EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

NLP-Based Robust Object Detection and Recognition Through Multimodal Relation Graph Construction Using MPST AND DSWDCNN

Bo Feng*, Jiayi Yang, Jingyue Xue

College of Mathematics and Computer Science, Yan'an University, Yan'an, 716000, Shannxi, China

Abstract

INTRODUCTION: Nowadays, object detection and recognition play an important role in various applications such as surveillance, autonomous driving, robotics, and medical imaging. However, none of the traditional works focuses on analyzing the explicit relationships, logical dependencies, and semantic conflicts in text-rich or complex scenes, affecting object recognition accuracy.
OBJECTIVES: A Natural Language Processing (NLP)-based robust object detection and recognition framework is developed through multimodal relation graph construction using Minimum Persistence Spanning Tree (MPST) and Deep Swim Wishart Distribution Convolutional Neural Network (DSWDCNN).
METHODS: Initially, image with their corresponding captions is collected. Then, the image preprocessing and text preprocessing are done independently. From the preprocessed image, the object is detected using You Aspect-ratio Adaptive Anchors Only Look Once version-8 (YAAAOLOv8), followed by visualization. Meanwhile, from the preprocessed text, entity relations are identified. The multimodal relation graph is constructed using MPST. Further, the features from preprocessed text, relation graphs, detected objects, and visualized-images are extracted. Next, the multimodal analysis is carried out.
RESULTS: In the meantime, the word embedding is performed on the preprocessed texts. Finally, the object recognition is carried out using DSWDCNN.
CONCLUSION: The proposed framework achieves an object recognition accuracy of 98.8569%, demonstrating its effectiveness under weakly supervised conditions.

Keywords: Object Detection and Recognition, Deep Learning (DL), Natural Language Processing (NLP), Human Computer Interface (HCI) Applications, Multimodal Relation Graph Construction, Entity Relation Identification, and Artificial Intelligence (AI).

Received on 03 February 2026, accepted on 10 April 2026, published on 27 April 2026

doi: 10.4108/eetsis.11813

*Corresponding author. Email: fengbo1982@163.com

1. Introduction

In recent years, Artificial Intelligence (AI) helps to revolutionize multiple domains by enabling machines to learn from data, adapt to complex environments, and perform human intelligence tasks [1], [2]. Among the most influential branches of AI, computer vision and Natural Language Processing (NLP) have gained significant attention for their ability to allow systems to observe visual information and understand human language [3]. NLP enables machines to process, analyze, and generate human language in a meaningful way, allowing users to communicate with intelligent systems

using natural expressions instead of rigid commands [4] [5].

Also, NLP plays a major role in the development of object detection and recognition techniques. Object detection refers to the process of identifying and localizing multiple objects within an image or video stream by assigning class labels and bounding boxes to each detected instance [6], [7]. Further, the NLP in object detection connects visual perception with semantic understandings by interpreting user queries, generating descriptive captions, and providing meaningful explanations about the detected entries [8] [9]. Also, the AI-based techniques, such as Machine Learning (ML) and Deep Learning (DL) approaches in NLP enhance the performance of object detection and recognition [10].

Traditional ML methods rely on handcrafted features and shallow classifiers, which often struggle with complex linguistic structures and visual variability [11]. In contrast, the DL methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformer-based methods automatically learn hierarchical representations from large-scale data, enabling more accurate language understanding and visual-semantic alignment [12], [13]. However, they face several challenges like high computational complexity and training data requirement, when applied to real-world object detection tasks [14]. Additionally, some traditional works' 3performances are affected in terms of real-time performance, robustness to noise, and generalization across diverse environments [15]. Also, none of the traditional works analyzes the explicit relationships, logical dependencies, and semantic conflicts, especially in text-rich or complex scenes, leading to more false positives and negatives [16]. Therefore, an enhanced and robust NLP-based object detection and recognition model is proposed via constructing the multimodal relation graphs using MPST and DSWDCNN.

1.1. Research Overview

In this phase, the main aim and objectives of the proposed model and the limitations of the traditional object detection models are elaborated as follows,

Research Gap: None of the traditional works concentrates on analyzing explicit relationships, logical dependencies, and semantic conflicts, especially in text-rich or complex scenes, affecting object recognition accuracy. Despite the progress achieved by existing weakly supervised and oriented object detection approaches, significant limitations remain in handling spatial ambiguity, feature inconsistency, and noisy pseudo-label propagation under sparse annotations. In particular, the lack of robust multi-scale feature representation and the absence of adaptive learning mechanisms hinder reliable localization in complex remote sensing environments. To address these challenges, the proposed framework introduces two key components: the Multi-Phase Spatial Transformation (MPST) module, which enhances feature representation across varying spatial scales and orientations, and the Deep Spatial-Weighted Dynamic Convolutional Neural Network (DSWDCNN), which adaptively refines feature learning and improves robustness against noisy supervision. These components are designed to directly tackle the identified research gaps by improving spatial consistency, feature discrimination, and detection reliability under weak supervision settings.

Problem Statement: The limitations of the traditional object recognition works are listed as follows,

- ☑ Existing models struggle to effectively model interactions between visual and textual modalities, limiting their ability to achieve deep semantic understanding and generate accurate object recognition.
- ☑ Some conventional works fail to highlight a salient image region that leads to hallucinated objects and reduced interpretability.
- ☑ Most of the traditional works recognize the object from unprocessed images without relying on traditional image preprocessing steps, increasing false positives.

Significance and Scope of this Study: The proposed work develops a robust NLP-based object detection and recognition model through multimodal relation graph construction. Here, by using the proposed MPST, the explicit relationships, logical dependencies, and semantic conflicts among image and text are efficiently analyzed via ensuring feature fusion.

Research Aim and Objective: The main objective of the proposed work is to develop an enhanced and robust NLP-based object recognition model through multimodal relation graph construction. The objective of the proposed work is listed as follows,

- To design an intelligent weakly supervised framework for precise and reliable orientation-aware object detection in high-resolution remote sensing imagery.
- To evaluate the effectiveness of weak supervision on large-scale aerial object detection using the DOTA-v2 dataset.
- To develop a YOLO-OBB-based detection architecture for accurate rotation-aware localization under sparse annotations.
- To incorporate Bayesian inference for modeling predictive uncertainty and improving detection reliability.

The structure of the paper is depicted as follows, Section 2 indicates the literature survey, Section 3 elaborates the proposed methodology, Section 4 signifies the result and discussion, and Section 5

concludes the proposed work with future recommendations.

2. Literature Survey

[17] introduced an object detection model through an attention mechanism. Here, for multi-scale detection, the You Only Look Once version-5 (YOLOv5) was used. Further, the Convolutional Block Attention Module (CBAM) was applied to capture the channel and spatial features. As per the validation, this model improved the precision and recall by 4.52% and 1.18%, correspondingly. Yet, this model failed to capture the deep semantic understanding from both image and text, affecting the object recognition accuracy.

[18] presented a model for contextual object detection. Initially, the visual encoder extracted high-level image features and produced both local and fully visual tokens. Further, the pre-trained Large Language Model (LLM) was applied to decode multimodal context. Finally, the cross-attention was applied between the contextual queries and full visual tokens to predict bounding boxes. As per the results, this model attained 43.4 average precision. However, this model led to hallucinated objects and reduced interpretability due to the lack of visualizing the detected object regions.

[19] implemented an automated image captioning system through object detection. Initially, the images and their captions were collected, followed by preprocessing. Further, the HybridNet was applied to extract discriminative and visual features. Finally, the captions were generated using a Bidirectional Gated Recurrent Unit (BiGRU) via optimally tuning the hyperparameters using the Salp Swarm Algorithm (SSA). As per the results, this model achieved 69.06 BiLingual Evaluation Understudy-1(BLEU-1). Yet, this model faced a challenge in capturing long-term dependencies, affecting the model's accuracy during object detection.

[20] demonstrated an image captioning model for Natural Language Processing (NLP) and object detection. Primarily, the images were collected and preprocessed. Further, the ShuffledNet was applied as the encoder to extract features. Finally, the captions were generated by the decoder using the Hybrid Convolutional Neural Network (HCNN). As per the validation, this model attained 69.23 BLEU-1. Yet, this model was computationally expensive due to the large images and complex architectures.

[21] presented a method on rotation-aware 3D vehicle recognition that addresses spatial and orientation variability in aerial imagery. This concept is incorporated into the proposed framework through adaptive anchor design and multimodal fusion, enhancing robustness and accuracy in complex object detection tasks.

[22] introduced an automated image captioning model combining NLP and object detection. Here, the combiner module was used to fuse outputs from three models using Intersection over Union (IoU) thresholds, selecting the best candidates. Further, the unified feature vectors were obtained. Based on the vectors, the Gated Recurrent Unit (GRU) decoder with NLP rules produced a caption in a structured syntax. Yet, this model failed to consider the spatial information of the objects, affecting the object recognition performance.

[23] presented an NLP-based image captioning model by developing a fusion approach. Here, the multiple captioning models were used to generate candidate captions. Based on the captions, the postgeneration fusion stage selected the best captions using three strategies. As per the results, this ensemble approach improved robustness across varying datasets. Owing to the selection of irrelevant captions, this model struggled to detect objects with relevant captions accurately.

