2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

HG-DETR: Image-Level Few-Shot Object Detection with Cross-Category and Query-Level Heterogeneous Graphs

Liangchen Qu and Hongru Zhao*

School of Computer and Artificial Intelligence, Zhengzhou University,
Zhengzhou 450001, China
zhaohongru@zzu.edu.cn

Abstract. Few-shot object detection (FSOD) aims to detect novel objects with limited annotated examples, yet existing methods face critical challenges in handling low-quality region proposals, leading to suboptimal generalization. And current meta-learning approaches often rely on pairwise region-class matching, which neglects contextual relationships among proposals and fails to leverage cross-class semantic dependencies, resulting in misclassification over similar classes and limited adaptability to novel categories. To address these limitations, we propose HG-DETR, a novel FSOD framework that integrates image-level detection with heterogeneous relational reasoning. Our method bypasses error-prone region proposal networks by directly operating on holistic image features through a Transformer-based architecture, enabling end-to-end optimization. By considering these multi-faceted relationships between proposals and classes, we propose(1) a cross-category semantic relationship graph that dynamically models semantic dependencies among base and novel classes to enhance prototype representations through knowledge transfer, (2) a query-level context aggregation graph models spatial relation- ships within a query image by connecting top-confidence proposals and a class node, using a GCN layer to aggregate features and refine proposals, and (3) bidirectional class-query adaptation via attention mechanisms to align feature distributions and bridge domain gaps. Qualitative and quantitative results demonstrate that our method achieves superior performance in few-shot object detection on Pascal VOC and MS COCO datasets compared with existing methods.

Keywords: Object Detection, Few-Shot Learning, Few-Shot Object Detection, Heterogeneous Graph Convolutional Networks.

1 Introduction

In recent years, computer vision technology has made remarkable advancements. However, a considerable gap remains between current systems and human vision in terms of the ability to learn new concepts from only a few examples. Most existing methods depend heavily on large volumes of labeled data, whereas humans can accurately recognize novel objects with just a handful of examples. Few-Shot Object Detection (FSOD), a key task aimed at bridging this gap, seeks to enable efficient object detection using limited annotated samples from novel categories. Although meta-learning-based methods [7],[22] have shown progress by integrating region-based detection frameworks with feature reweighting strategies, their performance remains constrained by two critical challenges.

Firstly, traditional methods rely heavily on region proposal networks to generate candidate frames, but the quality of region proposals for new classes significantly degrades in low-sample scenarios, because new category proposals under limited supervision often contain a large amount of noise or missed detections, hindering the effective migration of base class knowledge. Secondly, existing meta-learning methods mostly adopt a category-by-category independent processing model, i.e., feature matching is performed for a single support category at a time. This isolated learning strategy ignores the semantic associations between categories, e.g., the similarity between ‘cow’ and ‘horse,’ for example, may exacerbate misclassification, while the commonality between ‘sheep’ and ‘cow’ is not used for knowledge transfer. This neglect of category relevance makes it difficult for the model to distinguish similar categories and limits the ability to generalise across categories.

To address the above challenges, this paper proposes HG-DETR, an innovative few-shot object detector based on heterogeneous relational reasoning. The framework is based on DETR, abandons the region proposal mechanism, directly achieves end-to-end detection through image-level features, and introduces hierarchical graph structure modeling to systematically solve the problems of category isolation and context fragmentation. Its core innovation contains the fol- lowing two aspects: 1) Cross-category semantic relationship graph, a dynamic semantic relationship graph that is designed to explicitly model the semantic associations between base categories and new categories, and between new categories. The nodes in the graph are prototypical features of all categories, and the edge weights are computed by prototypical cosine similarity to reflect the semantic tightness between categories. The discriminative features of the base class are migrated to the prototypes of the new classes through graph convolutional message passing, while the interactions between the prototypes of similar new classes enhance the differentiation. 2) Query-level context aggregation graph. In the decoding phase, a query-level relationship graph is constructed. The graph takes proposals and novel class prototypes as nodes and optimizes detection through two connection strategies: one is to establish local connections based on the intersection and union ratio (IoU) of prediction frames, where spatially overlapping query nodes share localization information to improve the detection consistency of occluded or small target scenes; the other is to introduce class prototypes, which calibrates the different statistical distribution between the proposal feature and the class prototype.

In light of the above analysis, The main contributions of this paper can be summarized as follows: 1) We propose the first FSOD framework that integrates image-level detection with hierarchical graph inference to systematically address region dependency and category isolation. 2) We design a hierarchical category semantic propagation network. By combining cross-category semantic graphs with query-level context graphs, our method captures inter-category semantic dependencies and intra-query spatial associations, thereby significantly improving the differentiation of similar categories and the accuracy of object localization. 3) We introduce a Bidirectional Feature Adaptation Mechanism: a symmetric attention module is embedded in the Transformer decoder to enable bidirectional interaction between category prototypes and query features, dynamically aligning feature distributions and reducing the domain gap.

