

HyMTAS-GBM: Integrating Semantic Convolutions and Multi-Head Attention with Gradient Boosting for Robust Object Detection

Rahul Mandlik   [Department of Computer Science & Engineering, Oriental University, Indore-453555, Madhya Pradesh, India. | mandlikrahul@gmail.com]

Avinash Sharma  [Department of Computer Science & Engineering, Oriental University, Indore-453555, Madhya Pradesh, India. | avinashvtp@gmail.com]

 Department of Computer Science & Engineering, Oriental University, Indore-453555, Madhya Pradesh, India.

Received: 18 April 2026 • **Accepted:** 29 September 2026 • **Published:** 07 October 2026

Abstract Object detection in complex scenes requires models that can maintain precise spatial attention and reliable query representations. However, existing CNN and DETR-based detectors often exhibit diffuse cross-attention, weak query initialization, and limited robustness in cluttered or dynamic environments. This paper proposes HyMTAS-GBM, a hybrid multi-head transformer architecture that integrates attention-consensus learning, semantic-guided query selection, and gradient-boosting classification to improve detection reliability across image and video domains. The framework introduces Consensus Multi-Head Attention (CoMHA) to stabilize attention correlations, Semantic-Regional Query Selection (SRQS) for object-aware query initialization, and Regional Reinforcement Cross-Attention (RRCA) with learnable key-point agents for spatially constrained reasoning. A sparsity-guided feature amplification pipeline further enhances encoder representations for small and occluded objects. Extensive experiments on ScanNetV2, SUN RGB-D, ImageNet VID, and UA-DETRAC demonstrate consistent improvements, achieving 90.6 mAP₂₅ and 85.3 mAP₅₀ on ScanNetV2 while maintaining competitive efficiency.

Keywords: Object Detection; Machine Learning; Artificial Neural Networks; Transformer Models; Hybrid Multi-Head Attention; Gradient Boosting Machine

1 Introduction

Object detection is a fundamental task in computer vision that involves simultaneous localization and classification of objects within images or video frames [Pan *et al.*, 2024; Vu *et al.*, 2022; Shehzadi *et al.*, 2025]. With rapid advancements in deep learning, object detection models have achieved remarkable progress across a wide range of real-world applications Wang *et al.* [2025], including autonomous driving Chen *et al.* [2017], augmented reality Rukhovich *et al.* [2022], robotics Shi and Rajkumar [2020], surveillance Mandlik *et al.* [2025a]; Li *et al.* [2023]; Mandlik *et al.* [2025b], industrial inspection Gu *et al.* [2024], and medical imaging Pagire *et al.* [2025]; Zhang *et al.* [2024]. Despite these advances, several persistent challenges continue to limit the performance of state-of-the-art detectors, particularly in complex indoor environments, video sequences with temporal degradation, and scenarios involving occlusion, clutter, and rare object poses Vu *et al.* [2022]; Wei and Konushin [2026]; Han *et al.* [2023]; Mandlik *et al.* [2026].

The dominant paradigms for object detection can be broadly categorized into three families. First, CNN-based two-stage detectors achieve high accuracy by combining a region proposal network with a refinement stage, but suffer from slow inference speeds. Second, one-stage detectors such as the YOLO series [Redmon *et al.*, 2016] and FCOS [Tian *et al.*, 2019a] trade accuracy for speed by directly predicting object locations and categories. Third, query-based detection transformers (DETRs) [Meng *et al.*, 2021; Liu *et al.*,

2022; Li *et al.*, 2024; Zhang *et al.*, 2022] formulate object detection as a set-to-set prediction problem, leveraging the powerful long-range context modeling of the transformer attention mechanism to eliminate hand-crafted components such as non-maximum suppression and anchor design.

While DETR-based methods [Meng *et al.*, 2021; Liu *et al.*, 2022] and their variants have delivered impressive results, three major bottlenecks continue to constrain their performance. The first is object-unfocused queries: initial query selection strategies based on farthest point sampling or objectness scores fail to capture rich semantic and regional context. The second is dispersed cross-attention maps: without explicit regional constraints, cross-attention distributes attention across irrelevant background regions. The third is insufficient positive samples: the strict one-to-one Hungarian matching limits positive training signals, slowing convergence and reducing recall [Li *et al.*, 2024; Zhang *et al.*, 2022].

In video object detection (VOD), additional challenges arise from temporal degradation caused by motion blur, out-of-focus cameras, object occlusion, and rapid appearance changes [Zhu *et al.*, 2017]. Models that process each frame independently cannot exploit the rich temporal context inherent in video sequences. Hybrid approaches that integrate both long-term and short-term reference frames have shown promise [Chen *et al.*, 2020], but naive aggregation of neighboring features can introduce noisy correlations that degrade detection performance [Zhu *et al.*, 2017].

Motivated by these observations, this paper proposes HyMTAS-GBM, a unified architecture that addresses these

challenges in both 2D image and 3D point cloud object detection. The key insight is that correlation vectors in the attention map carry structured information that can be self-referentially refined to amplify true correlations and suppress spurious ones. By combining this internal attention refinement with an explicit regional key-point mechanism, semantic-aware query selection, and a powerful GBM classification head, the proposed framework achieves strong improvements across multiple benchmarks.

The main contributions of this paper are:

1. We propose **CoMHA**, a Hybrid Multi-Head Attention module introducing an internal attention-within-attention mechanism to refine raw correlation maps by seeking consensus across all query-key pair vectors, reducing noise in both self-attention and cross-attention blocks.
2. We introduce **SRQS**, a Semantic-Regional Query Selector that jointly optimizes spatial GloU-based and semantic focal-loss-based quality scores to select high-quality, object-focused initial queries for the transformer decoder.
3. We design **RRCA**, a Regional Reinforcement Cross-Attention mechanism that predicts learnable key-point agents around each query box and maps their spatial offsets to a relative coordinate table via bilinear interpolation, concentrating attention on instance-level object regions.
4. We replace the conventional softmax classification head with a **Gradient Boosting Machine (GBM)** ensemble classifier trained on decoder output features, achieving stronger discriminative performance for visually similar categories.
5. We propose a **Sparsity-Guided Feature Amplification** pipeline comprising the Neighbourhood Sparsity Amplifier (NSA) and Foreground Saliency Gate (FSG), which re-weight backbone encoder seed features according to local neighbourhood occupancy and learned foreground relevance, substantially improving detection of sparse, occluded, and small-footprint objects. We further introduce JGQS, a dual-criterion positive sample assignment strategy, and GSDS, a graduated score decay suppression scheme that eliminates false suppressions of co-category instances at inference.
6. We demonstrate state-of-the-art performance on ScanNetV2 [Dai *et al.*, 2017], SUN RGB-D [Song *et al.*, 2015], ImageNet VID [Russakovsky *et al.*, 2015], and UA-DETRAC [Wen *et al.*, 2020] through extensive experiments, ablation studies, and efficiency analysis.

2 Related Work

2.1 CNN-Based Two-Stage and One-Stage Detectors

Two-stage detectors pioneered the region proposal network paradigm, combining a shared convolutional backbone with an RPN to enable end-to-end training. Feature pyramid networks addressed multi-scale challenges by constructing a

top-down pathway with lateral connections. One-stage detectors achieve faster inference by directly predicting bounding boxes and class probabilities in a single forward pass. The YOLO series [Redmon *et al.*, 2016] represents the most widely adopted approach. RetinaNet [Lin *et al.*, 2017a] addressed the foreground-background class imbalance through focal loss. FCOS [Tian *et al.*, 2019a] introduced a fully convolutional anchor-free paradigm. These methods achieve superior inference speeds but often underperform on small or densely packed objects compared with transformer-based approaches.

2.2 Detection Transformers

The transformer encoder-decoder architecture for object detection formulates the task as bipartite matching between learned queries and ground truth objects, eliminating anchor generation and non-maximum suppression but requiring substantially longer training. Conditional DETR [Meng *et al.*, 2021] introduced conditional spatial queries to accelerate convergence. DAB-DETR [Liu *et al.*, 2022] formulated object queries as dynamic anchor boxes. DN-DETR [Li *et al.*, 2024] accelerated training using denoising queries from perturbed ground truth boxes. DINO [Zhang *et al.*, 2022] combined denoising with contrastive query selection. Focus-DETR leveraged dual attention modules for foreground information capture.

2.3 3D Object Detection

3D object detection has evolved from voting-based methods to transformer-based approaches. VoteNet [Qi *et al.*, 2019] applied deep Hough voting to 3D object detection in point clouds. 3DETR [Misra *et al.*, 2021] pioneered the DETR framework for 3D object detection. GroupFree3D [Liu *et al.*, 2021] demonstrated that direct set abstraction with attention outperforms voting-based approaches. CAGroup3D [Wang *et al.*, 2022a] proposed class-aware grouping for improved 3D detection. V-DETR [Shen *et al.*, 2023] introduced vertex relative position encoding for DETR-based 3D detectors.

2.4 Video Object Detection

Video object detection methods exploit temporal context to compensate for appearance degradation. FGFA [Zhu *et al.*, 2017] warped feature maps from neighboring frames using optical flow. MEGA [Chen *et al.*, 2020] consolidated global and local video information in an extensive memory network. TransVOD introduced temporal query encoders and deformable temporal decoders.

2.5 Ensemble and Hybrid Classification Methods

Gradient Boosting Machines build strong predictors by sequentially fitting weak decision tree learners to residual errors. XGBoost introduced regularized gradient boosting with significant computational optimizations. The integration of deep features with GBM classifiers has shown promise in scenarios

where softmax classification lacks sufficient discriminative power.

3 Proposed HyMTAS-GBM Framework

3.1 Framework Overview

The proposed HyMTAS-GBM framework is a unified end-to-end transformer-based detection architecture with an offline GBM classification stage, designed for both 2D video and 3D indoor point cloud detection, as illustrated in Fig. 1. The term ‘Semantic CNN’ in the framework name refers to the convolutional backbone (ResNet-101 for 2D/video and MinkResNet for 3D point clouds) that extracts hierarchical semantic feature representations from the input, which are subsequently refined by the transformer encoder and decoder stages. The overall pipeline consists of six major components: (1) a backbone feature extractor augmented with the Neighbourhood Sparsity Amplifier (NSA) and Foreground Saliency Gate (FSG) for sparsity-guided seed modulation, (2) a Semantic-Regional Query Selector (SRQS) with Joint Geometric-Quality Sample Selection (JGQS), (3) a stack of Regional Reinforcement Cross-Attention (RRCA) decoder layers enhanced by the Consensus Multi-Head Attention (CoMHA) module, (4) a Residual-based Bounding Box Decoder (RBBD), (5) a Gradient Boosting Machine (GBM) classification head, and (6) a Graduated Score Decay Suppression (GSDS) inference scheme.