[24] implemented an image captioning model for mimicking human image understanding. Initially, the Xception CNN was pertained to extract the spatial image features. Further, the You Only Look Once version-4 (YOLOv4) detected the objects. Then, the Bahdanau attention mechanism was applied to highlight the most relevant features. Finally, the GRU decoder generated captions through fully connected layers. As per the validation, this model attained ~0.48 BLEU-1. However, this model's effectiveness was affected due to the lack of capturing the object semantic information from captions.

3. Proposed Methodology for Nlp-Based Object Detection and Recognition Through Multimodal Relation Graph Construction

In this section, the proposed methodology for the development of an NLP-based enhanced and robust object detection and recognition model is demonstrated. In the proposed YAAAOLOv8 framework, the input image is processed through a multi-scale feature pyramid, where each feature map implicitly defines a grid structure responsible for localized object prediction. Each grid cell corresponds to a spatial region of the image and predicts bounding boxes for objects whose centers fall within that region, enabling effective detection of both small and large objects. Furthermore, to handle the diverse scale and orientation variations in aerial imagery, an adaptive anchor strategy is employed. The anchor boxes are initialized based on dataset-specific statistics and refined during training by optimizing their overlap with ground-truth bounding boxes. This adaptation allows the model to

better capture variations in object size, aspect ratio, and orientation, thereby improving localization accuracy and bounding box refinement under complex scene conditions.

In the multimodal inference stage, semantic inconsistencies between visual feature representations and textual embeddings are addressed through an adaptive alignment mechanism to ensure decision stability. Specifically, discrepancies between visual predictions and textual cues are identified by evaluating their semantic agreement in terms of object category relevance and contextual consistency. When inconsistencies arise, priority is assigned to the modality with higher confidence, while complementary information from the secondary modality is retained to refine the final prediction. Additionally, a consistency-aware fusion strategy is employed, where mutually reinforcing features are strengthened, and conflicting signals are attenuated to prevent erroneous predictions. This process effectively balances visual evidence and linguistic context, reducing ambiguity in complex scenarios.

Here, an effective DL-based object recognition model is developed by efficiently analyzing the explicit relationships among images and texts through multimodal relation graph construction. The architecture of the proposed work is depicted in Figure 1.

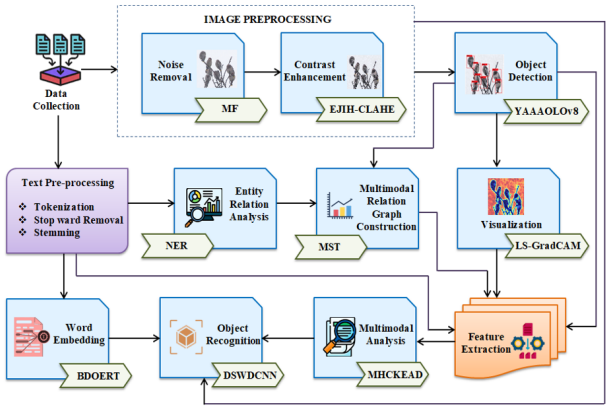


Figure 1. Architecture of the Proposed Work

3.1. Data Collection

Initially, images with their corresponding text/captions are collected from publicly available sources to train the object detection model.

$$Z_{inp}^{a'} = \left\{ Z_{inp}^{1}, Z_{inp}^{2}, Z_{inp}^{3}, \cdots, Z_{inp}^{a''} \right\} \qquad ; a' = 1 \to a'' \tag{1}$$

Where, a'' indicates the number of $Z_{inp}^{a'}$.

3.2. Image Pre-processing

To ensure reproducibility and clarity of the proposed EJIH-CLAHE enhancement process, the key parameter settings used in this study are summarized in Table 1.

Table 1. Parameter settings and configuration details of the proposed EJIH-CLAHE for image enhancement

Parameter	Description	Value/Range
Tile Size	Size of local regions for histogram equalization	8 × 8
Clip Limit	Contrast limiting threshold to avoid over-amplification	0.01 – 0.03
Histogram Bins	Number of bins for intensity distribution	256
Interpolation Method	Method for merging neighboring tiles	Bilinear interpolation
Enhancement Iterations	Number of enhancements passes	1–2
Noise Suppression	Optional smoothing applied after enhancement	Gaussian filter (σ = 1)

The selected parameter configuration provides a balance between contrast enhancement and noise suppression. The clip limit prevents excessive amplification of noise in homogeneous regions, while the chosen tile size ensures preservation of local spatial details. These settings are empirically determined to optimize visual quality and feature clarity in high-resolution remote sensing imagery.

The proposed EJIH-CLAHE is formulated as an optimization-driven contrast enhancement technique that aims to maximize the informational content of the image while controlling noise amplification. Specifically, the method seeks to enhance image entropy by improving the distribution of pixel intensities, ensuring that finer details and textures are preserved. At the same time, a clip-limit constraint is imposed to restrict excessive histogram amplification, preventing over-enhancement and noise boosting in homogeneous regions. The clip limit is adaptively determined based on the joint intensity distribution, ensuring a balance between contrast improvement and structural preservation.

Now, the images $\left(Z_{b'}^{img}\right)$ in $Z_{inp}^{a'}$ often contain noise and inconsistent illuminations, affecting the image quality during object detection. Hence, the prep-processing regarding noise removal and contrast enhancement is performed.

- **Noise Removal:** Initially, due to the presence of sensor imperfection, compression artifacts, and environmental conditions, $Z_{b'}^{img}$ may contain various noises, reducing the object detection efficiency. Therefore, to eliminate such noises from $Z_{b'}^{img}$, the MF is applied. The MF replaces the pixels (p_a, p_b) with the median value (v_{med}) of their neighboring pixels, effectively eliminating noise while preserving edges.

$$Z_{noi}^{img} = \Sigma \frac{(p_a + p_b)}{2} v_{med} * \left(Z_{b'}^{img}(p_a, p_b)\right) \quad (2)$$

Where, Z_{noi}^{img} indicates the noise-free images.

- **Contrast Enhancement:** Further, the visual clarity of Z_{noi}^{img} is enhanced by performing contrast enhancement using the proposed EJIH-CLAHE. Traditional Contrast Limited Adaptive Histogram Equalization (CLAHE) effectively highlights the fine details in various image regions while enhancing the contrast equality rather than globally. However, the improper setting of a clip limit parameter, CLAHE lead to over or under-smoothing. Therefore, the Entropy of Joint Intensity Histogram (EJIH) is used to set the clip limit by optimizing the balance between the image information and noise amplification.

Initially, the EJIH is used to set the clip limit $(\hat{\lambda}^{clip})$ parameter with respect to Z_{noi}^{img}.

$$\hat{\lambda}^{clip} = -\Sigma\Sigma f_{jpm}(p_x, p_y) \log f_{jpm}(p_x, p_y) \quad (3)$$

Where, $(\hat{\lambda}^{clip}, f_{jpm})$ depict the tuned $\hat{\lambda}^{clip}$ and joint probability mass function, respectively, and (p_x, p_y) indicates the pixels of Z_{noi}^{img}. Now, Z_{noi}^{img} is decomposed into c'' number of small tiles $(T_{c'})$, followed by histogram (λ^{his}) computation.

$$\lambda^{\text{his}}(T_{c'}) = \sum_{c'=1}^{c''} T_{c'}\left[p^{\text{int}}(c_c, c_d)\right] \quad (4)$$

Here, p^{int} and (c_c, c_d) demonstrates the pixel intensities and pixel coordinates of $T_{c'}$, correspondingly. Then, based on $\hat{\lambda}^{clip}$, λ^{his} is clipped to prevent over- or under-enhancement.

$$\tilde{\lambda}^{\text{his}} = \min\left(\lambda^{\text{his}}; \left(\hat{\lambda}^{clip}\right)\right) \quad (5)$$

Where, $\tilde{\lambda}^{\text{his}}$ indicates the clipped histogram. Next, the cumulative distribution function (f_{cum}) of $\tilde{\lambda}^{\text{his}}$ is calculated, ensuring contrast equalization across the pixels (p_c).

$$f_{cum} = \frac{1}{d''} \Sigma (p_c)_{d'} \left(\tilde{\lambda}^{\text{his}}\right); d' = 1 \rightarrow d'' \quad (6)$$

Here, d'' demonstrates the number of pixels (p_c) in $\tilde{\lambda}^{\text{his}}$. Finally, contrast-enhanced or preprocessed images $\left(Z_{pre}^{e'}\right)$ are obtained by normalizing f_{cum}.

$$Z_{pre}^{e'} = \left(\frac{f_{cum} - \min(f_{cum})}{\max(f_{cum}) - \min(f_{cum})}\right) \quad (7)$$

Thus, by using the proposed EJIH-CLAHE, the visual clarity of $Z_{pre}^{e'}$ is enhanced while preventing the edge information.

