE International Conference on Robotics and Automation, Seattle, WA, USA, 26–30 May 2015; pp. 1329–1335.

[16] Cheng, Y.; Zhao, X.; Huang, K.; Tan, T. Semi-supervised learning and feature evaluation for RGB-D object recognition. Comput. Vis. Image Underst. 2015, 139, 149–160. [CrossRef]

[17] Socher, R.; Huval, B.; Bhat, B.; Manning, C.D.; Ng, A.Y. Convolutional-Recursive Deep Learning for 3D Object Classification. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 656–664.

[18] Rahman, M.M.; Tan, Y.; Xue, J.; Lu, K. RGB-D object recognition with multimodal deep convolutional neural networks. In Proceedings of the IEEE International Conference on Multimedia and Expo, Hong Kong, China, 10–14 July 2017; pp. 991–996.

[19] Tang, L.; Yang, Z.X.; Jia, K. Canonical Correlation Analysis Regularization: An Effective Deep Multi-View Learning Baseline for RGB-D Object Recognition. IEEE Trans. Cogn. Dev. Syst. 2018. [CrossRef]

[20] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. 2015. Multimodal deep learning for robust RGB-D object recognition. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE Press, 681–687. https://doi.org/10.1109/IROS.2015.7353446

[21] Y. Zhang, M. Yin, H. Wang and C. Hua, "Cross-Level Multi-Modal Features Learning With Transformer for RGB-D Object Recognition," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 33, no. 12, pp. 7121-7130, Dec. 2023, doi: 10.1109/TCSVT.2023.3275814.

[22] Gao, Junyu & Chen, Mengyuan & Xiang, Liangyu & Xu, Changsheng. A Comprehensive Survey on Evidential Deep Learning and Its Applications. (2024) 10.48550/arXiv.2409.04720.

[23] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18). Curran Associates Inc., Red Hook, NY, USA, 3183–3193.

[24] Juan Baz, Mikel Ferrero-Jaurrieta, Irene Díaz, Susana Montes, Gleb Beliakov, and Humberto Bustince. 2024. Probabilistic study of Induced Ordered Linear Fusion Operators for time series forecasting. Inf. Fusion 103, C (Mar 2024). https://doi.org/10.1016/j.inffus.2023.102093.

[25] Mendonça, Hildeberto & Lawson, Jean-Yves & Vybornova, Olga & Macq, Benoit & Vanderdonckt, Jean. (2009). A fusion framework for multimodal interactive applications. 161-168. 10.1145/1647314.1647344.

[26] Gupta, S.; Girshick, R.; Arbeláez, P.; Malik, J. Learning Rich Features from RGB-D Images for Object Detection and Segmentation. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 345–360.

[27] Gupta, S.; Arbeláez, P.; Malik, J. Perceptual Organization and Recognition of Indoor Scenes from RGB-D Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 564–571.

[28] Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D scene understanding benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576.

[29] Gupta, S.; Hoffman, J.; Malik, J. Cross Modal Distillation for Supervision Transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2827–2836.

[30] Song, S.; Xiao, J. Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 808–816

[31] Lai, K.; Bo, L.; Ren, X.; Fox, D. A large-scale hierarchical multi-view RGB-D object dataset. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 1817–1824

[32] Zeng, H., Yang, B., Wang, X., Liu, J., & Fu, D. (2019). RGB-D Object Recognition Using Multi-Modal Deep Neural Network and DS Evidence Theory. *Sensors*, *19*(3), 529. https://doi.org/10.3390/s19030529

[33] Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.

[34] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.

[35] Bo, L.; Ren, X.; Fox, D. Depth kernel descriptors for object recognition. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, USA, 25–30 September 2011; pp. 821–826

[36] Socher, R.; Huval, B.; Bhat, B.; Manning, C.D.; Ng, A.Y. Convolutional-Recursive Deep Learning for 3D Object Classification. In Proceedings of the International Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 656–664.

[37] Zia, S.; Yüksel, B.; Yüret, D.; Yemez, Y. RGB-D Object Recognition Using Deep Convolutional Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 887–894.