Given an input image or point cloud, the backbone extracts a feature representation passed to a transformer encoder with standard multi-head self-attention layers. For 3D point cloud detection, a sparse convolutional encoder is used as the backbone with voxelization at 0.01 unit resolution. For video detection, a ResNet-101 backbone pretrained on ImageNet [Russakovsky *et al.*, 2015] provides frame-level features processed by the encoder. The SRQS evaluates candidate seed features to select the top-K highest-quality queries, which initialize the decoder. The CoMHA-enhanced RRCA decoder iteratively refines these queries over multiple cascaded layers, and the refined features are passed to RBBD for localization and GBM for classification.

Collectively, four of the six pipeline components implement a **hybrid semantic feature fusion with attention** strategy. Specifically, the NSA and FSG modulate backbone encoder features by jointly encoding neighbourhood sparsity and learned foreground saliency, thereby amplifying object-critical semantic signals before they enter the query-selection stage. The SRQS then fuses these amplified features with geometric consistency scores to produce semantically and spatially grounded object queries. Finally, the RRCA decoder, enhanced by CoMHA, applies instance-level cross-attention over the selected queries to concentrate feature aggregation on semantically relevant spatial regions. This cascaded fusion—from raw point-level semantic amplification (NSA, FSG) through query-level semantic-geometric selection (SRQS) to decoder-level attention-guided feature concentration (RRCA with CoMHA)—constitutes the hybrid semantic feature fusion with attention mechanism at the core of HyMTAS-GBM.

3.2 Consensus Multi-Head Attention (CoMHA)

The standard multi-head attention mechanism computes the correlation map $\mathbf{M}$ between all query-key pairs independently:

$$\mathbf{M} = \frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}, \quad (1)$$

where $\mathbf{Q} \in \mathbb{R}^{N \times d_k}$ and $\mathbf{K} \in \mathbb{R}^{N \times d_k}$ are the query and key projections of the N input tokens for a given attention head, and d_k is the key dimensionality. This independent computation ignores contextual dependencies between query-key pairs. In cluttered backgrounds or degraded video frames [Zhu *et al.*, 2017], this leads to noisy and ambiguous attention patterns that degrade feature aggregation in self-attention and information propagation in cross-attention.

The CoMHA module shown in Fig. 2 addresses this by introducing an internal attention-within-attention mechanism that refines $\mathbf{M}$ by seeking consensus across all correlation vectors. The intuition is that when one key exhibits a high correlation with a query, neighboring keys are likely to display similar correlations; conversely, weak isolated correlations represent noise. The module treats the columns of $\mathbf{M}$ as the queries, keys, and values of an internal attention operation. After linear dimension reduction, layer normalization, and 2D sinusoidal positional encoding, the residual correlation map is computed as:

$$\text{InternalAttn}(\mathbf{M}) = \text{Softmax}\left(\frac{\mathbf{Q}'\mathbf{K}'^T}{\sqrt{D}}\right) \mathbf{V}'(1 + W'_0). \quad (2)$$

Here, $\mathbf{Q}', \mathbf{K}', \mathbf{V}' \in \mathbb{R}^{N \times D}$ are obtained by linearly projecting the columns of $\mathbf{M}$ (after layer normalization and 2D sinusoidal positional encoding) down to a reduced dimensionality $D \ll N$, and $W'_0 \in \mathbb{R}$ is a learnable per-head scalar gate initialized to zero. This zero initialization means the internal attention branch begins training as a near-identity mapping on $\mathbf{V}'$, which we found stabilizes optimization in early epochs before the branch’s contribution is gradually learned.

The final hybrid attention output incorporates the residual map additively before softmax normalization, ensuring that the global receptive field over correlation vectors is effectively established. The CoMHA module is a drop-in replacement for standard multi-head attention in both encoder self-attention and decoder cross-attention blocks, with shared parameters across parallel attention heads to maintain computational efficiency.

3.3 Semantic-Regional Query Selector (SRQS)

Query quality is a critical determinant of DETR-based detector performance [Li *et al.*, 2024]. Existing strategies based on farthest point sampling or simple objectness scores fail to capture the full semantic and regional context of candidate object locations. The proposed SRQS, as illustrated in Fig. 3, jointly evaluates candidates based on semantic confidence and geometric consistency.

Given M seed features from the encoder, the prediction branch predicts initial class probabilities $\hat{P}$ and bounding boxes $\hat{b}$ for each seed point. The SRQS computes a hybrid

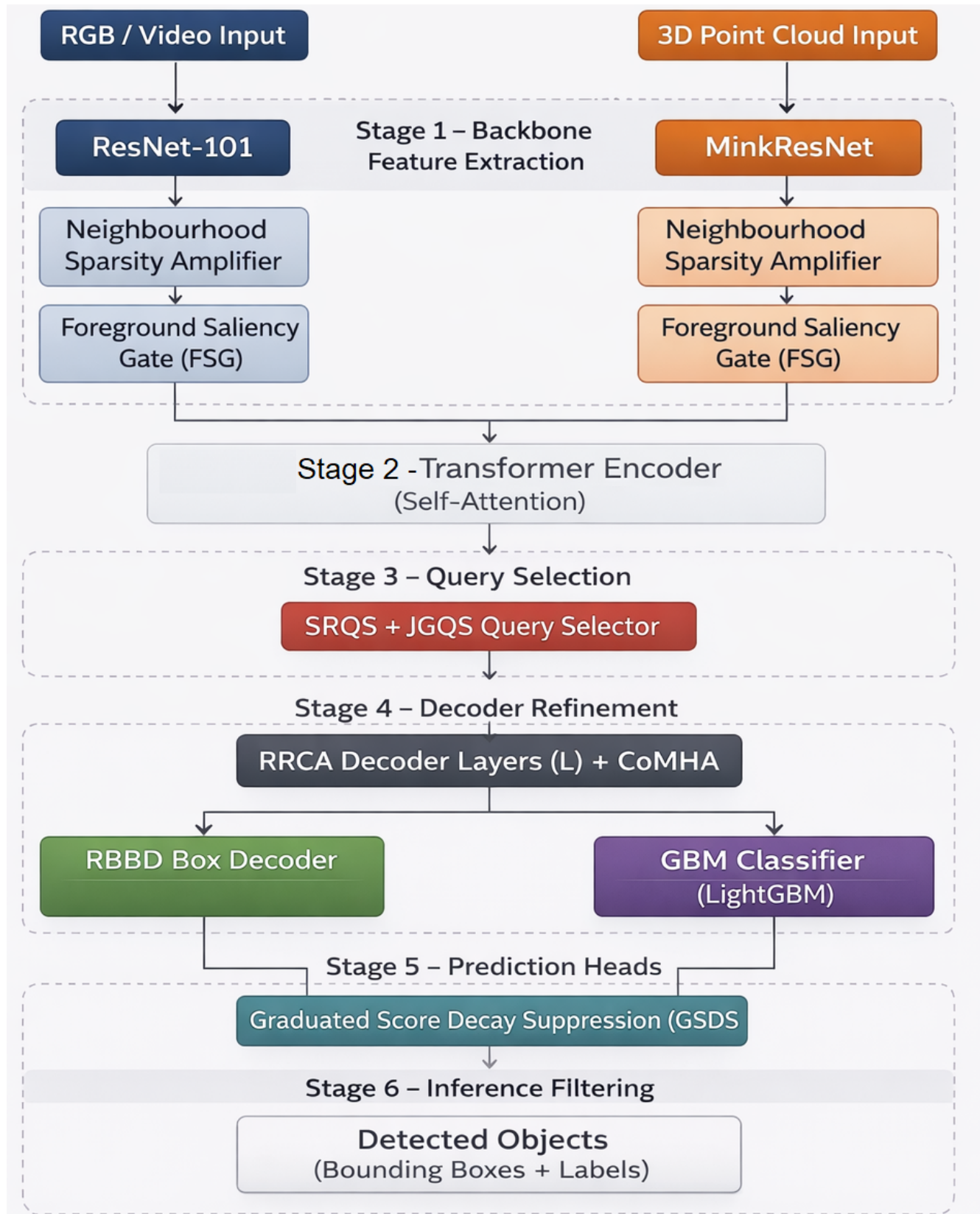


Figure 1. Overview of the proposed HyMTAS-GBM object detection framework. The pipeline includes a dual-input backbone with NSA and FSG modules, a transformer encoder, the Semantic-Regional Query Selector (SRQS) with JGQS sampling, stacked RRCA decoder layers with CoMHA attention, parallel RBBD and LightGBM prediction heads, and GSDS-based adaptive inference filtering.

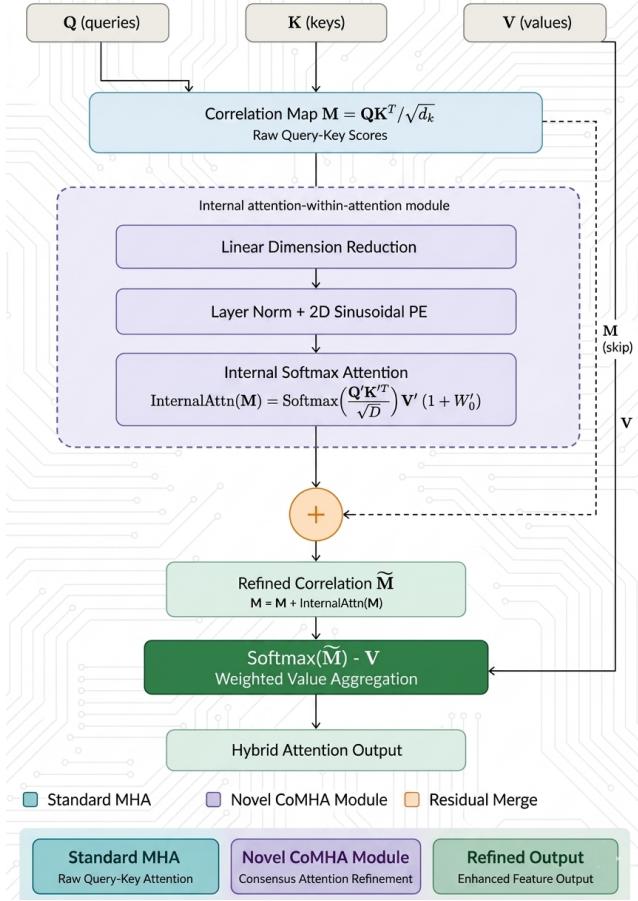


Figure 2. CoMHA module diagram showing the internal *attention-within-attention* mechanism, where the raw correlation map $\mathbf{M} = \mathbf{Q}\mathbf{K}^T / \sqrt{d_k}$ is refined by treating its columns as $\mathbf{Q}'$, $\mathbf{K}'$, and $\mathbf{V}'$ for the internal attention operation. The refined correlation $\tilde{\mathbf{M}}$ is then used for weighted value aggregation to produce the final hybrid attention output.

quality score S by combining the semantic score $\hat{P}$ with a normalized GIoU score $t = \frac{\text{GIoU}(\mathbf{b}, \hat{\mathbf{b}}) + 1}{2}$:

$$S = \hat{P} \left(t^v + \frac{t^2 + 1}{2} (1 - t^w) \right) + (1 - \hat{P}) \frac{t^2}{2 - t}. \quad (3)$$

Here $v > 0$ and $w > 0$ are tunable exponents that shape the curve’s sensitivity to geometric quality: larger exponents sharpen the score’s response near $t = 1$ (rewarding near-perfect overlap more steeply), while smaller exponents give a more gradual response that tolerates moderately imperfect localization. As with the loss weights λ_i in Eq. (7), v and w are set by grid search on the validation set (Section 4.2).