The sequential preprocessing operations significantly influence the feature-space distribution prior to model learning. Initially, contrast enhancement techniques such as EJIH-CLAHE improve the dynamic range of pixel intensities, leading to a more uniform distribution of feature values and enhanced visibility of object boundaries. This reduces the overlap between foreground and background feature representations. Subsequent filtering and normalization steps suppress noise and stabilize intensity variations, resulting in smoother feature transitions and reduced intra-class variability. Furthermore, edge and texture enhancement operations introduce higher discriminative gradients, enabling clearer separation of structural patterns in the feature space. As a result, the transformed feature distribution exhibits improved clustering of similar object classes and increased separability between distinct categories. This structured feature representation facilitates more effective learning in downstream detection and classification stages, ultimately improving model convergence,

robustness, and localization accuracy under weak supervision conditions.

3.3. Object Detection

Now, from $Z_{pre}^{e'}$, the objects detection is carried out to localize the objects with their labels using the proposed You Aspect-ratio Adaptive Anchors Only Look Once version-8 (YAAAOLOv8). Traditional You Only Look Once version-8 (YOLOv8) is extremely fast and predicts both bounding boxes and class probabilities in one pass over the entire image. However, it may produce less precise bounding boxes, particularly for irregularly shaped or densely packed objects. Hence, the Aspect-ratio Adaptive Anchors (AAA) is used as a post-processing technique in object detection that combines overlapping bounding boxes for the same object into a single, more accurate box by weighting coordinates according to confidence scores, eliminating redundant detections while retaining the most reliable predictions.

- Initially, $Z_{pre}^{e'}$ divided into f'' number of grid cells $\left(c_{f''}^{pre}\right)$.
- Further, the CNN acts as the backbone that extracts the high, low, and medium-level features from each $c_{f''}^{pre}$.

$$f_{FM}\left[c_{f''}^{pre}\right] = \text{conv}\left[c_{f''}^{pre}\right] \tag{8}$$

Where, f_{FM} indicates the feature maps from CNN.

- Now, f_{FM} from various layers of backbone are combined and refined in the neck component to enhance the semantic and positional information.

$$\ddot{f}_{FM} = f_{neck}\left[f_{FM}\right] \tag{9}$$

Here, $\left(\ddot{f}_{FM}, f_{neck}\right)$ illustrates the refined f_{FM} and neck layer function, respectively.

- Further, the prediction head provides the predicted class probabilities with boundary box coordinates $\left(c_{wid}, c_{hei}, c_{x-c}, c_{y-c}\right)$ for object detection based on $\ddot{f}_{FM}$,

$$\mathrm{B}^{box} = \left\{c_{wid} \quad c_{hei} \quad c_{x-c} \quad c_{y-c}\right\} \tag{10}$$

Where, B^{box} indicates the bounding box.

- Finally, objects-detected images with labels $\left(Z_{Obj}\right)$ are obtained via eliminating the redundant bounding boxes using the AAA, ensuring that each detected object has a single, accurate bounding box.

$$\ddot{\mathrm{B}}^{box} = \frac{w_{\mathrm{B}^{box}}}{h_{\mathrm{B}^{box}}} \Rightarrow Z_{Obj} \tag{11}$$

Where, $\left(w_{\mathrm{B}^{box}}, h_{\mathrm{B}^{box}}\right)$ signify the width and height of B^{box}, respectively and $\ddot{\mathrm{B}}^{box}$ illustrates the AAA-based B^{box}.

The pseudo-code for the proposed YAAAOLOv8 is demonstrated as follows,

Pseudocode: YAAAOLOv8 Object Detection

Input:
{Ii}\{I_i\} {Ii} → Set of preprocessed input images

Output:
DDD → Detected objects with class labels

1: Initialize model parameters θ, grid size GGG, and detection thresholds

2: for each I_i do

3: Divide I_i into G×G grid cells

4: Extract feature maps Fi from backbone network

5: Compute multi-scale feature representations

6: Fuse and refine features:
F^i=Ffusion(Fi)\hat{F}_i = \mathcal{F}_{fusion}(F_i)F^i=Ffusion(Fi)

7: Predict bounding boxes BiB_iBi and class probabilities PiP_iPi

8: Apply adaptive attention/aggregation module (AAA)

9: if detection confidence ≥τ\geq \tau≥τ then

10: Retain detection BiB_iBi

11: else

12: Discard detection

13: end if

14: end for

15: Return final detection set DDD with labeled objects

Thus, the proposed YAAAOLOv8 accurately detects the objects with their labels.

The Aspect-ratio Adaptive Anchor mechanism in the proposed YAAAOLOv8 framework is designed to dynamically adjust anchor configurations based on the geometric diversity of objects present in aerial imagery. Unlike fixed anchor schemes, the proposed approach adapts anchor scales and aspect ratios by analyzing the distribution of object dimensions during training, enabling better alignment with heterogeneous object shapes such as elongated structures, compact vehicles, and arbitrarily oriented instances. The adaptive optimization principle is guided by the objective of maximizing overlap between anchor boxes and ground-truth object regions while minimizing localization error. During training, anchor configurations are iteratively refined based on their matching quality with annotated objects, allowing the model to prioritize anchor shapes that better represent dominant geometric patterns in the dataset. This dynamic scaling ensures that anchors expand or contract in response to variations in object size and aspect ratio, thereby improving detection sensitivity for both small-scale and large-scale objects.

3.4. Visualization

Now, detected objects are localized in the corresponding Z_{Obj} using the proposed LS-GradCAM for accurate object recognition, enhancing the model's interpretability. Traditional Gradient-weighted Class Activation Mapping (Grad-CAM) highlights the regions that are most relevant to a specific target class, improving interpretability. However, it often misses highlighting the important regions when gradients are vanishing or saturating. Therefore, the Logit Scaling (LS) is applied before the backpropagation, preventing the gradients from vanishing or saturating by sharpening the probability distribution.

- Primarily, the class probabilities (p_{cls}) are extracted from Z_{Obj} by passing it to the CNN.

$$p_{cls} = \text{conv}[Z_{Obj}] \tag{12}$$

- Further, the LS is applied on p_{cls} to highlight the important regions while preventing the gradients from vanishing or saturating.

$$F_{LS} = f_{sca} * p_{cls} \tag{13}$$

Where, (F_{LS}, f_{sca}) indicate the outcome of LS and scaling factor, respectively.

- Now, with the help of F_{LS}, the backpropagation is performed to calculate the gradients (g_{gra}) of the target class score with respect to the feature map (l_{fin}^{FM}) of final convolutional layer.

$$g_{gra} = \frac{\partial p_{cls}}{\partial l_{fin}^{FM}} * F_{LS} \tag{14}$$

Where, ∂ indicates the derivative function.

- Further, the weight value (ϖ^{wei}) for each p_{cls} is determined based on g_{gra}.

$$\varpi^{wei} = \frac{1}{h''} \Sigma \frac{\partial l_{spa}}{\partial p_{cls}} \tag{15}$$

Here, (h'', l_{spa}) represent the number of pixels in Z_{Obj} and spatial locations in p_{cls}, respectively.

- Based on ϖ^{wei}, the class activation maps (A^{cls}) are generated.

$$A^{cls} = \max[0, \varpi^{wei}] \tag{16}$$

- Finally, by upsampling and superimposing A^{cls} on Z_{Obj}, the objects are effectively visualized. Then, the objects-visualized images are indicated as (Z_{Vis}).

Thus, the proposed LS-GradCAM efficiently localizes and visualizes the objects from the skeletonized images for ensuring precise object recognition.

3.5. Text Pre-processing

In the meantime, the texts/captions $(Z_{i'}^{txt})$ from $(Z_{inp}^{a'})$ are collected to ensure accurate and precise recognition of objects. Here, $Z_{i'}^{txt}$ often contains irrelevant symbols, spelling variations, and redundant words, affecting the semantic understanding during object recognition. Therefore, text pre-processing regarding tokenization, stopword removal, and stemming are performed.

- **Tokenization:** Primarily, each $Z_{i'}^{txt}$ is decomposed into smaller meaningful units called

tokens/words/subwords $\left(T_{j''}\right)$, ensuring faster computation.

- **Stopword Removal:** Further, the words that do not provide much meaning are removed from $T_{j''}$, reducing irrelevant symbols and text dimensionality.
- **Stemming:** Finally, the preprocessed texts $\left(Z_{k'}^{pre}\right)$are obtained by reducing the words to their root forms, eliminating complexity.

$$Z_{k'}^{pre} = \left\{Z_1^{pre}, Z_2^{pre}, Z_3^{pre}, \cdots, Z_{k''}^{pre}\right\}; k' = 1 \to k'' \quad (17)$$

Where, k'' indicates the number of $Z_{k'}^{pre}$.

3.6. Entity Relation Analysis

Afterwards, the entity relation analysis is performed on each word in to analyze their relationships and contextual dependencies for effective and robust object recognition using the Named Entity Recognition (NER) approach. NER is the process of locating and classifying key information (i.e., entities) in text into predefined categories (i.e., persons, organizations, and locations), transforming unstructured text into structured data. Then, the entity-relation analyzed outcome is represented in Z^{ER}.