2 Related Work

2.1 Object Detection

Modern object detection architectures are broadly categorized into two-stage, single-stage, and image-level transformer-based paradigms. Two-stage detectors (e.g., Faster R-CNN [15] and its variants employ a Region Proposal Network (RPN) to generate candidate regions, followed by region-wise classification and regression. While effective in many-shot scenarios, their reliance on proposal quality and multi-stage pipelines introduces computational overhead and limits adaptability to few-shot settings. Single-stage detectors (e.g., YOLO [14], SSD[11]) bypass the RPN by densely sampling anchors, enabling faster inference but sacrificing precision in occluded or small-object cases. Recent advancements in Transformer-based models, such as DETR [2] and its derivatives [26],[4],[10], have redefined detection paradigms through pure image-level frameworks. By replacing handcrafted components (e.g., anchors, NMS) with learnable queries and attention mechanisms, these methods achieve end-to-end optimization and competitive performance. In the context of FSOD, existing frameworks inherit the limitations of their base detectors. Two-stage methods struggle with noisy proposals under low-sample regimes, while single-stage and transformer-based approaches lack explicit mechanisms to exploit inter-class dependencies or contextual coherence. Our work builds upon the strengths of image-level detection paradigms, addressing these challenges through structured relational reasoning to enable robust generalization with minimal supervision.

2.2 Few-Shot Object Detection

Few-shot object detection (FSOD) aims to detect novel objects with minimal supervision while retaining robustness against background interference and semantic confusion among similar categories. Existing approaches predominantly follow meta-learning or transfer learning paradigms. Meta-learning approaches, aim to learn task-agnostic feature matching through episodic training. While effective, these methods rely on region proposal networks (RPNs), leading to performance degradation due to noisy proposals for novel classes. Recent advancements like Meta-DETR [24] redefined meta-learning by integrating DETR's image-level detection framework, eliminating proposal dependency, and enabling end-to-end optimization. Transfer learning-based methods (e.g., TFA [19], FSCE [16]) fine-tune pre-trained detectors on novel classes, balancing simplicity and efficiency. However, they struggle to decouple class-agnostic localization from category-specific features, resulting in biased generalization. Our work builds on Meta-DETR's image-level meta-learning paradigm but introduces hierarchical graph reasoning to address its limitations. It is the first framework to integrate graph convolutional networks (GCNs) into image-level meta-learning.

2.3 Graph Convolutional Networks(GCNs)

Graph Convolutional Networks [8] and their variants, such as Graph Attention Networks (GATs) [17], have been widely adopted in computer vision for modeling relational structures, including action localization [13], visual relation reasoning [12], and object proposal interactions [3]. While existing studies primarily focus on many-shot scenarios or rely on predefined ontologies [18], their application to Few-Shot Object Detection (FSOD) remains limited. Re- cent efforts, such as QA-FewDet [6], have pioneered the use of heterogeneous GCNs to model class–proposal and proposal–proposal relationships, yet these approaches still depend on region proposal mechanisms. In contrast, HG-DETR is the first to integrate hierarchical graph reasoning into a proposal-free, image- level detection framework. By combining this proposal-free design with two novel modules—Inter-Class Semantic Graphs and Intra-Query Context Graphs—our framework explicitly captures cross-category dependencies and spatial-semantic coherence, effectively addressing the limitations of both region-based and conventional image-level methods. To the best of our knowledge, this is the first study to unify GCN-based relational modeling with holistic image-level detection in the FSOD setting, enabling robust knowledge transfer and accurate localization without the noise introduced by proposal mechanisms. The detailed methodology is presented in Section 5.

3 Problem Definition

In few-shot object detection (FSOD), we consider two disjoint sets of object classes: base classes Cbase and novel classes Cnovel, where:

$$Cbase \cap Cnovel = \emptyset$$

The goal is to train a detector using:

A richly annotated base dataset Dbase (with abundant instances for each c ∈ Cbase)

A novel dataset Dnovel (with only K annotated instances per c ∈ Cnovel) such that the model can generalize to detect objects from Cnovel in unseen query images.

Formally, for each class c, the annotations include instance labels and bounding boxes:

$$T_c = \{(c, u, I) \mid u \in U, I \in \mathbb{R}^{H_I \times W_I \times 3}\}$$

where U defines the bounding box coordinates (x, y, w, h). In the K-shot setting, each novel class has exactly K support instances in Dnovel. Given a query image Iq the task is to output a set of detections：

$$S_q = \{(c, u) \mid c \in Cnovel, u \in U\}$$

accurately localizing and classifying novel-class objects while suppressing background regions.