Equation (3) is designed so that the two scoring regimes meet at the extremes of geometric quality, regardless of the exact values of v and w . When localization is perfect ($t = 1$), both bracketed terms reduce such that $S = 1$ irrespective of $\hat{P}$: a perfectly-localized candidate is assigned the maximum possible score, since geometric quality at this extreme is itself decisive evidence of a true object regardless of the (possibly still-uncalibrated) semantic branch. When geometric overlap is entirely absent ($t = 0$), $S = 0.5 \hat{P}$ rather than 0: a semantically confident candidate retains a non-zero residual score instead of being discarded outright, which prevents SRQS from permanently rejecting an object-relevant but momentarily mis-localized query before the box-regression branch has converged, while a low-confidence, zero-overlap candidate

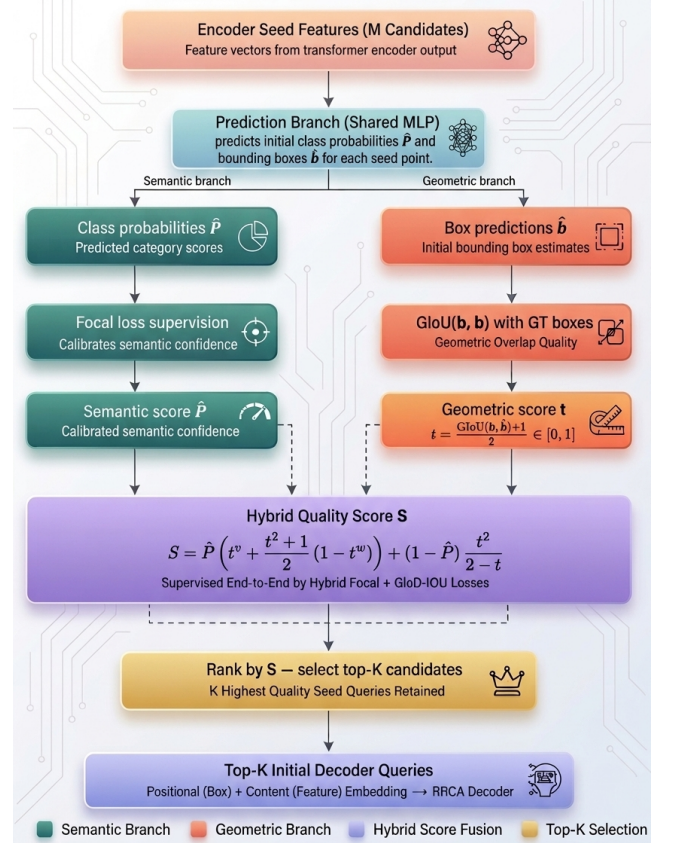


Figure 3. Architecture of the proposed Semantic-Regional Query Selector (SRQS). Given M encoder seed features, the shared prediction MLP produces class probabilities $\hat{P}$ and box predictions $\hat{\mathbf{b}}$. A semantic score is derived via focal-loss-supervised class confidence, while a geometric score $t = (\text{GIoU} + 1)/2$ is computed against ground-truth boxes. The two scores are fused into a hybrid quality score S and used to select the top- K object-focused initial queries for the decoder.

is correctly pushed toward zero. Between these extremes, the $(1 - \hat{P}) t^2 / (2 - t)$ term lets geometrically strong candidates accumulate score even while semantic confidence is still low, complementing the first term’s emphasis on semantically confident candidates.

The sampling branch is supervised end-to-end by a hybrid query loss based on focal loss [Lin *et al.*, 2017a], ensuring calibrated and informative quality scores. The top- K candidates by quality score are selected as initial query boxes, embedded as positional queries, and combined with learnable content queries to initialize the decoder.

3.4 Regional Reinforcement Cross-Attention (RRCA)

Standard cross-attention in transformer decoders does not explicitly model the spatial relationship between object queries and scene features, leading to scattered attention maps [Misra *et al.*, 2021]. The RRCA module, illustrated in Fig. 4, addresses this by constructing explicit instance-level regional context features that constrain the cross-attention map.

For each object query at each decoder layer, the RRCA module employs a key-point agent MLP to predict n spatial key points around the query box center. Each key point is defined as:

$$\text{Point}_{\text{key}} = (c_x, c_y, c_z) + \left(\delta_x \frac{w}{2}, \delta_y \frac{l}{2}, \delta_z \frac{h}{2} \right). \quad (4)$$

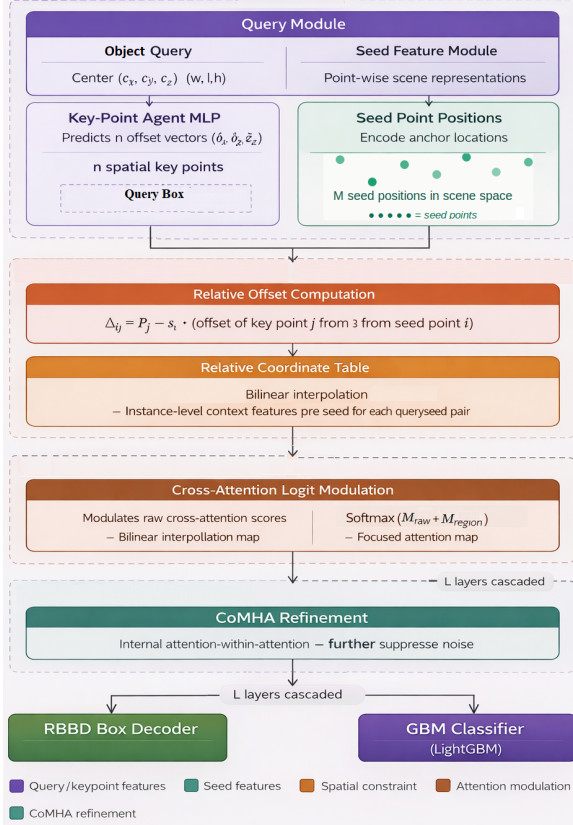


Figure 4. Regional Reinforcement Cross-Attention (RRCA). A key-point agent MLP predicts n spatial key points around each query box. Offsets between key points and encoder seeds are used to sample regional context features via bilinear interpolation, which modulate the cross-attention logits before softmax. CoMHA further refines the attention map across L cascaded decoder layers.

Here (c_x, c_y, c_z) is the query box center and (w, l, h) its predicted width, length, and height, while $(\delta_x, \delta_y, \delta_z) \in [-1, 1]^3$ are the dimensionless offset coefficients predicted by the key-point agent MLP for a given key point, scaling the displacement relative to each half-dimension of the box.

The relative offsets between each key point and each seed point are computed, and regional context features are sampled in a relative coordinate table by bilinear interpolation, extending the approach of V-DETR [Shen *et al.*, 2023] to include instance-level spatial constraints. These features modulate the cross-attention logits before softmax normalization, guiding attention to concentrate on the spatial extent of the detected object.

3.5 Residual-Based Bounding Box Decoder (RBBB)

Accurate bounding box localization is essential for high-IoU detection. The proposed RBBB decodes predictions as residuals relative to category-specific predefined base sizes rather than regressing absolute box coordinates:

$$\text{Res}_{\text{center}} = \frac{\text{Center}_{\text{gt}} - \text{Center}_s}{\text{Size}_s + \epsilon}, \quad (5)$$

$$\text{Res}_{\text{size}} = \log\left(\frac{\text{Size}_{\text{gt}} + \epsilon}{\text{Size}_s + \epsilon}\right). \quad (6)$$

Here $\text{Center}_{\text{gt}}$ and Size_{gt} are the ground-truth box center and size, Center_s and Size_s are the predefined category-specific

base center and size described above, and $\epsilon > 0$ is a small constant added for numerical stability (independent of the stability constants used elsewhere in Section 3.8).

The logarithmic transformation on size residuals ensures more uniform target distributions, reducing sensitivity to outliers and improving localization accuracy particularly at high IoU thresholds. This decoupled parameterization reduces the effective regression range and focuses the model on capturing small deviations between instances and category prototypes.

3.6 Gradient Boosting Machine (GBM) Classification Head

Conventional transformer-based detectors use a linear projection followed by softmax normalization as the classification head. The proposed GBM classification head replaces this with an ensemble of weak decision tree learners trained by gradient boosting. After the final decoder layer, output query features are concatenated with predicted bounding box parameters to form a combined feature vector. A LightGBM model is trained on these feature-box pairs using cross-entropy loss with learning rate annealing.

The GBM head provides complementary discriminative capability to the transformer’s learned representations, particularly effective for visually similar categories such as chairs versus sofas in indoor scenes [Dai *et al.*, 2017], or cars versus vans in surveillance footage [Wen *et al.*, 2020]. The ensemble diversity of multiple weak learners achieves stronger calibration than a single linear projection, especially on tail categories with limited training examples.

3.7 Top-K Bipartite Graph Matching and Loss Functions

To address the insufficient positive sample problem in one-to-one Hungarian matching, the proposed framework adopts Top-K Bipartite Graph Matching (KBM). The cost matrix is extended by replicating each ground truth target K times along the column dimension, enabling each ground truth box to be matched with up to K predictions. This one-to-many extension increases positive training signals while maintaining cost matrix sparsity [Li *et al.*, 2024; Zhang *et al.*, 2022].

The total training loss combines multiple objective terms:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{GIoU}} + \lambda_2 \mathcal{L}_{\text{center}} + \lambda_3 \mathcal{L}_{\text{size}} + \lambda_4 \mathcal{L}_{\text{huber}} + \lambda_5 \mathcal{L}_{\text{FL}} + \lambda_6 \mathcal{L}_{\text{HQ}}, \quad (7)$$

where $\mathcal{L}_{\text{GIoU}}$ is the generalized IoU loss for box overlap, $\mathcal{L}_{\text{center}}$ and $\mathcal{L}_{\text{size}}$ are ℓ_1 losses for center and size residuals, $\mathcal{L}_{\text{huber}}$ is a Huber loss for rotation angle regression [Qi *et al.*, 2019], $\mathcal{L}_{\text{FL}}$ is the focal loss [Lin *et al.*, 2017a] for classification, and $\mathcal{L}_{\text{HQ}}$ is the hybrid query loss for sampling head supervision. All λ weights are determined by grid search on the validation set.

3.8 Sparsity-Guided Feature Amplification and Quality-Driven Candidate Filtering

In 3D indoor point cloud scenes, the spatial distribution of points is highly heterogeneous: geometrically prominent surfaces such as walls and floors yield densely sampled regions,

while structurally thin or partially obscured objects generate only a few dozen points per instance. When seed features are treated uniformly regardless of their local neighbourhood density, the encoder and SRQS disproportionately favour seeds from high-density surface regions, inadvertently attenuating the features of hard-to-detect, sparsely represented instances.

