Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

HST-FSOD: Hierarchical semantic transfer ofbase-to-novel and fine-grained semantic prototypeencoding for few-shot object detection

Junchi Su
Beijing University of Posts and Telecommunications
Xin Gao
x1hhh74@bupt.edu.cn

Beijing University of Posts and Telecommunications
Xinping Diao
China Electric Power Research Institute Company Limited
Yuan Li
China Electric Power Research Institute Company Limited
Tianyang Chen
China Electric Power Research Institute Company Limited
Yukun Lin
China Electric Power Research Institute Company Limited

Research Article

Keywords: Few-shot object detection, Vision-language models, Hierarchical semantic transfer, Fine-grained semantic prototype encoding module, Knowledge transfer

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8706735/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

HST-FSOD: Hierarchical semantic transfer of base-to-novel and fine-grained semantic prototype encoding for few-shot object detection

Junchi Su[1], Xin Gao[1*], Xinping Diao[2], Yuan Li[2], Tianyang Chen[2], Yukun Lin[2]

[1*]School of Intelligent Engineering and Automation, Beijing University of Posts and Telecommunications, Beijing, 100876, China.
[2]Metrology Research Institute, China Electric Power Research Institute Company Limited, Beijing, 100192, China.

*Corresponding author(s). E-mail(s): xlhhh74@bupt.edu.cn;
Contributing authors: sujunchi@bupt.edu.cn; diaoxp@epri.sgcc.com.cn; liyuan3@epri.sgcc.com.cn; chentianyang@epri.sgcc.com.cn; linyukun@epri.sgcc.com.cn;

Abstract

Few-shot object detection (FSOD) aims to recognize and localize objects of novel classes using only a very limited number of annotated samples. Existing FSOD methods based on vision-language models map textual and visual data into a shared embedding space, implicitly establishing a linear association between class semantics and visual features. However, these approaches overlook the hierarchical semantic relationships among categories and struggle to extract discriminative fine-grained features from limited textual descriptions, thereby constraining model generalization under data-scarce scenarios. To address these limitations, this paper proposes HST-FSOD, a novel framework that incorporates **H**ierarchical **S**emantic **T**ransfer of base-to-novel and fine-grained semantic prototype encoding for **F**ew-**S**hot **O**bject **D**etection. The Hierarchical Semantic Transfer module first constructs a visual-semantic embedding similarity matrix to assign all classes into superclass labels. Features of subclasses within the same superclass group are aggregated to form a superclass representation. Subsequently, similarities among superclass features are computed recursively to build a hierarchical tree, yielding multi-level category sets with progressively finer granularity. Each novel class is then associated with a semantically similar base class as its superclass label, establishing semantic connections between

data-scarce novel classes and data-rich base classes within a multi-level structure, thereby facilitating effective knowledge transfer via inter-class commonalities. The fine-grained semantic prototype encoding module generates per-image textual descriptions that include both class names and contextual details, which are encoded by a pre-trained text encoder to produce fine-grained semantic embeddings. This enhances the model's capacity to perceive intra-class variations and fine-grained characteristics. Experiments on two benchmark datasets demonstrate that HST-FSOD significantly improves detection accuracy in few-shot settings and outperforms state-of-the-art baseline methods.

Keywords: Few-shot object detection, Vision-language models, Hierarchical semantic transfer, Fine-grained semantic prototype encoding module, Knowledge transfer

1 Introduction

Traditional object detection methods achieve high precision and real-time multi-class detection in complex natural scenes, significantly accelerating the transition of computer vision from laboratory research to industrial deployment[1]. However, these methods heavily rely on large-scale high-quality annotations. In data-scarce scenarios, models are prone to overfitting, leading to severe performance degradation. Moreover, in the real world, object categories often follow a long-tailed distribution—common classes are well-represented, while rare categories (e.g., endangered animals, specialized industrial components, or rare lesions) have very few samples[2]. Consequently, traditional detectors perform poorly on tail classes. Few-shot object detection alleviates this issue by leveraging mechanisms such as meta-learning, feature alignment, and knowledge transfer to learn novel categories from minimal samples[3].

The core objective of FSOD is to accurately detect objects of novel classes given only a few annotated instances (typically 1–10 per class). Existing FSOD approaches are broadly categorized into meta-learning and fine-tuning paradigms[4]. Meta-learning methods simulate test-time scenarios during training by constructing episodic tasks: in each episode, a support set (containing a few samples from randomly selected base classes) and a query set (with labeled instances of the same classes) are sampled. The model learns to adapt rapidly to new tasks by optimizing detection performance on the query set. At test time, it generalizes to novel classes without retraining. While this "task-driven" paradigm enhances generalization, it suffers from high training complexity due to frequent data sampling, limiting practical deployment[5].

In contrast, fine-tuning methods adopt a two-stage training process: first pretraining on base classes, then fine-tuning on novel classes[6]. This approach avoids complex inference procedures and is more deployment-friendly. The key challenge lies in effectively transferring knowledge from base to novel classes. Current fine-tuning strategies include parameter transfer, knowledge distillation, and feature transfer. Parameter transfer methods treat base-class knowledge as model weights, initializing novel-class training from a base-class-pretrained model and fine-tuning a subset of

parameters. For example, TFA[6] freezes part of the network and fine-tunes on a balanced dataset of base and novel classes to mitigate catastrophic forgetting. However, this parameter transfer is implicit and lacks explicit modeling of semantic relationships, making it susceptible to overfitting or forgetting. Knowledge distillation methods use a pre-trained "teacher" model to provide supervision signals during fine-tuning. PA-BoVW[7] uses an independent pre-trained model to offer "non-overfitting" semantic priors via a visual bag-of-words representation. D&R[8] introduces causal intervention theory to address confounding bias in distillation by decoupling semantics from spurious correlations. While effective, these methods require maintaining a teacher model (increasing computational cost) and still transfer knowledge implicitly without explicitly modeling inter-class associations.