The Named Entity Recognition (NER) process is guided by operational criteria that prioritize object-relevant semantic categories. Specifically, extracted entities are filtered based on their correspondence to visually identifiable object classes, such as vehicles, buildings, infrastructure elements, and environmental features commonly present in remote sensing imagery. Entities that do not have a direct visual representation or lack spatial relevance are excluded from further processing. Additionally, contextual relevance is considered by evaluating the co-occurrence of entities within descriptive annotations, ensuring that only semantically meaningful relationships contributing to object localization are retained. Priority is given to entities that exhibit strong spatial and categorical consistency with annotated object classes, thereby improving the alignment between textual semantics and visual targets. This selection strategy enables the model to effectively integrate linguistic cues into the object recognition pipeline, enhancing detection accuracy and contextual understanding within the proposed framework.

3.7. Multimodal Relation Graph Construction

Now, the multimodal relation graphs are constructed based on Z^{ER} and Z_{Obj} to determine the interaction, semantic, and contextual dependencies and support higher-level reasoning for object recognition using the proposed MPST. Traditional Minimum Spanning Tree (MST) avoids redundant paths and unnecessary connections and connects all nodes with the least possible sum of edge weights. However, frequent node/edge updates require recomputation, increasing computational cost in real-time systems. Therefore, the Persistence function is applied to efficiently handle frequent updates by preserving the history of the graph.

- ✧ Initially, from Z^{ER} and Z_{Obj}, the entities are considered as nodes, and their relations are considered as edges $\left(N_{edg}\right)$.
- ✧ Now, the weight value $\left(\varpi_{edg}\right)$ for each N_{edg} is applied, attaining weight-assigned edges $\left(\tilde{N}_{edg}\right)$.
- ✧ Finally, the multimodal relation graph $\left(M_{MRG}\right)$is constructed by tracing N_{edg} with the minimum ϖ_{edg} and efficiently handling the frequent node/edge updates using the Persistence function $\left(f_{Per}\right)$.

$$M_{MRG} = \arg\min\left[\varpi_{edg}\left(N_{edg}\right)\right] * f_{Per} \quad (18)$$

$$f_{Per} = \frac{1}{m''} \int f_{ind}\left(\tilde{N}_{edg}\right) dt \quad (19)$$

Where, $\left(m'', f_{ind}\right)$ indicate the number of N_{edg} and indicator function, respectively.

Thus, the proposed MPST efficiently constructs the graph by analyzing the multimodal relation among the entity relations and detected objects. The Minimum Persistence Spanning Tree (MPST) is an extension of the traditional Minimum Spanning Tree that incorporates temporal stability into graph construction. In this work, the multimodal relation graph consists of nodes representing entities from text and detected objects, while edges represent their semantic and contextual relationships. Unlike conventional approaches, MPST considers both the current relationship strength and historical importance of connections, ensuring stable graph formation over time.

3.8. Feature Extraction

Further, the features that are crucial for accurate and robust object recognition are extracted from $\left(Z_{k'}^{pre}, M_{MRG}, Z_{Vis}, Z_{Obj}\right)$. The extracted features

from Z_{Vis} includes spatial features, channel-wise important features, object/shape boundaries, location cues, and contextual features. Similarly, features like Node degree, Edge weight statistics, and Betweenness centrality are extracted from M_{MRG}. Likewise, the NLP features such as bag of words, Term Frequency-Inverse Document Frequency (TF-IDF), word count, and character count are extracted from $Z_{k'}^{pre}$. Further, from Z_{Obj}, the features like edges, textures, shapes, and contextual cues, bounding box coordinates, objectness score, and class probabilities.

$$Z_{l'}^{fea} = \left\{Z_1^{fea}, Z_2^{fea}, Z_3^{fea}, \cdots, Z_{l''}^{fea}\right\} \quad ; l' = 1 \to l'' \tag{20}$$

Where, l'' illustrates the number of extracted features $\left(Z_{l'}^{fea}\right)$ from $\left(Z_{k'}^{pre}, M_{MRG}, Z_{Vis}, Z_{Obj}\right)$.

3.9. Multimodal Analysis

Then, the multimodal analysis is carried out on $Z_{l'}^{fea}$ to ensure interpretation of visual and linguistic information for effective object recognition through feature fusion using the proposed MHCKEAD. The multimodal analysis is the process of fusing the visual features extracted from images with the semantic features extracted from texts to improve object recognition and understanding. Attention maps from different heads in traditional Multi-Head Cross Attention (MHCA) provide insights into cross-modal interactions, helping to analyze predictions-influence features. However, MHCA exhibits an overfitting issue due to the usage of a large parameter space. Therefore, the Kapur Entropy-based Attention Dropout (KEAD) is applied to tune the model parameters by dynamically adjusting the dropout probability for each head.

- ⇔ Primarily, the model parameters are dynamically tuned $\left(P_{tun}\right)$ with respect to $Z_{l'}^{fea}$ using the KEAD to eliminate the overfitting issue.

$$P_{tun} = \frac{1}{-\sum w_{att} \log\left(w_{att}\right)} \tag{21}$$

Where, w_{att} indicates the attention weights.

- ⇔ Further, based on P_{tun}, $Z_{l'}^{fea}$ are transformed into Query $\left(V_Q\right)$, key $\left(V_K\right)$, and value $\left(V_V\right)$ vectors. Here, the query indicates the semantic intent of the text, the key represents the visual relevance cues of image regions, and the value signifies the contextual visual information.

$$\begin{cases} V_Q = m_{wei}^{Q} Z_{l'}^{fea} \\ V_K = m_{wei}^{K} Z_{l'}^{fea} \\ V_V = m_{wei}^{V} Z_{l'}^{fea} \end{cases} \tag{22}$$

Where, $\left(m_{wei}^{Q}, m_{wei}^{K}, m_{wei}^{V}\right)$ indicate the learnable weight matrices for $\left(V_Q, V_K, V_V\right)$, respectively.

- ⇔ Now, the cross-attention weights score $\left(W_{crs}\right)$ are computed by the multiple attention heads to capture the cross-model relevance between $Z_{l'}^{fea}$.

$$W_{crs}\left(V_Q, V_K, V_V\right) = \tilde{\Psi}^{sof}\left[\frac{V_Q\left[V_K\right]^{f_{tra}}}{\sqrt{d_{\dim}\left(V_K\right)}}\right]V_V \tag{23}$$

Here, $\left(\tilde{\Psi}^{sof}, f_{tra}, d_{\dim}\right)$ represent the softmax function, transpose function, and dimensionality of V_K, respectively.

- ⇔ Finally, by concatenating W_{crs} from the multiple attention heads, the proposed MHCKEAD provides the unified multimodal representations through linear transformation, representing fused features $\left(M^{IT}\right)$.

Thus, the proposed MHCKEAD efficiently analyzes the multimodal features and provides its unified representations.

3.10. Word Embedding

In the meantime, $Z_{k'}^{pre}$ are transformed into vector representations for better understanding during object recognition using the proposed Bidirectional Dice Overlap Encoder Representations from Transformers (BDOERT), capturing semantic meaning and contextual relationships among words. Traditional Bidirectional Encoder Representation from Transformers (BERT) captures deep contextual relationships and is excellent for tasks that require understanding of context, co-reference, and other linguistic nuances. In the BDOERT embedding architecture, the Dice Overlap Distance is employed as an optimization criterion to guide contextual representation learning. Specifically, the Dice-based measure evaluates the overlap between predicted semantic representations and ground-truth annotations, thereby providing a direct signal for aligning learned embeddings with target object structures. This enables the model to emphasize regions with higher semantic relevance while suppressing irrelevant or noisy features.

The architectural configuration of the proposed BDOERT model is defined to ensure consistency and reproducibility. The transformer encoder comprises 6 stacked layers, each with an embedding dimension of 512. Multi-head self-attention is implemented using 8 attention heads, enabling the model to capture diverse contextual dependencies across feature representations. The position-wise feed-forward network within each transformer block has a dimensionality of 2048, providing sufficient capacity for nonlinear feature transformation.

To enhance generalization and prevent overfitting, a dropout rate of 0.1 is applied to both the attention and feed-forward layers. Residual connections and layer normalization are incorporated to stabilize training and improve gradient flow. This configuration balances computational efficiency and representational power, making it suitable for multimodal contextual embedding within the proposed framework.

However, due to the cosine similarity between word embedding and document embedding-based results, BERT is computationally intensive and may extracts truly irrelevant words as keywords. Therefore, the Dice Overlap Distance (DOD) is applied to find the top similarity by comparing the dice similarity scores.