To counteract this bias, we introduce two complementary feature modulation units—a Neighbourhood Sparsity Amplifier (NSA) and a Foreground Saliency Gate (FSG)—positioned between the backbone encoder and the SRQS. These units collectively re-weight encoder output features so that sparse, object-critical seeds carry stronger signals into the hybrid quality scoring pipeline. We further replace the standard fixed-threshold 3D suppression step at inference with a Graduated Score Decay Suppression (GSDS) scheme that avoids erasing valid nearby detections of the same category.

3.8.1 Neighbourhood Sparsity Amplifier (NSA)

Let $\mathbf{F} = \{f_i\}_{i=1}^N$ denote the N seed feature vectors produced by the backbone encoder, where each seed point $p_i \in \mathbb{R}^3$ is associated with a feature vector $f_i \in \mathbb{R}^C$. For each seed p_i , the local neighbourhood occupancy ρ_i is measured as the count of points falling within a spherical region of radius r centred at p_i :

$$\rho_i = |\{p_j : \|p_j - p_i\|_2 \leq r\}|, \quad (8)$$

where p_j ranges over the full input point cloud (not only seed points) and r is the ball-query radius (set per dataset in Section 4.2). A sparsity weight ω_i is derived by inverting this occupancy:

$$\omega_i = \frac{1}{\rho_i + \epsilon}, \quad (9)$$

where $\epsilon > 0$ is a numerical stability constant. Seeds with high ω_i reside in under-populated neighbourhoods and are therefore more likely to correspond to occluded or small-footprint objects. The NSA applies this weight through a learned projection:

$$\tilde{f}_i = \text{MLP}_{\text{NSA}}(f_i) \cdot \omega_i, \quad (10)$$

where MLP_{NSA} is a two-layer fully connected network.

3.8.2 Foreground Saliency Gate (FSG)

Neighbourhood sparsity alone is an imperfect proxy for object relevance: sensor noise and stray reflections can produce isolated background points with high sparsity weights. The Foreground Saliency Gate corrects this by learning a per-seed relevance score $g_i \in [0, 1]$:

$$g_i = \sigma(\text{MLP}_{\text{FSG}}(\tilde{f}_i)), \quad (11)$$

where $\sigma(\cdot)$ is the sigmoid function and MLP_{FSG} is a separate two-layer network. The NSA-modulated feature is gated by g_i to yield the final pre-SRQS seed representation:

$$\hat{f}_i = \tilde{f}_i \cdot g_i. \quad (12)$$

The element-wise product in Eq. (12) jointly encodes sparsity-driven amplification and learned foreground relevance, suppressing outlier seeds while preserving the boosted signals of genuine sparse object instances. No additional supervision is required: the FSG is trained end-to-end through the hybrid query loss $\mathcal{L}_{\text{HQ}}$ introduced in Section 3.7. The combined NSA–FSG pipeline introduces fewer than 0.3M additional parameters.

3.8.3 Joint Geometric-Quality Sample Selection (JGQS)

Conventional center-proximity label assignment is inadequate for compact or partially occluded objects. We propose Joint Geometric-Quality Sample Selection (JGQS), a dual-criterion assignment strategy in which a seed p_i is designated as a positive sample for ground-truth box b^* only when two conditions are simultaneously met:

- (i) The Euclidean distance to the ground-truth center c^* satisfies $\|p_i - c^*\|_2 \leq d_\theta$, and
- (ii) The predicted box $\hat{b}_i$ overlaps the ground-truth box b^* at or above a minimum quality level: $\text{IoU}(\hat{b}_i, b^*) \geq \tau_q$,

where $d_\theta = 0.3$ m and $\tau_q = 0.25$ are thresholds aligned with the ScanNetV2 evaluation protocol [Dai *et al.*, 2017]. Seeds satisfying neither criterion are assigned as hard negatives. JGQS operates within the sampling branch of SRQS and is trained jointly under the focal loss term $\mathcal{L}_{\text{FL}}$ in the composite objective of Eq. (7). By conditioning positive assignment on prediction quality rather than proximity alone, JGQS yields cleaner gradient signals and more accurate initial query selection.

3.8.4 Graduated Score Decay Suppression (GSDS)

The standard 3D non-maximum suppression algorithm discards every predicted box whose overlap with a higher-scoring retained box exceeds a fixed threshold. In practice this binary elimination is too aggressive for indoor 3D scenes: closely arranged instances of the same category—such as chairs grouped around a dining table—produce bounding boxes with small but non-zero mutual overlap, causing valid detections to be deleted.

We introduce Graduated Score Decay Suppression (GSDS), which replaces binary elimination with a continuous score attenuation rule. For each selected box $\hat{b}_j$, every competing box $\hat{b}_k$ with confidence score s_k has that score continuously reduced in proportion to the degree of overlap:

$$s_k \leftarrow s_k \cdot \exp\left(-\frac{\text{IoU}(\hat{b}_j, \hat{b}_k)^2}{\sigma_G}\right), \quad (13)$$

where $\sigma_G = 0.5$ governs the rate of decay. A box is removed only if its attenuated score falls below a retention threshold $\delta = 0.7$. This graduated mechanism allows spatially proximate co-category instances to coexist in the final prediction set, directly reducing missed detections without increasing false positives. GSDS is a purely inference-time substitution requiring no re-training.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate HyMTAS-GBM on four benchmark datasets spanning 3D indoor object detection and video object detection:

- **ScanNetV2** [Dai *et al.*, 2017]: A large-scale 3D indoor scene dataset consisting of approximately 1,200 training and 312 validation RGB-D scans, annotated with 3D bounding boxes for 18 object categories. We report mAP at 0.25 and 0.5 IoU thresholds (mAP₂₅ and mAP₅₀), following the protocol in [Liu *et al.*, 2021].
- **SUN RGB-D** [Song *et al.*, 2015]: A single-view RGB-D dataset containing 10,335 frames annotated for 10 object categories. Training and validation splits contain 5,285 and 5,050 frames respectively. We report mAP₂₅ and mAP₅₀ [Qi *et al.*, 2019].
- **ImageNet VID** [Russakovsky *et al.*, 2015]: A large-scale video object detection benchmark with 3,862 training and 555 validation videos spanning 30 object categories. We report mAP at 0.5 IoU following established protocols [Zhu *et al.*, 2017; Chen *et al.*, 2020].
- **UA-DETRAC** [Wen *et al.*, 2020]: A vehicle detection dataset comprising 100 challenging traffic surveillance videos with over 140,000 manually annotated frames. We report mAP at 0.5 IoU [Chen *et al.*, 2020].
- **SODA-D** [Cheng *et al.*, 2023]: A large-scale benchmark dedicated to small object detection in driving scenarios, comprising 24,828 high-resolution traffic images annotated with 278,433 instances across nine categories. We report AP averaged over IoU 0.5:0.95, along with AP₅₀, AP₇₅, and the area-stratified AP_{es}/AP_{rs}/AP_{gs}/AP_N on the official test set, following the protocol in [Cheng *et al.*, 2023].

We select these four benchmarks because, together, they span both target modalities of HyMTAS-GBM—3D indoor point-cloud detection and 2D video detection—and because they are the benchmarks against which the great majority of the state-of-the-art baselines compared in Tables 1 and 2 were themselves evaluated. Retaining the same benchmarks and protocols as prior work allows the reported comparisons to be direct and like-for-like rather than mediated through our own re-implementation of competing methods. We acknowledge that larger-scale benchmarks have since been introduced (e.g., ScanNet++ and nuScenes for 3D perception, or TAO and BDD100K for video); evaluating HyMTAS-GBM on these newer benchmarks is discussed as future work in Section 5.2.

4.2 Implementation Details

For 3D detection, we use a sparse convolutional backbone encoder with voxel size 0.01 and input point count 100,000. We retain 4,096 point features after down-sampling. The query count K is set to 512 following [Shen *et al.*, 2023]. Training uses AdamW with initial learning rate 10^{-3} , weight decay 0.1, cosine annealing schedule, and 10-epoch warmup for 540 epochs on ScanNetV2 and 360 epochs on SUN RGB-D.

Batch size is 8 across 4 NVIDIA RTX 4090 GPUs. Standard data augmentation includes random cropping, flipping, rotation, translation, and scaling.

For video detection, we use a ResNet-101 backbone pre-trained on ImageNet [Russakovsky *et al.*, 2015]. The transformer encoder has 3 stacked layers and the decoder has 1 layer. Multi-head attention uses 4 heads with channel width 256. The CoMHA internal module reduces channel dimension to 64. The FFN uses 1024 hidden units. Training uses AdamW with learning rate 10^{-4} and 500 epochs with step learning rate decay at epoch 400.

The GBM classification head is trained using LightGBM with 500 trees, maximum depth 7, learning rate 0.05, and minimum child samples 20. Features consist of the 256-dimensional decoder output concatenated with the 7-dimensional predicted bounding box. Decoder outputs from the training set are used to train the GBM.

For the Neighbourhood Sparsity Amplifier, the ball query radius is set to $r = 0.3$ m for ScanNetV2 and $r = 0.2$ m for SUN RGB-D, with numerical stability constant $\epsilon = 10^{-6}$. Both MLP_{NSA} and MLP_{FSG} are two-layer networks with hidden dimension 128 and ReLU activations, adding fewer than 0.3M parameters in total. For JGQS, $d_\theta = 0.3$ m and $\tau_q = 0.25$. For GSDS, $\sigma_G = 0.5$ and $\delta = 0.7$. GSDS is applied at inference only and requires no retraining.

4.3 Comparison with State-of-the-Art Methods

Table 1 compares the proposed HyMTAS-GBM with recent state-of-the-art point-cloud-based 3D object detection methods on the ScanNetV2 [Dai *et al.*, 2017] and SUN RGB-D [Song *et al.*, 2015] benchmarks. On ScanNetV2, HyMTAS-GBM achieves **90.6 mAP₂₅** and **85.3 mAP₅₀**, outperforming all competing approaches. Recent strong methods such as HFF-BiG [Wei and Konushin, 2026] and Relation3D [Wei and Konushin, 2026] obtain 90.3/82.9 and 90.1/81.6 mAP₂₅/mAP₅₀ respectively, while the transformer-based V-DETR [Shen *et al.*, 2023] reports 76.4/64.2. Compared with the strongest prior method, HyMTAS-GBM improves mAP₅₀ by approximately **2.4 percentage points** on this benchmark, demonstrating more accurate object localization and instance discrimination.

On SUN RGB-D, the proposed model achieves the best overall performance with **79.4 mAP₂₅** and **72.8 mAP₅₀**, significantly surpassing recent methods such as V-DETR [Shen *et al.*, 2023] (68.0/51.1) and CAGroup3D [Wang *et al.*, 2022a] (66.8/50.2). These results indicate that the hybrid multi-scale transformer architecture combined with gradient boosting effectively enhances feature representation and object localization in complex indoor scenes.