4 The Baseline FSOD Model

Our baseline few-shot object detection (FSOD) framework, Meta-DETR [24], employs the Deformable DETR [26], a fully end to-end Transformer-based detector, as the basic detection framework. The model begins by extracting features from both the query image and support images using a shared backbone network (ResNet101). For the query image, multi-scale features are encoded through the backbone, while support images are processed to generate class prototypes via global average pooling, aggregating features from the limited annotated instances of each novel class. These prototypes represent the semantic essence of novel categories. The query features are then fed into a Transformer encoder to model global spatial dependencies via self-attention, capturing long-range contextual relationships critical for detecting occluded or small objects.

In the decoding phase, learnable object queries—initialized with positional encodings—interact with the encoded query features through cross-attention layers. Task-specific embeddings are injected into the decoder to align queries with the support class prototypes adaptively, guiding the model to focus on novel-class instances while suppressing background regions. The decoder directly predicts bounding box coordinates and class probabilities, bypassing traditional region proposal networks (RPNs) entirely.

While Meta-DETR [24] eliminates the noise introduced by RPN-based proposals and leverages global attention for robust context modeling, it faces challenges outlined in Section 2, including fragmented semantic reasoning and limited adaptability to domain shifts. These issues arise from isolated class matching and static prototype alignment. In Section 5, we introduce a hierarchical graph-augmented framework that addresses these challenges through Cross-category semantic relationship graph and Query-level context aggregation graph, systematically bridging the gaps in the baseline model.

5 Methodology

5.1 Overview

Our HG-DETR framework extends the DETR architecture into a unified pipeline for few-shot object detection, eliminating region proposals while integrating hierarchical context modeling and dynamic feature adaptation. As depicted in Fig.1, the model begins by processing a query image and (K)-shot support images through a shared ResNet-101 backbone, extracting visual features. Class prototypes for novel categories are initialized by averaging support features, forming the foundation for subsequent relational reasoning. The class prototypes then flow into the Cross-category semantic relationship graph, which constructs a global graph over all base and novel classes. Here, edges encode semantic correlations through sparse top-(k) connections (k=20) based on cosine similarity, allowing a GCN layer to propagate features across classes. This global enhancement mitigates prototype bias caused by the limited number of novel-class samples.

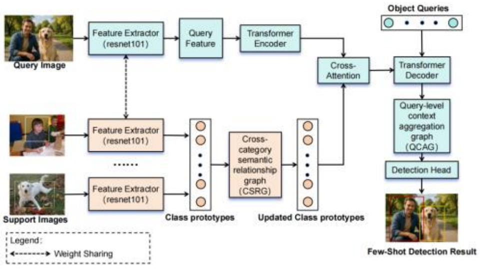

Object Queries
Query Image
Feature Extractor (resnet101)
Query Feature
Transformer Encoder
Cross-Attention
Transformer Decoder
Query-level context aggregation graph (QCAG)
Feature Extractor (resnet101)
......
Feature Extractor (resnet101)
Cross-category semantic relationship graph (CSRG)
Detection Head
Support Images
Class prototypes
Updated Class prototypes
Legend:
Weight Sharing
Few-Shot Detection Result

Fig. 1: The overall architecture of our proposed model. Query Image and Sup- port Images are processed by a weight-shared feature extractor to generate query image features and few-shot class prototypes. Next, in order to capture semantic dependencies between categories, a cross-category semantic relationship graph is designed to enhance few-shot class prototypes. Finally, the Transformer Encoder & Decoder is used to implement few-shot detection, with a query-level context aggregation graph designed to refine proposal features by leveraging the contextual relationships between proposals

Next, a Transformer decoder generates 300 candidate proposals from the query features, with the top 100 proposals feeding into the Query-level context aggregation graph. This component builds a query-level graph, incorporating the enhanced class prototype and candidate proposals. Edges model spatial relationships and scene-level context, enabling a GCN to refine proposal features by aggregating local neighbor information.

Finally, the bidirectional adaptation module dynamically aligns the enhanced prototypes and refined proposals through dual cross-attention pathways. The resulting features are processed by a lightweight detection head to predict bounding boxes and class labels.

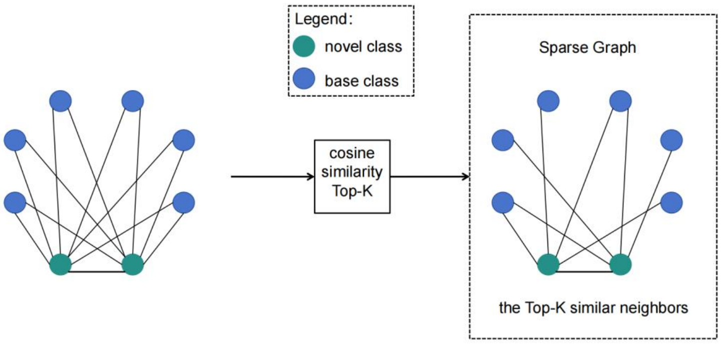


Fig. 2: The architecture of the Cross-category semantic relationship graph(CSRG). Nodes in the graph represent class prototypes for both base and novel classes. Edges between nodes are determined by calculating the cosine similarity between each pair of class prototypes, with edge weights normalized using a softmax function over the top-K highest similarities. This results in a sparsely connected graph that emphasizes the most semantically related classes.