- Primarily, the dimensional vector representations $\left(d^{vec}\right)$ of each $Z_{k'}^{pre}$ are obtained via performing positional embedding $\left(e_{pos}\right)$, token embedding $\left(e_{pos}\right)$, and segment embedding $\left(e_{pos}\right)$.

$$d^{vec}\left(Z_{k'}^{pre}\right)=e_{pos}\left(Z_{k'}^{pre}\right)+e_{tok}\left(Z_{k'}^{pre}\right)+e_{seg}\left(Z_{k'}^{pre}\right) \tag{24}$$

- Now, to extract the truly relevant words from $Z_{k'}^{pre}$, the similarity $\left(\aleph_{DOD}\right)$ between d^{vec} for each $Z_{k'}^{pre}$ is computed using the DOD.

$$\aleph_{DOD}\left(d_1^{vec},d_2^{vec}\right)=\frac{-\left(d_1^{vec}\cap d_2^{vec}\right)}{\left|d_1^{vec}\right|+\left|d_2^{vec}\right|+\min\left(\left|d_1^{vec}\right|,\left|d_2^{vec}\right|\right)} \quad ;\left(d_1^{vec},d_2^{vec}\right)\in d^{vec} \tag{25}$$

- Further, based on $\aleph_{DOD}$, the bidirectional self-attention mechanism is applied to capture the contextual representations $\left(r_{\cot}\right)$ from d^{vec}.

$$r_{\cot}=\Sigma\,\omega_{bid}\cdot v_{val}\left(d^{vec}\right) \tag{26}$$

Where, $\left(v_{val},\omega_{bid}\right)$ represent the random value and weight value of self-attention mechanism, correspondingly.

- Finally, based on $r_{\cot}$, the $Z_{k'}^{pre}$ are updated $\left(\ddot{Z}_{k'}^{pre}\right)$, providing word-embedded/vector representations $\left(Z_{WEO}\right)$.

$$\ddot{Z}_{k'}^{pre}=\Sigma\, r_{\cot}\left[d^{vec}\right]\Rightarrow Z_{WEO} \tag{27}$$

Thus, the proposed BDOERT efficiently provides a vector representation for the preprocessed text while effectively analyzing the semantic and contextual meaning of the words.

3.11. Object Recognition

Finally, the classification model is trained based on $\left(M^{IT},Z_{WEO},Z_{pre}^{e'}\right)$ using the proposed DSWDCNN to accurately recognize objects with captions. Traditional Deep Convolutional Neural Network (DCNN) learns features directly from raw data and provides hierarchical, translation-invariant feature learning with strong representational capacity, making them highly effective for complex spatial pattern recognition tasks. However, the poor weight initialization leads to slow learning or divergence, and DCNN suffers from the vanishing gradient problem due to the poor learning efficiency. Hence, the Wishart Distribution Initialization (WDI) is applied to initialize the model weights by sampling initial weights from a Wishart distribution rather than a traditional Gaussian or uniform distribution. Also, the learning efficiency in object recognition is enhanced by using the Swim Activation function. The architecture of the proposed DSWDCNN is depicted in Figure 2.

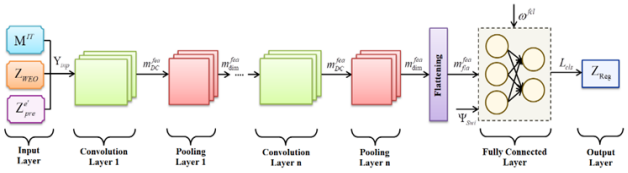


Figure 2. Architecture of the Proposed DSWDCNN

- Initially, $\left(M^{IT},Z_{WEO},Z_{pre}^{e'}\right)$ are given as input $\left(Y_{inp}\right)$ to the input layer of the proposed DSWDCNN.
- Further, the multiple convolutional and pooling operations are performed to capture the hierarchical features from Y_{inp}. Here, feature maps are extracted from Y_{inp} by applying filters at the convolutional layer.

$$m_{DC}^{fea}=\left[\frac{Y_{inp}+2s_{pad}-s_{\ker}}{s_{str}}\right]+1 \quad (28)$$

Where, $\left(m_{DC}^{fea},s_{pad},s_{\ker},s_{str}\right)$indicate the features map extracted from multiple convolutional layers, convolution padding size, convolution kernel size, and convolution stride size, correspondingly.

- Further, the dimensionality of each m_{DC}^{fea} is reduced while preserving important details using the pooling layer.

$$m_{\dim}^{fea}=f_{\text{Pool}}\left(m_{DC}^{fea}\right) \quad (29)$$

Here, $\left(m_{\dim}^{fea},f_{\text{Pool}}\right)$represent the dimensionality-reduced m_{DC}^{fea} and pooling function, respectively.

- Next, for accurate and robust object recognition, $m_{\dim}^{fea}$ are flattened $\left(m_{fla}^{fea}\right)$ into a one-dimensional vector.

$$m_{\dim}^{fea}\xrightarrow{flatening}m_{fla}^{fea} \quad (30)$$

- Subsequently, the predicted class labels $\left(L_{cls}\right)$ for object recognition are obtained by the fully connected-layer with the help of Swim activation function $\left(\Psi_{Swi}\right)$. Here, the WDI-based model weights initialization is carried out, eliminating the slow learning or divergence.

$$L_{cls}=\Psi_{Swi}\left[\omega^{fcl}\cdot m_{fla}^{fea}+b^{fcl}\right] \quad (31)$$

$$\omega^{fcl}=M_{WD}\left[M_{\text{cov}},d_{fre}\right] \quad (32)$$

$$\Psi_{Swi}\left(m_{fla}^{fea}\right)=\frac{m_{fla}^{fea}}{2}\left[1+\frac{p_{par}m_{fla}^{fea}}{\sqrt{1+\left(m_{fla}^{fea}\right)^{2}}}\right] \quad (33)$$

Where, $\left(\omega^{fcl},b^{fcl}\right)$indicate the weight and bias value of the fully-connected layer, respectively and $\left(M_{WD},M_{\text{cov}},d_{fre},p_{par}\right)$illustrate the Wishart distribution matrix, covariance matrix, degree of freedom, and trainable parameter, correspondingly.

- Finally, based on L_{cls} , the output layer produces the classified outcome $\left(Z_{\text{Reg}}\right)$as the recognized objects with its corresponding captions.

The pseudo-code for the proposed DSWDCNN is demonstrated as follows,

Pseudo-code for DSWDCNN

Input:
F^i\hat{F}_iF^i → Fused feature representations
viv_ivi → Vectorized feature values
{Ii}\{I_i\}{Ii} → Set of preprocessed input images

Output:
D → Recognized objects with class labels (captions)

1: Initialize model parameters θ\thetaθ, weight matrix WWW, and bias bbb

2: for each IiI_iIi do

3: Apply convolution and pooling operations to extract feature maps:
Fi=Fconv(Ii)F_i = \mathcal{F}_{conv}(I_i)Fi=Fconv(Ii)

4: Reduce dimensionality of feature maps:
F~i=Fdim(Fi)\tilde{F}_i = \mathcal{F}_{dim}(F_i)F~i=Fdim(Fi)

5: Flatten feature maps into vector form:
vi=Flatten(F~i)v_i = \text{Flatten}(\tilde{F}_i)vi=Flatten(F~i)

6: Initialize/update weights (WDI):
W=Winit(vi)W = \mathcal{W}_{init}(v_i)W=Winit(vi)

7: Apply Swim activation function:
ai=σswim(W·vi+b)a_i = \sigma_{swim}(W \cdot v_i + b)ai=σswim(W·vi+b)

8: Compute class probabilities:
Pi=Softmax(ai)P_i = \text{Softmax}(a_i)Pi=Softmax(ai)

9: if classification confidence ≥τ\geq \tau≥τ then

10: Assign label yi=arg⁡max⁡(Pi)y_i = \arg\max(P_i)yi=argmax(Pi)

11: else

12: Mark as uncertain / reject

13: end if

14: end for

15: Return final recognized set D={yi}D = \{y_i\}D={yi} with captions

Thus, the proposed DSWDCNN accurately recognizes the objects with their correspondingly class labels.

4. Result and Discussion

In this phase, the proposed framework's performance in object detection is demonstrated by comparing it with several traditional methodologies. The implementation of this work is conducted in the working platform of PYTHON. In LS-GradCAM, the localization maps are derived from gradients of class-specific logits with respect to feature maps. The magnitude of these gradients is inherently influenced by the scale of the logits, which introduces sensitivity to logit scaling factors. Specifically, when logits are scaled by a factor α, the resulting gradients are proportionally amplified, leading to sharper and more concentrated activation maps. However, excessive scaling may over-amplify dominant regions while suppressing secondary contextual cues, thereby reducing spatial generalization.

The proposed framework is implemented using the Python programming platform with deep learning libraries to ensure computational efficiency and scalability. The experiments are conducted on a system equipped with an NVIDIA GPU (e.g., RTX 3080) with 10 GB memory, an Intel Core i7 processor, and 16 GB RAM. The model is developed using PyTorch (version 1.12) along with supporting libraries such as NumPy (1.21), OpenCV (4.5), and torch vision (0.13). Training and evaluation are performed under a Linux/Windows environment with CUDA acceleration enabled to facilitate efficient model optimization and inference. These configurations ensure reproducibility and consistent performance evaluation across experiments.
The average inference time is computed over the test dataset under consistent hardware settings. The results indicate that the proposed model maintains efficient processing time while achieving high accuracy, demonstrating its scalability for practical deployment. Although the integration of multimodal components introduces additional computational steps, the optimized design of MPST and DSWDCNN ensures that latency remains within acceptable limits for near real-time applications such as surveillance and autonomous systems.