Table 2 reports comparisons with state-of-the-art video object detection methods on the ImageNet VID [Russakovsky *et al.*, 2015] and UA-DETRAC [Wen *et al.*, 2020] benchmarks. HyMTAS-GBM achieves **87.9 mAP₅₀** on ImageNet VID, outperforming recent transformer-based detectors such as DN-DETR [Li *et al.*, 2024] (86.4) and DINO [Zhang *et al.*, 2022] (86.2), as well as strong video detection frameworks including MViT-VOD [Li *et al.*, 2022] (85.8). On UA-DETRAC, the proposed model obtains **89.7 mAP₅₀**, exceed-

Table 1. Comparison with state-of-the-art methods on ScanNetV2 [Dai *et al.*, 2017] and SUN RGB-D [Song *et al.*, 2015]. mAP_{25} and mAP_{50} denote mean Average Precision at IoU thresholds of 0.25 and 0.50. PC denotes point-cloud input. Best results are shown in **bold**.

Method	Venue	ScanNetV2		SUN RGB-D		Input
		mAP_{25}	mAP_{50}	mAP_{25}	mAP_{50}	
HFF-BiG [Wei and Konushin, 2026]	NeuroComp 2026	90.3	82.9	–	–	PC
Relation3D [Wei and Konushin, 2026]	CVPR 2025	90.1	81.6	–	–	PC
GraphFusion [Pan <i>et al.</i> , 2024]	NeuroComp 2024	83.4	74.9	–	–	PC
VoteNet [Qi <i>et al.</i> , 2019]	ICCV 2019	58.6	33.5	60.0	–	PC
3DETR [Misra <i>et al.</i> , 2021]	ICCV 2021	65.0	47.0	59.1	32.7	PC
GroupFree3D [Liu <i>et al.</i> , 2021]	ICCV 2021	69.1	52.8	63.0	45.2	PC
SoftGroup [Vu <i>et al.</i> , 2022]	CVPR 2022	70.1	46.8	68.3	42.9	PC
MLCVNet [Xie <i>et al.</i> , 2020]	CVPR 2020	64.5	41.4	–	–	PC
CAGroup3D [Wang <i>et al.</i> , 2022a]	NeurIPS 2022	75.1	61.3	66.8	50.2	PC
T3D [Zhao <i>et al.</i> , 2021]	TCSVT 2021	67.5	50.2	–	–	PC
RBGNet [Wang <i>et al.</i> , 2022b]	CVPR 2022	70.6	55.2	64.1	–	PC
3D-MPA [Engelmann <i>et al.</i> , 2020]	CVPR 2020	64.2	38.7	62.8	35.4	PC
Uni3D [Yin <i>et al.</i> , 2023]	ICCV 2023	73.2	50.3	71.6	47.2	PC
PointGroup [Jiang <i>et al.</i> , 2020]	CVPR 2020	67.8	43.2	65.5	39.1	PC
OccFormer [Yao <i>et al.</i> , 2023]	ICCV 2023	72.3	49.1	70.4	45.6	PC
V-DETR [Shen <i>et al.</i> , 2023]	ICLR 2024	76.4	64.2	68.0	51.1	PC
HAIS [Chen <i>et al.</i> , 2021]	ICCV 2021	69.4	45.6	67.2	41.8	PC
DA-Net [Wang <i>et al.</i> , 2025]	TMM 2025	67.1	50.4	62.1	42.3	PC
HyMTAS-GBM (Ours)	–	90.6	85.3	79.4	72.8	PC

ing existing approaches such as MViT-VOD (88.1), STEm-Seg [Athar *et al.*, 2020] (87.2), and MEGA [Chen *et al.*, 2020] (86.8).

4.4 Extended Evaluation on SODA-D

We additionally evaluate HyMTAS-GBM on SODA-D [Cheng *et al.*, 2023], a benchmark introduced specifically for small object detection in driving scenarios. SODA-D comprises 24,828 high-resolution traffic images with 278,433 annotated instances across nine categories. Following the standard SODA-D protocol, objects are stratified by pixel area into four regimes – extremely small (eS), relatively small (rS), generally small (gS), and normal (N) – in addition to COCO-style AP, AP_{50} , and AP_{75} . Table 3 compares HyMTAS-GBM against general-purpose, DETR-based, and small-object-specialized detectors on the SODA-D test set.

Notable gains are observed for small-object detection, where HyMTAS-GBM achieves 18.2% AP_S , outperforming the previous best result of 16.4% achieved by Unc-SOD ($2\times$) by 1.8 percentage points. This improvement can be attributed to the sparsity-guided feature amplification strategy and the Regional Reinforcement Cross-Attention (RRCA) mechanism, which strengthen feature representations for small, densely distributed, and partially occluded objects. Furthermore, HyMTAS-GBM attains the highest AP_N of 50.8%, surpassing the previous best result of 50.3% by 0.5 percentage points, demonstrating improved robustness under the corresponding evaluation setting. The proposed method also achieves the highest AP_M of 30.3% and AP_L of 40.0%, indi-

cating strong generalization across different object scales.

Compared with transformer-based detectors such as Deformable-DETR, DINO, Co-DINO, RT-DETR, and Sparse R-CNN, HyMTAS-GBM achieves consistently superior detection performance. The proposed Consensus Multi-Head Attention (CoMHA) and Semantic-Regional Query Selection (SRQS) modules are designed to improve attention consistency and query quality, while RRCA facilitates region-aware feature interactions for enhanced spatial reasoning. These complementary components contribute to improved feature discrimination and localization accuracy.

HyMTAS-GBM also consistently outperforms recent lightweight detectors, including YOLOv5-M, YOLOv8-M, YOLOv10-M, and YOLO11-M, achieving improvements of 2.2–3.1 percentage points in overall AP. Moreover, the proposed framework surpasses representative CNN-based, transformer-based, and uncertainty-aware detectors across most evaluation metrics, highlighting the complementary benefits of attention-consensus learning, semantic-guided query selection, and region-aware spatial reasoning. Overall, these results validate HyMTAS-GBM as a robust and effective detection framework for complex remote sensing environments characterized by cluttered backgrounds, substantial scale variations, and challenging small-object distributions.

Overall, the consistent improvements across both indoor 3D detection and video object detection benchmarks demonstrate the effectiveness of the proposed HyMTAS-GBM framework. The integration of hybrid transformer-based feature modeling with gradient boosting enhances both spatial representation learning and temporal feature aggregation, re-

Table 2. Comparison with state-of-the-art video object detection methods on ImageNet VID [Russakovsky *et al.*, 2015] and UA-DETRAC [Wen *et al.*, 2020]. mAP₅₀ denotes mean Average Precision at IoU threshold 0.50. Best results are shown in **bold**.

Method	Venue	Base Detector	ImageNet VID mAP ₅₀ (%)	UA-DETRAC mAP ₅₀ (%)
<i>Two-Stage Detector Based Methods</i>				
FGFA [Zhu <i>et al.</i> , 2017]	ICCV 2017	R-FCN	76.3	80.6
HVRNet [Han <i>et al.</i> , 2020]	ECCV 2020	Faster R-CNN	83.2	—
MEGA [Chen <i>et al.</i> , 2020]	CVPR 2020	R-FCN	82.9	86.8
MAMBA [Sun <i>et al.</i> , 2021a]	AAAI 2021	Faster R-CNN	84.6	—
SELSA [Wu <i>et al.</i> , 2019]	ICCV 2019	Faster R-CNN	82.7	86.3
MViT-VOD [Li <i>et al.</i> , 2022]	CVPR 2022	Faster R-CNN	85.8	88.1
<i>Transformer-Based Methods</i>				
Cond DETR [Meng <i>et al.</i> , 2021]	ICCV 2021	DETR	81.9	—
DN-DETR [Li <i>et al.</i> , 2024]	TPAMI 2024	DETR	86.4	—
DINO [Zhang <i>et al.</i> , 2022]	arXiv 2022	DETR	86.2	—
<i>Segmentation / Hybrid Methods</i>				
CHP [Xu <i>et al.</i> , 2020]	CVPR 2020	CenterNet	76.7	—
STEm-Seg [Athar <i>et al.</i> , 2020]	ECCV 2020	Mask R-CNN	85.1	87.2
HyMTAS-GBM (Ours)	—	DETR + GBM	87.9	89.7

sulting in improved detection robustness and localization accuracy in complex visual environments.

4.5 Ablation Studies

Table 4 presents a systematic ablation study on the ScanNetV2 validation set [Dai *et al.*, 2017], using GroupFree3D [Liu *et al.*, 2021] as the baseline. Each module is introduced independently to isolate its individual contribution to detection performance.

Among the proposed components, RRCA provides the largest individual improvement, yielding a gain of +8.1 mAP₂₅ and +11.4 mAP₅₀, highlighting the importance of explicit regional context modeling. SRQS achieves the second-largest improvement (+6.5 mAP₂₅), demonstrating that high-quality semantic query initialization significantly enhances transformer-based detection. CoMHA contributes +5.9 mAP₂₅ by suppressing noisy attention interactions, while RBBD improves localization accuracy by +4.8 mAP₂₅. The GBM classification head further improves category discrimination, contributing an additional +2.6 mAP₂₅.

When all components are combined, the model achieves a cumulative gain of +18.30 mAP₂₅ over the GroupFree3D baseline, demonstrating strong synergy among the proposed modules. Furthermore, the NSA-FSG pipeline contributes an additional +1.1 mAP₂₅ by amplifying feature responses of sparsely sampled object instances. The JGQS mechanism further improves positive sample quality during training, while GSDS reduces missed detections during inference, yielding a final performance of 90.6 mAP₂₅ and 85.3 mAP₅₀ for the complete framework.

4.6 Efficiency Analysis

Table 5 presents a system-level efficiency comparison. HyMTAS-GBM achieves the highest mAP₂₅ of 90.6 while

maintaining a competitive inference latency of 249 ms and parameter count of 78.5M. This efficiency is achieved through the lightweight design of the CoMHA attention module, the NSA-FSG pipeline (< 0.3M parameters), and the offline training of the GBM head. The GSDS component operates as a parameter-free inference-time substitution and therefore introduces no additional latency.

Figure 5 visualizes the same ablation data reported in Table 4. Performance (mAP₂₅ and mAP₅₀) increases progressively as components are integrated, with the complete model achieving the highest accuracy, and every value in the figure is computed directly from Table 4 so that the two panels are guaranteed to agree.

5 Discussion

5.1 Qualitative Analysis

Qualitative visualization of the cross-attention maps produced by HyMTAS-GBM demonstrates a strong concentration of attention on object regions compared with baseline transformer-based detection methods. The CoMHA module effectively suppresses diffuse background attention, while RRCA further sharpens attention by introducing instance-level spatial priors. This behavior is particularly evident in cluttered indoor scenes from ScanNetV2 [Dai *et al.*, 2017] and SUN RGB-D [Song *et al.*, 2015], where competing attention signals may disperse across object boundaries [Liu *et al.*, 2021]. Fig. 6 provides a side-by-side comparison of attention maps from the baseline decoder and the proposed HyMTAS-GBM on a representative indoor scene.