5.2 Cross-category semantic relationship graph

To mitigate prototype bias in novel classes under limited supervision, we construct a Cross-category semantic relationship graph that globally models semantic relationships between base and novel classes. As shown in Fig.2, we define a heterogeneous graph Ginter = (Vinter, Einter) where: Nodes Vinter are class prototypes {f(ci)}, Edges Einter encode semantic correlations through sparse connections.

The semantic similarity between categories ci and cj is measured using cosine similarity:

$$e(c_i, c_j) = \frac{f(c_i)^\top f(c_j)}{\|f(c_i)\|_2 \|f(c_j)\|_2} \tag{1}$$

Edge weights are calculated using a softmax function over the top-20 highest similarities:

$$A_{inter}^{ji} = \frac{exp\big(e(c_i, c_j)\big)}{\sum_{k \in TopK(e(c_i, \cdot))} exp(\,e(c_i, c_k))} \tag{2}$$

where TopK(e(ci , ·)) is the set of indices for the top-K similarity scores of category ci.

A single-layer Graph Convolutional Network (GCN) then propagates features across the graph:

$$\hat{f}(c_i) = \sum_{j=1}^{C} A_{inter}^{ji} \cdot f(c_j) + f(c_i) \tag{3}$$

The enhanced prototype aggregates knowledge from semantically related classes (e.g., refining "cow" using "horse" and "sheep"), significantly reducing prototype bias compared to isolated class processing.

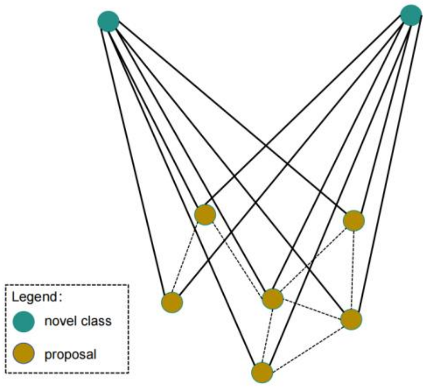


Fig. 3: The architecture of the Query-level context aggregation graph(QCAG). It consists of novel class nodes and proposal nodes. Edges connect proposals with an IoU > 0.7 and link the novel class node bidirectionally to all proposals. A GCN layer aggregates features from the novel class node and neighboring proposals, enhancing the proposal features by integrating local spatial relationships, thereby improving detection accuracy in few-shot scenarios.

5.3 Query-level context aggregation graph

To refine noisy candidate proposals, we design the Query-level context aggregation graph that models spatial relationships within a query image. As depicted in Fig.3, for ach novel class c, we build Gintra = (Vintra, Eintra), where: Nodes Vintra: Top 100 proposals {pi} (by confidence), Class node (enhanced by Eq.3); Edges Eintra: Proposal-Proposal: Connect pi and pj if IoU(pi, pj) > 0.7 (following [23]), Class-Proposal: Bidirectional edges between fˆ(c) and all pi.

A GCN layer aggregates contextual features:

$$\tilde{f}(p_i) = \left(A_{intra}^{cp_i} \cdot \hat{f}(c) + \sum_{p_j \in \mathcal{N}(p_i)} A_{intra}^{p_j p_i} \cdot f(p_j) \right) W + f(p_i) \qquad (4)$$

In this formula, we aggregate features from the class node and neighboring proposals. The class node's influence on proposal pi is weighted by $\mathrm{A}_{\mathrm{intra}}^{\mathrm{cp_i}}$, while the summation term aggregates features from neighboring proposals pj , weighted by $\mathrm{A}_{\mathrm{intra}}^{\mathrm{p_j p_i}}$. The learnable weight matrix W transforms these aggregated features, and the original proposal features f(pi) are added back to preserve the initial information. Here, N (pi) denotes neighboring proposals. This process effectively integrates local spatial context to suppress inconsistent detections.

The Query-level context aggregation graph effectively refines the quality of candidate proposals by integrating local context, leading to more accurate and reliable object detection in few-shot scenarios.