4.1. Dataset Description

The proposed work uses the "Common Objects in Context (COCO) Dataset 2017" to train the object recognition model. The source link to access this dataset is depicted under the reference section. This dataset is a large-scale object detection, segmentation, key-point detection, and captioning dataset. Also, this dataset contains 328,000 images with corresponding annotations. Among the whole data, the proposed work uses 80% (262,400) and 20% (65600) of the data for training and testing, respectively.

Table 2. Focused Object Categories and Sample Images from the DOTA-v2 Dataset

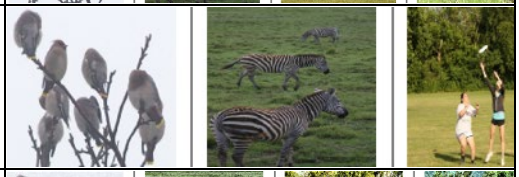
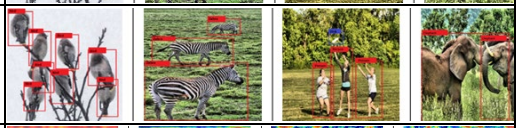
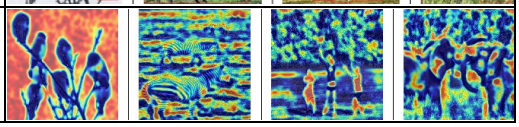
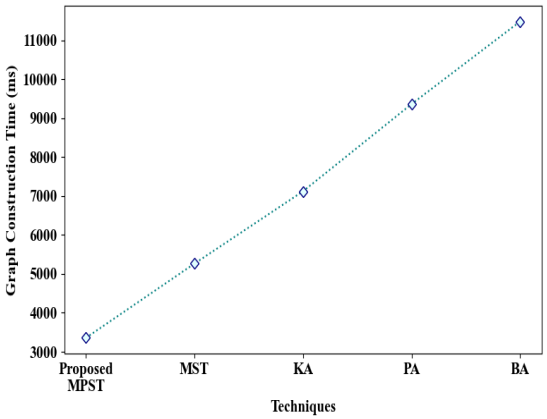
Steps	Sample Images
Input	
Noise Removal	
Contrast Enhancement	
Object Detection	
Visualization	

Table 2 clearly shows the sample image results of the proposed work for object recognition regarding input, noise removal, contrast enhancement, object detection, and visualization.

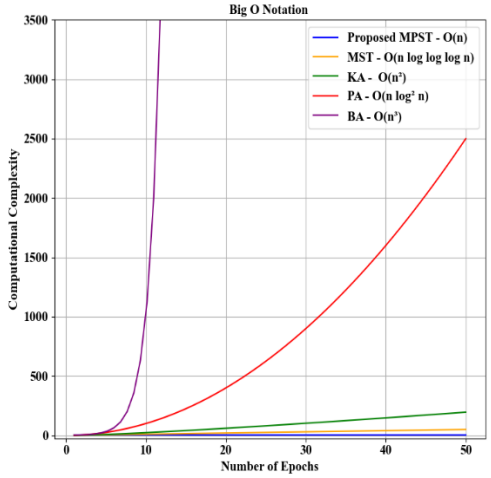
4.2. Performance Validation

In this phase, the proposed framework's performance regarding various performance metrics is compared with traditional methods. All experiments were conducted in a controlled computational environment to ensure fair benchmarking of the proposed MPST against traditional methods. The implementation was carried out using Python with relevant deep learning and scientific computing libraries. The system was equipped with a multi-core processor, sufficient main memory, and GPU acceleration to support efficient model training and graph computations. Consistent hardware and software configurations were maintained across all comparative methods to ensure unbiased evaluation. Additionally, identical input data, preprocessing steps, and execution settings were used for all algorithms, enabling a reliable assessment of computational complexity and graph construction time. This standardized setup ensures that the reported improvements in MPST performance are

attributed to the proposed methodology rather than variations in computational resources.



(a)



(b)

Figure 3 (a) and (b). Graph Construction Time and Computational Complexity Analysis of the Proposed MPST

As per Figure 3(a) and (b), it is proven that the proposed MPST significantly outperforms the traditional methods by requiring minimum time for multimodal relation graph construction (3358ms) and reduced computational complexity. This achievement is attained due to the integration of the persistence function within the MST, eliminating redundant recomputation during frequent node/edge updates. However, the traditional methods like MST, Kruskal's Algorithm (KA), Prim's Algorithm (PA), and Borůvka's Algorithm (BA) show comparatively lower performance in capturing the semantic and logical information among the image and tests, lacking the ability to retain temporal stability. Therefore, the proposed MPST achieves superior performance. Specifically, nodes that exhibit higher similarity in terms of spatial characteristics, texture patterns, and semantic content are assigned stronger edge connections, while less similar nodes receive lower weights. The similarity is computed using a combination of distance-based and directional measures, allowing the model to capture both magnitude differences and feature alignment across modalities. Additionally, normalization is applied to maintain consistent scaling of edge weights across the graph structure. This strategy enables effective encoding of contextual and semantic dependencies, ensuring that closely related features contribute more significantly to the learning process within the MPST framework.

The integration of persistence in the Minimum Persistence Spanning Tree (MPST) framework is designed to reduce computational overhead during dynamic graph updates. In conventional spanning tree construction, any modification in node features or edge weights requires complete recomputation of the tree structure, leading to increased computational complexity, particularly in large-scale multimodal graphs. In contrast, the proposed MPST leverages persistence by retaining previously computed structural relationships and selectively updating only the affected subregions of the graph.

Table 3. Numerical Evaluation of the Proposed DSWDCNN

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Proposed DSWDCNN	98.8569	98.8412	98.9578	98.6598
DCNN	95.6325	95.8471	95.4582	95.2674
DenseNet	93.1524	93.2695	93.1278	93.2295
EfficientNet	90.5628	90.2417	90.3265	90.7814
CNN	87.7845	87.5932	87.4872	87.6291

From Table 3, it is shown that the traditional methodologies suffer from limited performance in object recognition due to the lack of robust weight initialization and the presence of vanishing gradient issues. For instance, the traditional DCNN attains 95.6325% accuracy, 95.8471% precision, 95.4582% sensitivity, and 95.2674% specificity, which are very low when compared to the proposed DSWDCNN. The proposed DSWDCNN shows superior performance in object recognition by achieving higher accuracy (98.8569%), precision (98.8412%), sensitivity (98.9578%), and specificity (98.6598%). There, the proposed DSWDCNN outperforms the traditional methods by integrating WDI and Swim activation function with the traditional DCNN, ensuring stable and efficient weight convergence and enabling deep network training for robust object recognition.

Table 4. Comprehensive perceptual quality and contrast enhancement evaluation of EJIH-CLAHE using PSNR, SSIM, contrast metric, and entropy

Method	PSNR (dB)	SSIM	Contrast Metric	Entropy
Unprocessed Image	21.34	0.814	42.15	6.84
Median Filter (MF)	26.54	0.862	45.3	7.12
Traditional CLAHE	31.87	0.925	58.42	7.56
Proposed EJIH-CLAHE	36.92	0.978	64.88	7.91

Table 4 presents a comparative evaluation of image quality before and after enhancement using perceptual and contrast-based metrics. The unprocessed image shows low PSNR and SSIM values, indicating poor visual quality and structural degradation, along with limited contrast and entropy. The Median Filter (MF) improves noise reduction, resulting in moderate gains in PSNR and SSIM, but only slight contrast enhancement. The Traditional CLAHE significantly enhances contrast and structural similarity; however, its performance is limited by suboptimal clip limit control. In contrast, the proposed EJIH-CLAHE achieves the highest PSNR (36.92 dB) and SSIM (0.978), reflecting superior image fidelity and structure preservation. Additionally, it attains the maximum contrast metric and entropy values, demonstrating enhanced detail visibility and richer information content. Overall, the results confirm that EJIH-CLAHE provides more effective and balanced image enhancement compared to existing methods.

The Wishart Distribution Initialization (WDI) employed in the proposed DSWDCNN framework contributes to improved convergence behavior by providing a statistically structured initialization of network weights. Unlike conventional random initialization methods, which may lead to unstable gradient propagation and slow convergence, the Wishart-based approach generates positive semi-definite weight matrices with controlled variance and correlation properties. This ensures that the initial feature representations are well-conditioned and exhibit balanced activation distributions across layers.