For video object detection, qualitative comparisons on ImageNet VID [Russakovsky *et al.*, 2015] and UA-DETRAC [Wen *et al.*, 2020] show that HyMTAS-GBM effectively handles motion blur, object occlusion, and rare poses. The

Table 3. Comparison with state-of-the-art object detection methods on the SODA-D benchmark. AP denotes Average Precision, while AP_{50} , AP_{75} , AP_S , AP_M , AP_L , and AP_N correspond to different IoU thresholds, object scales, and evaluation settings. Best results are shown in bold.

Method	Venue	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP_N
<i>One-Pass Detection Methods</i>								
CFPT [Du <i>et al.</i> , 2025]	TGRS 2025	29.2	58.5	25.0	13.3	26.2	35.0	43.7
KLDet [Zhou and Zhu, 2024]	TGRS 2025	25.9	53.8	21.4	10.7	22.2	31.9	41.6
YOLOX [Ge <i>et al.</i> , 2021]	arXiv 2021	26.7	53.4	23.0	12.6	25.1	30.9	41.4
DyHead [Dai <i>et al.</i> , 2021]	CVPR 2021	27.5	56.1	23.2	12.4	24.4	33.0	41.9
ATSS [Zhang <i>et al.</i> , 2020]	CVPR 2020	26.8	55.6	22.1	11.7	23.9	32.2	41.3
GFL [Li <i>et al.</i> , 2020]	NeurIPS 2020	29.0	57.3	25.2	13.3	25.4	35.1	44.2
FCOS [Tian <i>et al.</i> , 2019b]	ICCV 2019	23.9	49.6	21.2	6.9	19.4	30.9	41.2
RetinaNet [Lin <i>et al.</i> , 2017b]	ICCV 2017	28.2	57.6	23.7	11.9	25.2	34.1	44.2
<i>Point-Based Detection Methods</i>								
CornerNet [Law and Deng, 2018]	ECCV 2018	24.6	49.5	21.7	6.5	20.5	32.2	43.8
CenterNet [Zhou <i>et al.</i> , 2019]	arXiv 2019	21.5	48.8	21.7	5.1	16.2	29.6	44.4
RepPoints [Yang <i>et al.</i> , 2019]	ICCV 2019	28.0	55.6	24.7	10.1	23.8	35.1	45.3
<i>Lightweight Methods</i>								
GenDist Zhou <i>et al.</i> [2026]	TGRS 2025	31.5	59.3	28.5	13.7	28.1	38.0	47.3
YOLOv5-M [Jocher, 2020]	GitHub 2020	30.6	58.4	27.6	10.3	25.3	38.7	49.3
YOLOv8-M [Jocher <i>et al.</i> , 2023]	GitHub 2023	31.3	59.2	28.3	10.8	26.0	39.5	50.3
YOLOv10-M [Wang <i>et al.</i> , 2024]	NeurIPS 2024	31.0	58.9	28.1	11.3	25.7	39.1	49.4
YOLO11-M [Jocher and Qiu, 2024]	GitHub 2024	31.5	59.5	28.5	11.4	26.6	39.6	49.3
RT-DETR-H [Lv <i>et al.</i> , 2023]	CVPR 2024	21.1	42.6	14.4	5.4	16.5	29.4	41.3
<i>Transformer-Based Methods</i>								
SO-DETR Fan <i>et al.</i> [2026]	TGRS 2026	33.3	62.8	30.0	18.5	31.0	38.7	45.7
Cross-DINO Cao <i>et al.</i> [2025]	TMM 2024	32.0	63.0	27.9	15.0	28.4	38.3	47.9
DINO [Zhang <i>et al.</i> , 2022]	ICLR 2023	28.9	59.4	22.4	12.5	22.7	34.7	42.8
Co-DINO [Zong <i>et al.</i> , 2023]	ICCV 2023	29.7	61.3	23.9	13.6	25.3	36.1	45.1
RT-DETR [Lv <i>et al.</i> , 2023]	arXiv 2023	29.3	60.2	25.2	13.2	26.9	35.4	44.6
Deformable-DETR [Zhu <i>et al.</i> , 2021]	ICLR 2020	19.2	44.8	13.7	6.3	15.4	24.9	34.2
Conditional DETR [Meng <i>et al.</i> , 2021]	ICCV 2021	25.7	52.8	15.0	7.9	20.3	28.0	36.5
DAB-DETR [Liu <i>et al.</i> , 2022]	ICLR 2022	27.2	55.1	20.6	10.3	22.5	31.9	37.2
Sparse R-CNN [Sun <i>et al.</i> , 2021b]	CVPR 2021	24.2	50.3	20.3	8.8	20.4	30.2	39.4
<i>Dual-Stage Methods</i>								
KL He <i>et al.</i> [2019]	CVPR 2019	29.4	59.2	24.7	11.0	26.9	36.0	44.2
RFLA [Xu <i>et al.</i> , 2022]	ECCV 2022	29.7	60.2	25.2	13.2	26.3	35.4	44.6
CFINet [Yuan <i>et al.</i> , 2023]	ICCV 2023	30.7	60.8	26.7	13.2	27.8	36.4	44.6
HS-FPN [Shi <i>et al.</i> , 2025]	AAAI 2025	29.6	56.8	26.7	13.6	26.4	35.3	45.3
SR-TOD [Cao <i>et al.</i> , 2024]	ECCV 2024	29.3	60.0	24.5	13.8	26.0	35.6	43.4
SimD [Shi <i>et al.</i> , 2024]	IROS 2024	27.8	58.0	21.8	12.2	24.6	33.6	41.9
F-CNN [Ren <i>et al.</i> , 2017]	NeurIPS 2015	28.9	59.4	24.1	13.8	25.7	34.5	43.0
Cascade RPN [Vu <i>et al.</i> , 2019]	NeurIPS 2019	29.7	56.5	25.9	12.5	25.5	35.4	44.7
BAFNet [Song <i>et al.</i> , 2025]	TGRS 2025	30.1	61.0	25.5	14.5	26.6	36.2	44.7
Unc-SOD (1 $\times$) [Yuan <i>et al.</i> , 2026]	TIP 2026	31.0	60.8	27.1	14.9	27.6	36.9	45.8
Unc-SOD (2 $\times$) [Yuan <i>et al.</i> , 2026]	TIP 2026	32.5	62.5	28.7	16.4	29.1	38.6	46.9
Better Sampling [Huang <i>et al.</i> , 2024]	arXiv 2024	31.4	63.7	28.1	15.6	29.7	38.4	46.4
SFDNet [Guo <i>et al.</i> , 2026]	arXiv 2026	31.3	62.1	26.8	15.1	27.8	37.3	46.2
HyMTAS-GBM (Ours)	–	33.7	64.1	30.4	18.2	30.3	40.0	50.8

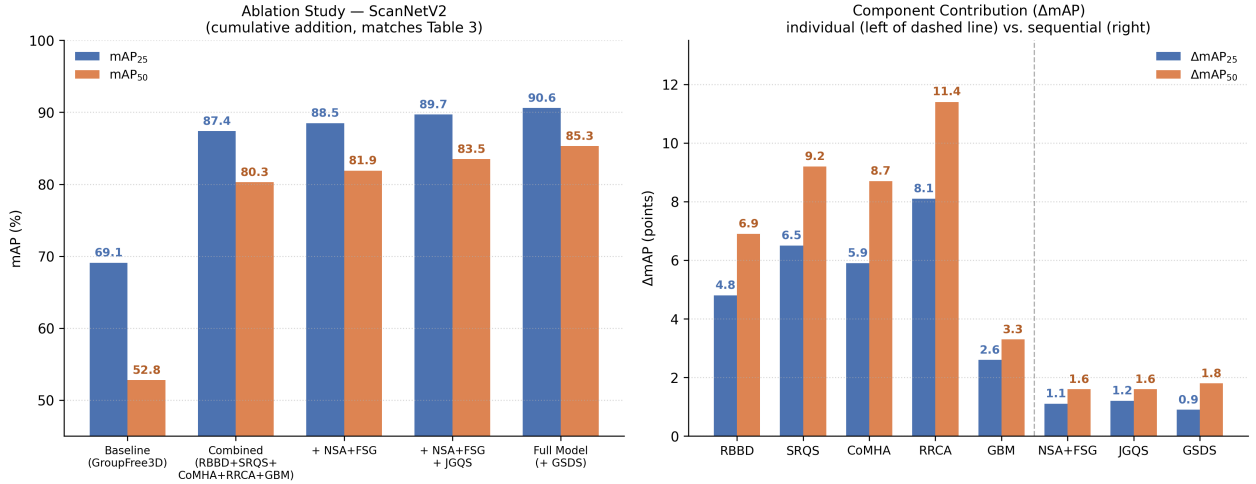


Figure 5. Ablation analysis on the ScanNetV2 validation set. **(Left)** Cumulative mAP₂₅ and mAP₅₀ obtained by progressively adding model components from the GroupFree3D baseline to the complete HyMTAS-GBM framework. **(Right)** Performance gain (ΔmAP) contributed by each component. RBBD, SRQS, CoMHA, RRCA, and GBM denote individual contributions measured relative to the baseline, whereas NSA+FSG, JGQS, and GSDS denote sequential gains measured after addition to the combined model. All values are derived directly from Table 4, ensuring consistency between the cumulative and gain analyses.

Table 4. Ablation study of HyMTAS-GBM components on the ScanNetV2 validation set [Dai *et al.*, 2017]. Results are reported as improvements over the GroupFree3D baseline [Liu *et al.*, 2021].

Configuration	ΔmAP_{25}	ΔmAP_{50}
Baseline (GroupFree3D)	69.1	52.8
<i>Individual Component Contribution</i>		
+ RBBD	+4.8	+6.9
+ SRQS	+6.5	+9.2
+ CoMHA	+5.9	+8.7
+ RRCA	+8.1	+11.4
+ GBM Head	+2.6	+3.3
<i>Combined Model</i>		
RBBD + SRQS + CoMHA + RRCA + GBM	+18.30	+27.50
<i>Additional Components (Section 3.8)</i>		
+ NSA + FSG	+19.4	+29.1
+ NSA + FSG + JGQS	+20.6	+30.7
Full Model + GSDS	+21.5	+32.5
Final Performance	90.6	85.3

Table 5. Efficiency comparison on the ScanNetV2 benchmark [Dai *et al.*, 2017].

Method	mAP ₂₅	Latency (ms)	Param (M)	GPU Mem (MB)
V-DETR [Shen <i>et al.</i> , 2023]	76.4	238	75.6	642
CAGroup3D [Wang <i>et al.</i> , 2022a]	75.1	187	47.2	498
HyMTAS-GBM	90.6	249	78.5	661

improved correlation refinement of CoMHA enables reliable feature aggregation across frames even when individual frame appearances are degraded [Zhu *et al.*, 2017]. The GBM classification head further enhances category discrimination in challenging cases such as visually similar object pairs, resulting in reduced false-positive detections across both datasets.