5.4 bidirectional class-query adaptation via attention

To address unidirectional alignment limitations in Meta-DETR, we propose Bidirectional Adaptation for mutual feature calibration. Given enhanced prototypes and refined proposals:

Class-to-Query Attention Prototypes attend to proposals to inject classaware semantics:

$$f_{cls2q} = Softmax\left(\frac{\hat{f}(c)(\tilde{f}(p))^{\top}}{\sqrt{d}}\right)\tilde{f}(p) \tag{5}$$

Query-to-Class Attention Proposals attend to prototypes to adapt prototypes to query context:

$$f_{q2cls} = Softmax\left(\frac{\tilde{f}(p)(\hat{f}(c))^{\top}}{\sqrt{d}}\right)\hat{f}(c) \tag{6}$$

Adaptive Fusion A learnable weight α balances both pathways:

$$f_{final} = \alpha \cdot f_{cls2q} + (1-\alpha) \cdot f_{q2cls}, \qquad \alpha \in [0,1] \tag{7}$$

This bidirectional design reduces distribution shifts between support and query domains, decreasing misclassification for similar classes (e.g., "cat" vs. "dog") compared to unidirectional approach.

6 Experimentals

6.1 Dataset

We evaluate our method on two widely adopted datasets for few-shot object detection: PASCAL VOC dataset and MS COCO dataset, following the established protocols from prior works [7],[22],[19]. Both datasets are split into base classes (with abundant annotations) and novel classes (with limited annotations) to simulate real-world scenarios where novel objects are sparsely annotated.

For PASCAL VOC dataset, we use the 20-class dataset with three predefined splits of 15 base classes and 5 novel classes. The novel classes in each split are: 1. bird, bus, cow, motorbike, sofa; 2. aeroplane, bottle, cow, horse, sofa; 3. boat, cat, motorbike, sheep, sofa. The model is trained on the ‘trainval’ sets from VOC 2007 and 2012 and evaluated on the ‘test‘ set of VOC 2007. For few-shot evaluation, we randomly sample K-shot (K = 1, 2, 3, 5, 10) support images for each novel class and report the mean average precision at IoU 0.5 (mAP@0.5) averaged over 10 independent runs.

For MS COCO dataset, we adopt the 20 PASCAL VOC classes as novel classes and the remaining 60 classes as base classes. Training is performed on the ’train2017’ subset, and evaluation is conducted on ’val2017’. In order to ensure compatibility with real-world data scarcity, we evaluate under (K = 1, 3, 5, 10, 30)-shot settings, where even 30-shot results remain far below fully supervised performance. We report standard COCO metrics (AP, AP50, AP75) averaged over 5 runs.

6.2 Comparison with State-of-the-Art Methods

All results are averaged over 10 runs for PASCAL VOC and 5 runs for MS COCO to reduce randomness, respectively. For fair comparison, we use the same support images and evaluation protocols as [7].

PASCAL VOC We evaluate our method on the PASCAL VOC benchmark under K = 1, 2, 3, 5, 10 -shot settings and compare it against state-of-the-art approaches, including FSRW [7], and TFA [19]. As shown in Table 1, Table 2, Table 3. our method achieves superior performance in most Few-shot settings. However, in the 1-shot setting, the performance is not as high as expected. This may be due to the limited novel-class samples, which can lead to unstable graph structure initialization, and the limited base classes in this dataset may also restrict the functionality of the cross-category semantic relationship graph. As the number of samples increases, the advantages of our method become more prominent.

MS COCO On the more challenging MS COCO benchmark, our method achieves strong performance under K = 1, 2, 3, 5, 10, 30 -shot settings. As shown in Table 4, Table 5, Table 6, our model outperforms region-based methods like MPSR [21] by significant margins. The performance gap widens in extremely low- shot scenarios. For 1-shot detection, our method achieves 13.7% AP50, +9.6% improvement over Meta-DETR.

6.3 Ablation Studies

We conduct comprehensive ablation studies on PASCAL VOC (Split 1: bird, bus, cow, motorbike, sofa) to validate the design choices in HG-DETR. All results are averaged over 10 independent runs with different support sets, and we report mean average precision at IoU 0.5 (mAP@0.5) for novel classes unless otherwise specified.

Table 1. Few-shot detection performance (mAP@0.5) on Pascal VOC for novel classes - Class Split 1

Method \ Shots	1	2	3	5	10
FSRW [7]	14.8	15.5	26.7	33.9	47.2
Meta Det [20]	18.9	20.6	30.2	36.8	49.6
Meta R-CNN [22]	19.9	25.5	35.0	45.7	51.5
TFA w/ fc [19]	36.8	29.1	43.6	55.7	57.0
TFA w/ cos [19]	39.8	36.1	44.7	55.5	56.0
MPSR [21]	41.7	43.1	51.4	55.2	61.8
Retentive R-CNN [5]	42.4	45.8	45.9	53.7	56.1
CME [9]	41.5	47.5	50.4	58.2	60.9
SRR-FSD [25]	47.8	50.5	51.3	55.2	56.8
FSCE [16]	44.2	43.8	51.4	61.9	63.4
QA-FewDet [6]	42.4	51.9	55.7	62.6	63.4
Meta-DETR [24]	40.6	51.4	58.0	59.2	63.6
FS-DETR [1]	45.0	48.5	51.5	52.7	56.1
HG-DETR (OURS)	39.1	53.3	54.3	63.2	64.6