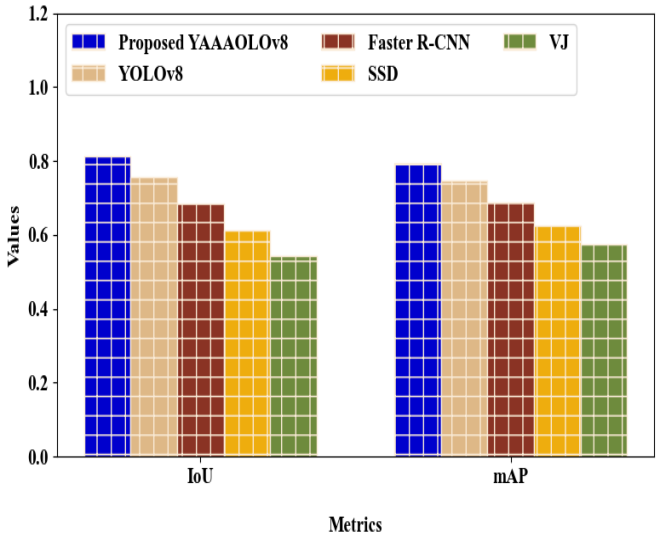


Figure 4. IoU and MAP analysis of the Proposed YAAAOLOv8

As per Figure 4, the proposed YAAAOLOv8's effectiveness in object detection is showcased regarding Intersection over Union (IoU) and mean Absolute Precision (mAP). Here, the proposed YAAAOLOv8 achieves higher IoU and mAP of 0.8142 and 0.7954, correspondingly, which are very high when compared to the traditional methods (YOLOv8, Faster Region-based Convolutional Neural Network (Faster R-CNN), Single Shot multibox Detector (SSD), and Viola–Jones (VJ). This achievement is attained by using the AAA, refining the bounding box predictions by dynamically adjusting the anchor dimensions according to the detected object's aspect ratio. However, the traditional methods rely on fixed dimensions, limiting their ability to ensure accurate object detection by increasing false positives and negatives.

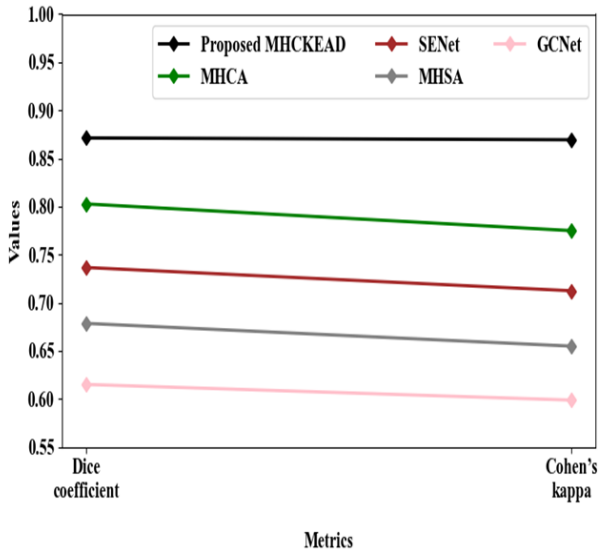


Figure 5. Performance Analysis of the proposed MHCKEAD

From Figure 5, it is noted that the prevailing techniques like MHCA, Squeeze-and-Excitation Network (SENet), Multi-Head Self-Attention (MHSA), and Global Context Network (GCNet) exhibit comparatively limited performance regarding dice coefficient and Cohen's kappa during multimodal analysis. For instance, the traditional MHCA attains 0.8026 Dice Coefficient and 0.7748 Cohen's Kappa due to the presence of overfitting in a large parameter space, failing to capture the fine-grained cross-model dependencies. In contrast, the proposed MHCKEAD uses the Kapur Entropy-based dropout to dynamically regulate the attention weights. This integration enhances the performance of proposed MHCKEAD in multimodal analysis by significantly achieving a higher dice coefficient (0.8712) and Cohen's kappa (0.8693) than the prevailing methods.

In the MHCKEAD module, Kapur entropy is utilized as a guiding measure to adaptively regulate attention-head dropout, enabling dynamic control of feature learning. Specifically, the entropy value reflects the information richness and distribution complexity of the feature space. Higher entropy indicates diverse and informative feature representations, while lower entropy suggests redundancy or limited variability. Based on this observation, an entropy-driven mapping strategy is employed to adjust the dropout probability of attention heads. When the entropy value is high, a lower dropout rate is applied to preserve informative features and maintain strong attention responses. Conversely, when entropy is low, the dropout probability is increased to suppress redundant activations and encourage feature diversification. This adaptive mechanism ensures that attention heads are selectively regularized based on the underlying information content.

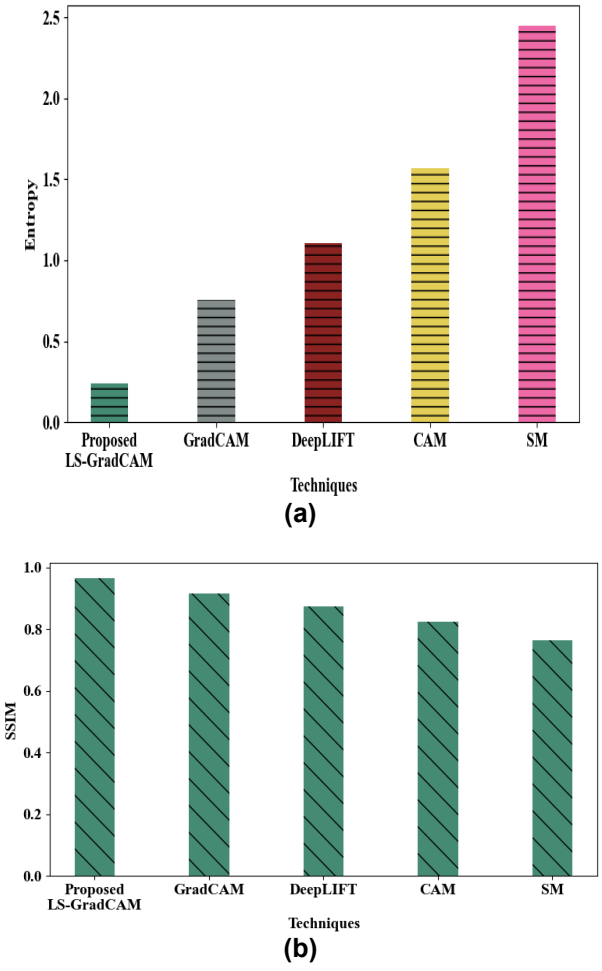


Figure 6. (a) Entropy and (b) SSIM analysis of the proposed LS-GradCAM

As per Figure 6, the proposed LS-GradCAM shows comparatively higher performance in visualizing salient regions of the detected objects by achieving 0.2415 entropy and 0.9658 Structural Similarity Index Measure (SSIM). This higher performance is attained by the integration of the LS mechanism into the traditional GradCAM, stabilizing the gradient flow and preventing saturation or vanishing during backpropagation. However, the traditional methods (GradCAM, Deep Learning Important FeaTures (DeepLIFT), Class Activation Mapping (CAM), and Saliency Map (SM)) suffer from limited performance in visualizing the detected object regions due to the instability in gradient flow. Therefore, by addressing the traditional methodologies' shortcomings, the proposed LS-GradCAM achieves superior performance in object region visualization.

Figure 7. Qualitative visualization of attention regions using Grad-CAM

In Figure 7, the GradCAM highlights a broad region around the central object, including irrelevant background areas such as surrounding structures. This indicates poor localization and attention dispersion. In contrast, the proposed LS-GradCAM produces a more concentrated heatmap, focusing primarily on the target human subject while suppressing background noise. This demonstrates improved interpretability and precise object localization due to stabilized gradient propagation.

Table 5. Quantitative Analysis of Image Enhancement using SSIM and Entropy

Image Stage	SSIM	Entropy	Visual Interpretation
Original Image	0.72	5.81	Low contrast with blurred edges and limited texture details
CLAHE	0.84	6.45	Improved contrast but slight noise amplification in homogeneous regions
EJIH-CLAHE (Proposed)	0.91	7.12	Enhanced edge sharpness, better texture clarity, and improved object delineation

Higher SSIM indicates better structural preservation, while increased entropy reflects enhanced information content and contrast. The proposed EJIH-CLAHE method achieves superior performance, consistent with improved visual quality is shown in table 5.

Table 6. Numerical Evaluation of the Proposed EJIH-CLAHE

Methods	MSE	RMSE	MAE
Proposed EJIH-CLAHE	0.2145	0.2587	0.1547
CLAHE	0.7562	0.8451	0.6982
BHE	1.2345	1.2612	1.1023
AHE	3.3695	3.4218	2.6358
HE	5.2147	5.2417	4.1287

From Table 6, it is proven that the proposed EJIH-CLAHE significantly outperforms the traditional methods by using the entropy-driven clip limit adjustment mechanism, enhancing the image contrast and achieving a limited Mean Square Error (MSE) (0.2145), Root Mean Square Error (RMSE) (0.2587), and Mean Absolute Error (MAE) (0.1547). However, the traditional methods like CLAHE, Bi-Histogram Equalization (BHE), Adaptive Histogram Equalization (AHE), and Histogram Equalization (HE) exhibit higher error values during image contrast enhancement due to the global enhancement artifacts, leading to over or under-enhancement. Thus, the proposed EJIH-CLAHE ensures superior and effective contrast enhancement for object detection and recognition.

4.3. Comparative Analysis

In this phase, the comparative analysis is carried out.