5.2 Limitations and Future Work

Despite its strong empirical performance, HyMTAS-GBM has several limitations. First, the GBM classification head requires offline training after the transformer has converged, introducing an additional training stage. Future work could explore end-to-end integration of differentiable gradient boost-

ing within the training pipeline to enable joint optimization. Second, the RRCA module introduces additional inference latency due to the correlation refinement operations. Efficient approximations, such as sparse key-point sampling, could reduce this overhead while preserving attention quality. Third, the current evaluation focuses primarily on indoor 3D point cloud benchmarks [Dai *et al.*, 2017; Song *et al.*, 2015] and standard video detection datasets [Russakovsky *et al.*, 2015; Wen *et al.*, 2020], which, while long-established and widely used for direct comparison with prior state-of-the-art methods, are not the most recently introduced benchmarks in their respective areas. Evaluating HyMTAS-GBM on newer, larger-scale benchmarks such as ScanNet++, nuScenes, TAO, and BDD100K, alongside extending the framework to outdoor LiDAR perception scenarios and multimodal sensor fusion, remains an important direction for future work.

6 Conclusion

This paper presented HyMTAS-GBM, a unified hybrid transformer framework that integrates attention-guided semantic convolutional features with gradient-boosting-based classification for robust object detection in both 3D indoor point clouds and video data. The proposed architecture enhances detection performance through several complementary mechanisms, including consensus multi-head attention for correlation refinement, semantic-regional query selection with joint geometric-quality assignment for improved query initialization, regional reinforcement cross-attention for instance-level spatial feature aggregation, and a gradient boosting classifier that strengthens discriminative decision boundaries. In addition, a sparsity-guided feature amplification pipeline with graduated score-decay suppression improves the detection of occluded and sparsely represented objects while reducing false suppression during inference.

Extensive experiments on ScanNetV2, SUN RGB-D, ImageNet VID, and UA-DETRAC demonstrate that HyMTAS-GBM consistently improves detection accuracy and robust-

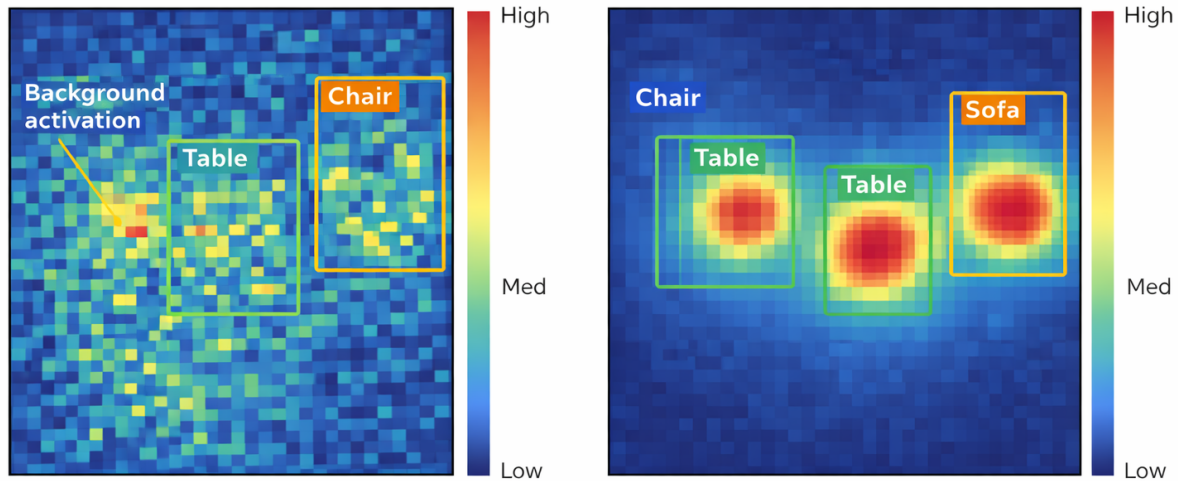


Figure 6. Qualitative attention map comparison on a representative indoor scene. (Left) Baseline DETR exhibits scattered cross-attention across background regions. (Right) HyMTAS-GBM with CoMHA produces spatially concentrated attention maps tightly localized to object instances (e.g., chair, table, sofa), illustrating the noise-suppression effect of the internal attention-within-attention mechanism.

ness compared with existing methods. Ablation studies further validate the effectiveness of each architectural component and show that their combination provides complementary performance gains. Efficiency analysis indicates that these improvements are achieved with only moderate computational overhead.

Overall, the proposed framework demonstrates the effectiveness of integrating transformer-based attention modeling with structured boosting-based classification within a unified detection pipeline. Future work will investigate end-to-end integration of differentiable boosting mechanisms, multimodal RGB-D fusion, and the extension of the framework to large-scale outdoor perception tasks such as autonomous driving and intelligent surveillance.

Declarations

Authors' Contributions

Rahul Mandlik: Conceptualization, Methodology, Software, Validation, Formal analysis, Resources, Investigation, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Avinash Sharma:** Supervision, Methodology, Validation, Resources, Investigation, Data curation, Project administration, Writing – review & editing. Author contributions are described using the CRediT Taxonomy: <https://credit.niso.org/>

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the Department of Computer Science and Engineering, Oriental University, Indore, for providing the necessary facilities and support to carry out this research work.

Funding

No funding was received for this research.

Availability of data and materials

All datasets used in this study are publicly available. ScanNetV2 can be accessed at <http://www.scan-net.org/ScanNet/>, SUN RGB-D is available at <https://rgbd.cs.princeton.edu/>, the ImageNet VID dataset can be obtained from <https://image-net.org/challenges/LSVRC/2015/>, and the UA-DETRAC dataset is available at <https://detrac-db.rit.albany.edu/>. No new datasets were generated or analyzed in this study. The source code implementing HyMTAS-GBM, including training and evaluation scripts, is publicly available at <https://github.com/mandlikrahul/hymtas-gbm>.

References

- Athar, A., Mahadevan, S., Osep, A., Leal-Taixé, L., and Leibe, B. (2020). Stem-seg: spatio-temporal embeddings for instance segmentation in videos. In *Lecture Notes in Computer Science*, pages 158–177. Springer. DOI: 10.1007/978-3-030-58621-8_10.
- Cao, B., Yao, H., Zhu, P., and Hu, Q. (2024). Visible and clear: Finding tiny objects in difference map. In *Proceedings of the European Conference on Computer Vision (ECCV)*. DOI: 10.48550/arXiv.2405.11276.
- Cao, G., Huang, W., Lan, X., Zhang, J., Jiang, D., and Wang, Y. (2025). Cross-dino: Cross the deep mlp and transformer for small object detection. *IEEE Transactions on Multimedia*, 27:7369–7379. DOI: 10.1109/tmm.2025.3599074.
- Chen, X., Ma, H., Wan, J., Li, B., and Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915. DOI: 10.1109/CVPR.2017.691.
- Chen, Y., Cao, Y., Hu, H., and Wang, L. (2020). Memory enhanced global-local aggregation for video object detection. In *Proceedings of the IEEE/CVF Conference on*

- Computer Vision and Pattern Recognition (CVPR)*, pages 10334–10343. DOI: 10.1109/CVPR42600.2020.01035.
- Chen, Y., Han, B., Li, C., Ko, T., Hebert, M., and Shi, J. (2021). Hais: hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15480–15489. DOI: 10.1109/ICCV48922.2021.01518.
- Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., and Han, J. (2023). Towards large-scale small object detection: survey and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13467–13488. DOI: 10.1109/TPAMI.2023.3290594.
- Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., and Nießner, M. (2017). Scannet: richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2432–2443. DOI: 10.1109/CVPR.2017.261.
- Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., and Zhang, L. (2021). Dynamic head: unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7373–7382. DOI: 10.1109/CVPR46437.2021.00729.
- Du, Z., Hu, Z., Zhao, G., Jin, Y., and Ma, H. (2025). Cross-layer feature pyramid transformer for small object detection in aerial images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:5625714. DOI: 10.1109/TGRS.2025.3572706.
- Engelmann, F., Bokeloh, M., Fathi, A., Leibe, B., and Nießner, M. (2020). 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9028–9037. DOI: 10.1109/CVPR42600.2020.00905.
- Fan, Z., Wang, C., Yu, Q., Min, L., Wang, S., and Wang, B. (2026). So-detr: Small object detection transformer via state space duality encoder and multigroup rank supervision. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5630013. DOI: 10.1109/tgrs.2026.3703859.
- Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). YoloX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*. DOI: 10.48550/arXiv.2107.08430.
- Gu, Y., Hu, Z., Zhao, Y., Liao, J., and Zhang, W. (2024). Mfgtn: a multi-modal fast gated transformer for identifying single trawl marine fishing vessel. *Ocean Engineering*, 303:117711. DOI: 10.1016/j.oceaneng.2024.117711.
- Guo, Y., Yang, Z., Kou, F., Hu, Y., Zhang, R., and Yao, S. (2026). Adaptive spectrum-aware feature disentangled network for small object detection. *arXiv preprint arXiv:2606.29029*. DOI: 10.48550/arXiv.2606.29029.
- Han, K. et al. (2023). A survey on vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:87–110. DOI: 10.1109/TPAMI.2022.3152247.
- Han, M., Wang, Y., Chang, X., and Qiao, Y. (2020). Mining inter-video proposal relations for video object detection. In *Lecture Notes in Computer Science*, volume 12366. Springer. DOI: 10.1007/978-3-030-58589-1_26.
- He, Y., Zhu, C., Wang, J., Savvides, M., and Zhang, X. (2019). Bounding box regression with uncertainty for accurate object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2883–2892. DOI: 10.1109/CVPR.2019.00300.
- Huang, Z., Zhang, C., Jin, M., Wu, F., Liu, C., and Jin, X. (2024). Better sampling, towards better end-to-end small object detection. *arXiv preprint arXiv:2407.06127*. DOI: 10.48550/arXiv.2407.06127.
- Jiang, J., Zhao, D., Shi, J., and Gao, Y. (2020). Pointgroup: dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876. DOI: 10.1109/CVPR42600.2020.00492.
- Jocher, G. (2020). Ultralytics yolov5. <https://github.com/ultralytics/yolov5>.
- Jocher, G., Chaurasia, A., and Qiu, J. (2023). Ultralytics yolov8. <https://github.com/ultralytics/ultralytics>.
- Jocher, G. and Qiu, J. (2024). Ultralytics yolol11. <https://github.com/ultralytics/ultralytics>.
- Law, H. and Deng, J. (2018). Cornernet: detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750. DOI: 10.1007/978-3-030-01264-9_45.
- Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M., and Zhang, L. (2024). Dn-detr: accelerate detr training by introducing query denoising. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2239–2251. DOI: 10.1109/TPAMI.2023.3335410.
- Li, H. et al. (2023). Multi-object tracking via deep feature fusion and association analysis. *Engineering Applications of Artificial Intelligence*, 124:106527. DOI: 10.1016/j.engappai.2023.106527.
- Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Yang, J., and Dai, J. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*. DOI: 10.48550/arXiv.2006.04388.
- Li, Y. et al. (2022). Mvitv2: improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4804. DOI: 10.1109/CVPR52688.2022.00476.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017a). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007. DOI: 10.1109/ICCV.2017.324.
- Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017b). Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. DOI: 10.1109/ICCV.2017.324.
- Liu, S. et al. (2022). Dab-detr: dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*. DOI: 10.48550/arXiv.2201.12329.
- Liu, Z., Zhang, Z., Cao, Y., Hu, H., and Tong, X. (2021). Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference*