Table 2. Few-shot detection performance (mAP@0.5) on Pascal VOC for novel classes - Class Split 2

Method \ Shots	1	2	3	5	10
FSRW [7]	15.7	15.3	22.7	30.1	40.5
Meta Det [20]	21.8	23.1	27.8	37.7	43.0
Meta R-CNN [22]	10.4	19.4	29.6	34.8	45.4
TFA w/ fc [19]	18.2	29.0	33.4	35.5	39.0
TFA w/ cos [19]	23.5	26.9	34.1	35.1	39.1
MPSR [21]	24.4	29.5	39.2	39.9	47.8
Retentive R-CNN [5]	21.7	27.8	35.2	37.0	40.3
CME [9]	27.2	30.2	41.4	42.5	46.8
SRR-FSD [25]	32.5	35.3	39.1	40.8	43.8
FSCE [16]	27.3	29.5	43.5	44.2	50.2
QA-FewDet [6]	25.9	37.8	46.6	48.9	51.1
Meta-DETR [24]	37.0	36.6	43.7	49.1	54.6
FS-DETR [1]	37.3	41.3	43.4	46.6	49.0
HG-DETR (OURS)	35.4	46.6	50.9	52.6	61.3

2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Table 3. Few-shot detection performance (mAP@0.5) on Pascal VOC for novel classes - Class Split 3

Method \ Shots	1	2	3	5	10
FSRW [7]	21.3	25.6	28.4	42.8	45.9
Meta Det [20]	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [22]	14.3	18.2	27.5	41.2	48.1
TFA w/ fc [19]	27.7	33.6	42.5	48.7	50.2
TFA w/ cos [19]	30.8	34.8	42.8	49.5	49.8
MPSR [21]	35.6	40.6	43.2	48.0	49.7
Retentive R-CNN [5]	30.2	37.6	43.0	49.7	50.1
CME [9]	34.3	39.6	45.1	48.3	51.5
SRR-FSD [25]	40.1	41.5	44.3	46.9	46.4
FSCE [16]	37.2	41.9	47.5	54.6	58.5
QA-FewDet [6]	35.2	42.9	47.8	54.8	53.5
Meta-DETR [24]	41.6	45.9	52.7	58.9	60.6
FS-DETR [1]	43.8	47.1	50.6	52.1	56.9
HG-DETR (OURS)	39.5	48.4	50.4	58.8	61.0

Table 4. Few-shot detection performance on COCO for novel classes - 1-shot and 2-shot

	1-Shot			2-shot		
method	AP0.5:0.95	AP0.5	AP0.75	AP0.5:0.95	AP0.5	AP0.75
TFA w/ fc [19]	2.9	5.7	2.8	4.3	8.5	4.1
TFA w/ cos [19]	3.4	5.8	3.8	4.6	8.3	4.8
MPSR [21]	2.3	4.1	2.3	3.5	6.3	3.4
QA-FewDet [6]	4.9	10.3	4.4	7.6	16.1	6.2
Meta-DETR [24]	7.5	12.5	7.7	-	-	-
FS-DETR [1]	7.0	13.6	7.5	8.9	17.5	9.0
HG-DETR (OURS)	7.9	13.7	7.8	9.3	16.4	9.2

Table 5. Few-shot detection performance on COCO for novel classes - 3-shot and 5-shot

	3-Shot			5-shot		
method	AP0.5:0.95	AP0.5	AP0.75	AP0.5:0.95	AP0.5	AP0.75
TFA w/ fc [19]	6.7	12.6	6.6	8.4	16.0	8.4
TFA w/ cos [19]	6.6	12.1	6.5	8.3	15.3	8.0
MPSR [21]	5.2	9.5	5.1	6.7	12.6	6.4
QA-FewDet [6]	8.4	18.0	7.3	9.7	20.3	8.6
Meta-DETR [24]	13.5	21.7	14.0	15.4	25.0	15.8
FS-DETR [1]	10.0	18.8	10.0	10.9	20.7	10.8
HG-DETR (OURS)	13.5	22.5	13.2	15.5	26	16

Table 6. Few-shot detection performance on COCO for novel classes - 10-shot and 30-shot