Table 7. Comparative Analysis

Author's Name	Objectives	Techniques	Advantages	Limitations
Proposed work	NLP-based object detection and recognition through multi-modal relation graph construction	DSWDCNN	Enhanced the classification accuracy by efficiently capturing the semantic and logical information among images and texts	The proposed work didn't support for extreme environmental conditions.
[22]	Hierarchical object detection through automatic image captioning	MR-CNN	Enhanced precision and richer semantic representation	Due to the rigid caption structure, the natural sentence variability and linguistic fluency were affected.
[25]	Infrared small object detection	HCF-Net	Demonstrated superior detection accuracy and robustness	Owing to the reliance on a specific dataset, this model had a high computational cost.
[26]	Multimodal object detection through cross-mamba interaction analysis	YOLOv8	Achieved offset robustness and increased efficiency	Yet, this model relied on high-level features, limiting the fine-grained detail extraction.
[27]	Multimodal object detection system for real-world applications	ViT	Achieved high classification accuracy	However, the real-time efficiency was affected due to the sensitivity to spatial information.
[3]	Remote sensing object detection and recognition	YOLO-SE	Efficiently detected the small objects	Yet, this model showed increased computational overhead due to additional prediction heads and transformer integration.

As per Table 7, it is proven that the proposed framework exhibits comparatively higher performance in object detection and recognition using DSWDCNN. However, the traditional works like [26] and [3] enhance the classification accuracy using the YOLOv8 and YOLO with Squeeze-and-Excitation (YOLO-SE), they suffer from limited fine-grained features and computational overhead, correspondingly. Likewise, by using the Mask

Region-based Convolutional Neural Network (MR-CNN), Vision Transform (ViT), and Hierarchical Context Fusion Network, the traditional methodologies show comparatively lower performance in object detection. Therefore, by effectively capturing the semantic relationships and explicit interactions among image and text using DSWDCNN, the proposed work accurately recognizes objects more than the traditional works.

Table 8. Ablation Study of the Proposed Framework

Model Configuration	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Full Proposed Model	98.8569	98.8412	98.9578	98.6598
Without MPST	96.7425	96.5184	96.8932	96.3057
Without MHCKEAD	95.9863	95.7421	96.1045	95.8876
Without BDOERT	95.2147	95.0068	95.3289	95.1024
Without DSWDCNN (Replaced with CNN)	93.4582	93.2475	93.5621	93.1986
Baseline Model (No Proposed Modules)	90.5628	90.2417	90.3265	90.7814

The ablation results clearly demonstrate that each component contributes to overall performance, with the removal of any module leading to a noticeable decline in accuracy and related metrics is shown in Table 8.

5. Conclusion

Here, the proposed work developed a robust model for NLP-based object detection and recognition through multimodal relation graph analysis. Here, the proposed EJIH-CLAHE was applied to enhance the image contrast. Further, the proposed MHCKEAD accurately detected the object regions with a 0.8712 dice coefficient. Moreover, by using the LS-GradCAM approach, the proposed work highlighted the object regions effectively. Additionally, the MPST efficiently constructed the multimodal relation graph within 3358ms, capturing the explicit relationships, logical dependencies, and semantic conflicts among image and text. Meanwhile, the multimodal analysis was carried out to model both image and textual features using the proposed MHCKEAD. Finally, the proposed work ensured accurate and robust object recognition using the proposed DSWDCNN with improved accuracy (98.8569%). Thus, the proposed work demonstrated superior performance in NLP-based object detection and recognition than the traditional works.

Practical Implication

Here, the robust multimodal fusion and efficient multimodal relation graph construction enable faster and meaningful semantic analysis for accurate object detection and recognition. Further, the proposed work is highly applicable in healthcare, surveillance, autonomous systems, industrial inspections, and document analysis.

Future Scope

In the future, the proposed work will be enhanced by integrating advanced methodologies for object detection under extreme environmental conditions.

Declarations

Conflict of Interest

The authors declare that they have no conflicts of interest regarding this work.

Data Availability

The data that support the findings of this study are not publicly available due to confidentiality agreements but are available from the corresponding author upon reasonable request.

Author Contributions

Bo Feng conceived and designed the study. Jiayi Yang collected the data and performed the experiments. Jingyue Xue analyzed the data and interpreted the results. All authors contributed to writing and revising the manuscript and approved the final version.

References

[1] Chang Y, Chen Y, Huang R, Yu Y. Enhanced image captioning with color recognition using deep learning methods. Appl Sci. 2022;12:209.

[2] Manakitsa N, Maraslidis GS, Moysis L, Fragulis GF. A review of machine learning and deep learning for object detection, semantic segmentation, and human action recognition in machine and robotic vision. Technologies. 2024;12:15.

[3] Wu T, Dong Y. YOLO-SE: Improved YOLOv8 for remote sensing object detection and recognition. Appl Sci. 2023;13:12977.

[4] Gui S, Song S, Qin R, Tang Y. Remote sensing object detection in the deep learning era: A review. Remote Sens. 2024;16:327.

[5] Ferreira LA, Meneghetti DDR, Lopes M, Santos PE. CAPTION: Caption analysis with proposed terms, image of objects, and natural language processing. SN Comput Sci. 2022;3(5):1–16.

[6] Turay T, Vladimirova T. Toward performing image classification and object detection with convolutional neural networks in autonomous driving systems: A survey. IEEE Access. 2022;10:14076–14119.

[7] Gu X, Lin T, Kuo W, Cui Y. Open-vocabulary object detection via vision and language knowledge distillation. arXiv. 2022;2104.13921.

[8] Amjoud AB, Amrouch M. Object detection using deep learning, CNNs and vision transformers: A review. IEEE Access. 2023;11:35479–35516.
[9] Arkin E, Yadikar N, Xu X, Aysa A, Ubul K. A survey: Object detection methods from CNN to transformer. Multimed Tools Appl. 2023;82:21353–21383.
[10] Rani S, Ghai D, Kumar S, Kantipudi MVVP, Alharbi AH, Ullah MA. Efficient 3D AlexNet architecture for object recognition using syntactic patterns from medical images. Comput Intell Neurosci. 2022;2022:7882924.
[11] Rani S, Lakhwani K, Kumar S. Three-dimensional object recognition and pattern recognition techniques: Related challenges. Multimed Tools Appl. 2022;1–44.
[12] Al Shamayleh AS, Adwan O, Alsharaiah MA, Hussein AH, Kharma QM, Eke CI. A comprehensive literature review on image captioning methods and metrics based on deep learning techniques. Multimed Tools Appl. 2024;1–50.
[13] Ondeng O, Ouma H, Akuon P. A review of transformer-based approaches for image captioning. Appl Sci. 2023;13:11103.
[14] Khan AS, Abbass MJ, Khan AH. Towards fault-aware image captioning: A review on integrating facial expression recognition and object detection. Sensors. 2025;25:5992.
[15] Abdulgalil HD, Basir OA. Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to multimodal large language models. Nat Lang Process J. 2025;12:1–20.
[16] Zhang Y. Intelligent edge caching and computing for scalable information systems. EAI Endorsed Trans Scalable Inf Syst. 2023;10(5).
[17] Jiang T, Li C, Yang M, Wang Z. An improved YOLOv5s algorithm for object detection with an attention mechanism. Electronics. 2022;11:2494.
[18] Zang Y, Li W, Han J, Zhou K, Loy CC. Contextual object detection with multimodal large language models. arXiv. 2024;2305.18279.
[19] Duhayyim MA, Alazwari S, Mengash HA, Marzouk R, Alzahrani JS, Mahgoub H, Althukair F, Salama AS. Metaheuristics optimization with deep learning enabled automated image captioning system. Appl Sci. 2022;12:7724.
[20] Alnashwan RO, Chelloug SA, Almalki NS, Issaoui I, Motwakel A, Sayed A. Lighting search algorithm with convolutional neural network-based image captioning system for natural language processing. IEEE Access. 2023;11:142643–142651.
[21] Gudivaka RK. Enhancing 3D vehicle recognition with AI: Integrating rotation awareness into aerial viewpoint mapping for spatial data. J Curr Sci Humanit. 2022;10(1):7–21.
[22] Rinaldi AM, Russo C, Tommasino C. Automatic image captioning combining natural language processing and deep neural networks. Results Eng. 2023;18:101107.
[23] Ricci R, Melgani F, Marcato J, Gonçalves WN. NLP-based fusion approach to robust image captioning. IEEE J Sel Top Appl Earth Obs Remote Sens. 2024;17:11809–11822.
[24] Al Malla MA, Jafar A, Ghneim N. Image captioning model using attention and object features to mimic human image understanding. J Big Data. 2022;9:1–16.
[25] Xu S, Zheng S, Xu W, Xu R, Wang C, Zhang J, Teng X, Li A, Guo L. HCF-Net: Hierarchical context fusion network for infrared small object detection. arXiv. 2024;2403.10778.
[26] Liu C, Ma X, Yang X, Zhang Y, Dong Y. COMO: Cross-Mamba interaction and offset-guided fusion for multimodal object detection. arXiv. 2024;2412.18076.
[27] Ikram S, Bajwa IS, Ikram A, Abdullah-Al-Wadud M, Haleema P. A transformer-based multimodal object detection system for real-world applications. IEEE Access. 2025;13:29162–29176.