- on Computer Vision (ICCV), pages 2929–2938. DOI: 10.1109/ICCV48922.2021.00294.
- Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., and Liu, Y. (2023). Detsr beat yolos on real-time object detection. *arXiv preprint arXiv:2304.08069*. DOI: 10.48550/arXiv.2304.08069.
- Mandlik, S., Labade, R., Chaudhari, S., and Agarkar, B. (2025a). Optimizing gait-based biometric recognition with deep dense networks. *Brazilian Journal of Biometrics*, 43:e-43857. DOI: 10.28951/bjb.v43i4.857.
- Mandlik, S., Labade, R., Chaudhari, S., et al. (2025b). Pose-invariant gait recognition using an enhanced inception-resnet and vision transformer framework. *Journal of Engineering and Applied Sciences*, 72:234. DOI: 10.1186/s44147-025-00820-8.
- Mandlik, S. et al. (2026). Dual attention embedded inception framework for robust biometric gait recognition. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 25(2):79–105. DOI: 10.5565/rev/elcvia.2402.
- Meng, D. et al. (2021). Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3631–3640. DOI: 10.1109/ICCV48922.2021.00363.
- Misra, I., Girdhar, R., and Joulin, A. (2021). An end-to-end transformer model for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2886–2897. DOI: 10.1109/ICCV48922.2021.00290.
- Pagire, V., Chavali, M., and Kale, A. (2025). A comprehensive review of object detection with traditional and deep learning methods. *Signal Processing*, 237:110075. DOI: 10.1016/j.sigpro.2025.110075.
- Pan, L., Luan, W., Zheng, Y., Li, J., Tao, L., and Xu, C. (2024). Graphfusion: integrating multi-level semantic information with graph computing for enhanced 3d instance segmentation. *Neurocomputing*, 602:128287. DOI: 10.1016/j.neucom.2024.128287.
- Qi, C. R., Litany, O., He, K., and Guibas, L. J. (2019). Deep hough voting for 3d object detection in point clouds. *arXiv preprint arXiv:1904.09664*. DOI: 10.48550/arXiv.1904.09664.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788. DOI: 10.1109/CVPR.2016.91.
- Ren, S., He, K., Girshick, R., and Sun, J. (2017). Faster r-cnn: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149. DOI: 10.1109/T-PAMI.2016.2577031.
- Rukhovich, D., Vorontsova, A., and Konushin, A. (2022). Imvoxelnet: image to voxels projection for monocular and multi-view 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2397–2406. DOI: 10.1109/WACV51458.2022.00133.
- Russakovsky, O., Deng, J., Su, H., et al. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211–252. DOI: 10.1007/s11263-015-0816-y.
- Shehzadi, T., Hashmi, K. A., Liwicki, M., Stricker, D., and Afzal, M. Z. (2025). Object detection with transformers: a review. *Sensors*, 25:6025. DOI: 10.3390/s25196025.
- Shen, Y., Geng, Z., Yuan, Y., Lin, Y., Liu, Z., Wang, C., Hu, H., Zheng, N., and Guo, B. (2023). V-detr: Detr with vertex relative position encoding for 3d object detection. *arXiv preprint arXiv:2308.04409*. DOI: 10.48550/arXiv.2308.04409.
- Shi, S., Fang, Q., Xu, X., and Zhao, T. (2024). Similarity distance-based label assignment for tiny object detection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13711–13718. DOI: 10.1109/IROS58592.2024.10801448.
- Shi, W. and Rajkumar, R. (2020). Point-gnn: graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1711–1719. DOI: 10.1109/CVPR42600.2020.00178.
- Shi, Z., Hu, J., Ren, J., Ye, H., Yuan, X., Ouyang, Y., He, J., Ji, B., and Guo, J. (2025). Hs-fpn: high frequency and spatial perception fpn for tiny object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6896–6904. DOI: 10.1609/aaai.v39i7.32740.
- Song, J., Zhou, M., Luo, J., Pu, H., Feng, Y., Wei, X., and Jia, W. (2025). Boundary-aware feature fusion with dual-stream attention for remote sensing small object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 63:5600213. DOI: 10.1109/TGRS.2024.3514376.
- Song, S., Lichtenberg, S. P., and Xiao, J. (2015). Sun rgb-d: a rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 567–576. DOI: 10.1109/CVPR.2015.7298655.
- Sun, G., Hua, Y., Hu, G., and Robertson, N. (2021a). Mamba: multi-level aggregation via memory bank for video object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2620–2627. DOI: 10.1609/aaai.v35i3.16365.
- Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., and Luo, P. (2021b). Sparse r-cnn: end-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14454–14463. DOI: 10.1109/CVPR46437.2021.01422.
- Tian, Z., Shen, C., Chen, H., and He, T. (2019a). Fcos: fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635. DOI: 10.1109/ICCV.2019.00972.
- Tian, Z., Shen, C., Chen, H., and He, T. (2019b). Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9627–9636. DOI: 10.1109/ICCV.2019.00972.
- Vu, T., Jang, H., Pham, T. X., and Yoo, C. D. (2019). Cascade rpn: delving into high-quality region proposal

- network with adaptive convolution. In *Advances in Neural Information Processing Systems (NeurIPS)*. DOI: 10.48550/arXiv.1909.06720.
- Vu, T., Kim, K., Luu, T. M., Nguyen, T., and Yoo, C. D. (2022). Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2708–2717. DOI: 10.48550/arXiv.2203.01509.
- Wang, A. *et al.* (2024). Yolov10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems (NeurIPS)*. DOI: 10.48550/arXiv.2405.14458.
- Wang, H., Ding, L., Dong, S., Shi, S., Li, A., Li, J., Li, Z., and Wang, L. (2022a). Cagroup3d: class-aware grouping for 3d object detection on point clouds. In *Advances in Neural Information Processing Systems (NeurIPS)*. DOI: 10.52202/068431-2173.
- Wang, H. *et al.* (2022b). Rbgnet: ray-based grouping for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1119. DOI: 10.1109/CVPR52688.2022.00118.
- Wang, S., Lu, K., Xue, J., and Zhao, Y. (2025). Dagnet: density-aware 3d object detection network for point clouds. *IEEE Transactions on Multimedia*, 27:665–678. DOI: 10.1109/TMM.2023.3245359.
- Wei, X. and Konushin, A. (2026). Hff-big: hierarchical feature fusion with bi-directional gating for 3d indoor instance segmentation. *Neurocomputing*, 680:133233. DOI: 10.1016/j.neucom.2026.133233.
- Wen, L., Du, D., Cai, Z., *et al.* (2020). Ua-detrac: a new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding*, 193:102907. DOI: 10.1016/j.cviu.2020.102907.
- Wu, H., Chen, Y., Wang, N., and Zhang, Z. (2019). Sequence level semantics aggregation for video object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9217–9225. DOI: 10.1109/ICCV.2019.00931.
- Xie, Q., Lai, Y.-K., Wu, J., Wang, Z., Zhang, Y., Xu, K., and Wang, J. (2020). Mlcvnet: multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10447–10456. DOI: 10.1109/CVPR42600.2020.01046.
- Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., and Xia, G.-S. (2022). Rfla: Gaussian receptive field based label assignment for tiny object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 526–543. DOI: 10.1007/978-3-031-20077-9_31.
- Xu, Z., Hrustic, E., and Vivet, D. (2020). Centernet heatmap propagation for real-time video object detection. In *Lecture Notes in Computer Science*, volume 12370. Springer. DOI: 10.1007/978-3-030-58595-2_14.
- Yang, Z., Liu, S., Hu, H., Wang, L., and Lin, S. (2019). Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9657–9666. DOI: 10.1109/ICCV.2019.00975.
- Yao, Y., Luo, Z., Li, S., Fang, T., and Zeng, A. (2023). Occformer: dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15590–15600. DOI: 10.1109/ICCV51070.2023.00865.
- Yin, X., Lian, X., Han, Z., Zhu, B., Li, H., and Nießner, M. (2023). Uni3d: exploring unified 3d representation at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20226–20237. DOI: 10.48550/arXiv.2310.06773.
- Yuan, X., Cheng, G., Cheng, J., Yao, R., and Han, J. (2026). Unc-sod: an uncertainty learning framework for small object detection. *IEEE Transactions on Image Processing*, 35:1127–1142. DOI: 10.1109/TIP.2026.3654892.
- Yuan, X., Cheng, G., Yan, K., Zeng, Q., and Han, J. (2023). Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6317–6327. DOI: 10.1109/ICCV51070.2023.00581.
- Zhang, H., Liu, H., and Kim, C. (2024). Semantic and instance segmentation in coastal urban spatial perception. *Sustainability*, 16:833. DOI: 10.3390/su16020833.
- Zhang, H. *et al.* (2022). Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*. DOI: 10.48550/arXiv.2203.03605.
- Zhang, S., Chi, C., Yao, Y., Lei, Z., and Li, S. Z. (2020). Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9759–9768. DOI: 10.1109/CVPR42600.2020.00978.
- Zhao, L., Guo, J., Xu, D., and Sheng, L. (2021). Transformer3d-det: improving 3d object detection by vote refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 31:4735–4746. DOI: 10.1109/tcsvt.2021.3102025.
- Zhou, X., Wang, D., and Krähenbühl, P. (2019). Objects as points. *arXiv preprint arXiv:1904.07850*. DOI: 10.48550/arXiv.1904.07850.
- Zhou, Z., Zhao, X., Zhao, X., Zhu, Y., Zhao, Z., and Li, L. (2026). Gendist: Generative-driven knowledge distillation for small object detection in remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–16. DOI: 10.1109/tgrs.2026.3676015.
- Zhou, Z. and Zhu, Y. (2024). Kldet: Detecting tiny objects in remote sensing images via kullback–leibler divergence. *IEEE Transactions on Geoscience and Remote Sensing*, 62:4703316. DOI: 10.1109/TGRS.2024.3382099.
- Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). Deformable detr: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2010.04159.
- Zhu, X., Wang, Y., Dai, J., Yuan, L., and Wei, Y. (2017). Flow-guided feature aggregation for video object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 408–417. DOI: 10.1109/ICCV.2017.52.

Zong, Z., Song, G., and Liu, Y. (2023). Detsr with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. DOI: 10.1109/ICCV51070.2023.00621.