	10-shot			30-shot		
method	AP0.5:0.95	AP0.5	AP0.75	AP0.5:0.95	AP0.5	AP0.75
FSRW [7]	5.6	12.3	4.6	9.1	19.0	7.6
Meta Det [20]	7.1	14.6	6.1	11.3	21.7	8.1
Meta R-CNN [22]	8.7	19.1	6.6	12.4	25.3	10.8
TFA w/ fc [19]	10.0	19.2	9.2	13.4	24.7	13.2
TFA w/ cos [19]	10.0	19.1	9.3	13.7	24.9	13.4
MPSR [21]	9.8	17.9	9.7	14.1	25.4	14.2
CME [9]	15.1	24.6	16.4	16.9	28.0	17.8
SRR-FSD [25]	11.3	23.0	9.8	14.7	29.2	13.5
FSCE [16]	11.1	-	9.8	15.3	-	14.2
QA-FewDet [6]	11.6	23.9	9.8	16.5	31.9	15.5
Meta-DETR [24]	19.0	30.5	19.7	22.2	35.0	22.8
FS-DETR [1]	11.3	21.7	11.1	-	-	-
HG-DETR (OURS)	18.8	30.1	19.3	21.5	35.2	22.7

Component Ablation on Graph Modules To further assess the individual contributions of the proposed Cross-category semantic relationship graph(CSRG) and Query-level context aggregation graph(QCAG), we conduct a detailed component ablation study on Pascal VOC Split 1, evaluating performance under 1, 2, 3, 5, and 10-shot settings using mAP@50. As shown in Table 7, using only the Cross-category semantic relationship graph yields solid performance, indicating that modeling semantic correlations among class prototypes benefits the support-to-query transfer, especially in low- data regimes. In contrast, employing only the Query-level context aggregation graph achieves slightly lower performance across all settings, suggesting that contextual modeling within the query image is helpful but less dominant on its own. When both modules are enabled, we observe consistent and noticeable improvements across the board, confirming that inter-class semantic reasoning and intra-image relational reasoning are complementary in enhancing few-shot object detection. These results demonstrate the necessity of jointly modeling both class-level and spatial-level structures to maximize generalization in few-shot detection tasks.

Table 7. Ablation studies on Pascal VOC Split 1

CSRG	QCAG	1-shot	2-shot	3-shot	5-shot	10-shot
×	×	36.2	45.8	51.3	58.0	60.5
✓	×	38.6	51.9	53.7	62.6	63.8
×	✓	37.5	49.4	52.4	58.8	62.0
✓	✓	39.1	53.3	54.3	63.2	64.6

2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

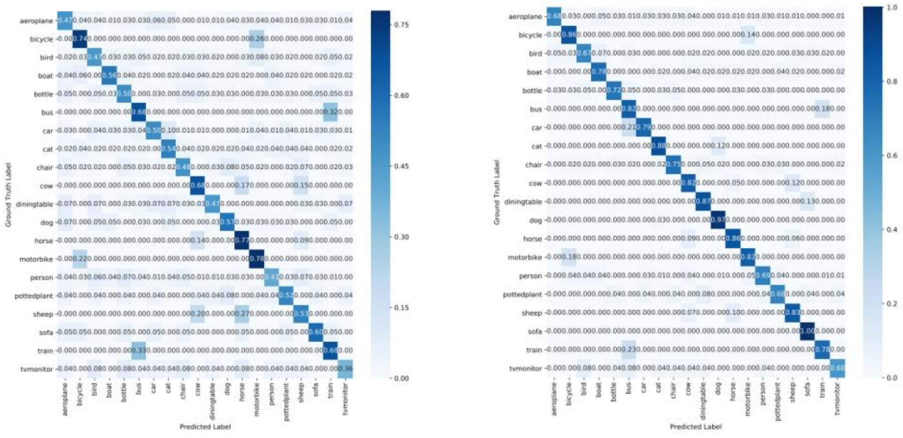
Effect of Bidirectional Class-Query Adaptation To validate the design of our Bidirectional Class-Query Adaptation (BCQA) mechanism, we conduct an ablation study on Pascal VOC Split 1. Specifically, we isolate the two directional interactions: (1) Class-to-Query adaptation, and (2) Query-to-Class adaptation. Table 8 reports the results. Experimental results show that the bidirectional design consistently outperforms both single-direction counterparts across all few-shot settings, highlighting the benefit of mutual and iterative feature alignment. This confirms the effectiveness of our design and the necessity of bidirectional adaptation for enhancing few-shot generalization.

Table 8. Bidirectional adaptation ablation on Pascal VOC Split 1

	1-shot	2-shot	3-shot	5-shot	10-shot
Class-to-Query Only	37.8	50.5	51.5	61.2	63.4
Query-to-Class Only	38.6	52.4	53.0	62.6	63.6
Full Bidirectional	39.1	53.3	54.3	63.2	64.6

6.4 Experimental Results Visualization

To demonstrate the superiority of our HG-DETR method, we visualize the con- fusion matrices of both the baseline model and our approach under the 10-shot setting on the PASCAL VOC split1. The results are presented in Figure 4.



(a) Baseline Model (b) HG-DETR Method

Fig. 4. Confusion matrices comparing the baseline model and HG-DETR method in a 10-shot setting on PASCAL VOC split1. Our HG-DETR demonstrates superior detection performance, particularly in distinguishing similar classes such as cow vs. sheep, motorbike vs. bicycle.

The confusion matrix of our HG-DETR method demonstrates remarkable detection performance. Particularly, the advantage becomes more pronounced when dealing with similar classes such as cow vs. sheep, motorbike vs. bicycle, and horse vs. cow. This enhanced performance in distinguishing similar categories can be attributed to the Cross-category semantic relationship graph we introduced, which effectively captures the relationships between different classes and improves the model's detection accuracy for objects with similar visual features.

7 Conclusion

In this paper, we present HG-DETR, a novel few-shot object detection frame- work that unifies image-level detection with hierarchical graph reasoning. Our method eliminates region proposals and introduces three key components: a cross-category semantic relationship graph to model cross-category dependencies, enhancing prototype discrimin ability by transferring knowledge from base classes; a query-level context aggregation graph to aggregate local overlaps and class prototype, refining proposals; and bidirectional feature adaptation to align prototypes and queries through attention mechanisms, reducing domain gaps. Our model, HG-DETR, outperforms existing methods, and we hope this work can offer good insights and inspire further researches in few-shot object detection and other related topics.

References

1. Bulat, A., Guerrero, R., Martinez, B., Tzimiropoulos, G.: Fs-detr: Few-shot detection transformer with prompting and without re-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11793–11802 (2023)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End- to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, J., Lei, B., Song, Q., Ying, H., Chen, D.Z., Wu, J.: A hierarchical graph network for 3d object detection on point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 392–401 (2020)
4. Dai, Z., Cai, B., Lin, Y., Chen, J.: Up-detr: Unsupervised pre-training for object detection with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1601–1610 (2021)
5. Fan, Z., Ma, Y., Li, Z., Sun, J.: Generalized few-shot object detection without forgetting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4527–4536 (2021)
6. Han, G., He, Y., Huang, S., Ma, J., Chang, S.F.: Query adaptive few-shot object detection with heterogeneous graph convolutional networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3263–3272 (2021)
7. Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., Darrell, T.: Few-shot object detection via feature reweighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8420–8429 (2019)
8. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
9. Li, B., Yang, B., Liu, C., Liu, F., Ji, R., Ye, Q.: Beyond max-margin: Class mar- gin equilibrium for few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7363–7372 (2021)

2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

10. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: Dab-detr: Dynamic anchor boxes are better queries for detr. arXiv preprint arXiv:2201.12329 (2022)
11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. pp. 21–37. Springer (2016)
12. Mi, L., Chen, Z.: Hierarchical graph attention network for visual relationship detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13886–13895 (2020)
13. Nawhal, M., Mori, G.: Activity graph transformer for temporal action localization. arXiv preprint arXiv:2101.08540 (2021)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. IEEE transactions on pattern analysis and machine intelligence 39(6), 1137–1149 (2016)
16. Sun, B., Li, B., Cai, S., Yuan, Y., Zhang, C.: Fsce: Few-shot object detection via contrastive proposal encoding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7352–7362 (2021)
17. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al.: Graph attention networks. stat 1050(20), 10–48550 (2017)
18. Wang, X., Ye, Y., Gupta, A.: Zero-shot recognition via semantic embeddings and knowledge graphs. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6857–6866 (2018)
19. Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly simple few-shot object detection. arXiv preprint arXiv:2003.06957 (2020)
20. Wang, Y.X., Ramanan, D., Hebert, M.: Meta-learning to detect rare objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9925–9934 (2019)
21. Wu, J., Liu, S., Huang, D., Wang, Y.: Multi-scale positive sample refinement for few-shot object detection. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XVI 16. pp. 456–472.Springer (2020)
22. Yan, X., Chen, Z., Xu, A., Wang, X., Liang, X., Lin, L.: Meta r-cnn: Towards general solver for instance-level low-shot learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9577–9586 (2019)
23. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7094–7103 (2019)
24. Zhang, G., Luo, Z., Cui, K., Lu, S., Xing, E.P.: Meta-detr: Image-level few-shot detection with inter-class correlation exploitation. IEEE transactions on pattern analysis and machine intelligence 45(11), 12832–12843 (2022)
25. Zhu, C., Chen, F., Ahmed, U., Shen, Z., Savvides, M.: Semantic relation reasoning for shot-stable few-shot object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8782–8791 (2021)
26. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159 (2020)