Feature transfer methods aim to learn a universal, discriminative feature space from base classes, which is directly used for novel-class classification or matching. During the initial training stage, FSCE [9] adopts contrastive learning to regularize the embedding space. This strategy strengthens within-class cohesion and promotes class-wise distinctness, providing a more robust foundation for subsequent few-shot fine-tuning. DeFRCN[10] decouples RPN and detection head features to ensure RPN proposals are class-agnostic yet semantically rich. However, these methods largely ignore the hierarchical structure inherent in real-world categories. For instance, "vehicle" is a broad superclass encompassing semantically related but visually distinct subclasses like cars, bicycles, and motorcycles. With only a few samples, models easily confuse visually similar subclasses (e.g., mistaking a moped for an e-bike). This fine-grained discrimination challenge is critical in FSOD, yet existing feature transfer approaches fail to exploit hierarchical inter-class relationships that could significantly boost generalization.

To address these gaps, this paper proposes HST-FSOD, a base-to-novel hierarchical semantic transfer and fine-grained semantic prototype encoding framework. Our Hierarchical Semantic Transfer Module first computes a visual-semantic similarity matrix to assign all classes into superclasses. Classes with high similarity are grouped into a cluster (subtree), which is recursively merged based on superclass similarity to construct a hierarchical tree. Each level of the tree corresponds to a superclass set with progressively finer semantic granularity. This enables data-scarce novel classes to establish multi-level semantic links with data-rich base classes, enhancing generalization. Moreover, while semantic encoding improves class sensitivity, generic textual descriptions often lack sufficient discriminative power for cross-category generalization. Thus, a fine-grained semantic prototype encoding module is introduced, comprising a contrastive description generator, a text encoder, and a semantic prototype bank. This module generates fine-grained visual descriptions for each image, encodes them into semantic embeddings, and stores these embeddings in the prototype bank. These embeddings offer tailored semantic guidance for visual-text alignment within the hierarchical module, thereby significantly enhancing the model's perception of intra-class variation and fine-grained features. The main contributions of this paper are as follows:

- A base-to-novel hierarchical semantic transfer module is proposed, comprising three key steps: constructing a visual-text similarity matrix via cosine similarity between image embeddings and semantic prototypes; recursively merging semantically similar categories into superclass clusters to build a multi-level hierarchical tree; and

performing hierarchical classification to establish semantic links between base and novel classes, thereby transferring category-agnostic knowledge from base to novel classes and enhancing recognition accuracy and generalization.

- A fine-grained semantic prototype encoding module is proposed. Its core components include a contrastive description generator, a text encoder, and a semantic prototype bank. The core function of this module is to first generate fine-grained visual descriptions containing intra-class difference information for the training images, then semantically encode these descriptions through the text encoder and store them in the semantic prototype bank, and finally provide effective support for the visual-text alignment of the hierarchical semantic module, thereby enhancing the model's ability to perceive intra-class differences and fine-grained features and its generalization ability.
- Single-round and multi-round experimental results show that the proposed method achieves state-of-the-art performance on the PASCAL VOC and MS-COCO benchmark datasets. The feature visualization experiments further demonstrate that the proposed method effectively enhances inter-class separability between base and novel categories, mitigates confusion during the transfer of shared semantic knowledge, and consequently boosts detection accuracy for both base and novel classes. This verifies effectiveness of our proposed superclass learning strategy and semantic description generation module in improving model performance.

The remainder of this paper is organized as follows: Section 2 introduces related work on few-shot object detection and hierarchical learning. Section 3 introduces the overall framework of the proposed method, including the proposed hierarchical superclass learning strategy for few-shot object detection and a fine-grained semantic prototype encoding module. Section 4 describes four carefully designed experiments, including quantitative comparison between the novel class and the base class, visual comparison, feature visualization comparison, ablation experiments. Section 5 summarizes the contributions of this paper.

2 Related work

2.1 Few-shot object detection

In recent years, few-shot object detection has attracted widespread attention from researchers due to its significant challenges. The core objective of this task is to achieve accurate classification and localization of objects from novel classes when only one or a few instances of each novel class are provided[5]. Based on differences in model training paradigms, existing approaches are primarily categorized into two main types: meta-learning based methods and fine-tuning based methods. Meta-learning based methods construct multiple episode-level tasks to acquire task-level knowledge, thereby enhancing generalization to novel-class samples. In contrast, fine-tuning based methods typically adopt a two-stage training strategy: first pre-training on a base-class dataset, followed by fine-tuning on a novel-class dataset[6].

2.1.1 Meta-learning based methods

Most meta-learning methods[11–18] treat tasks as the fundamental unit, with each task comprising a support set and a query set. The training process is further divided into meta-training and meta-testing phases. The goal of this learning paradigm is to acquire task-agnostic prior knowledge from a large number of tasks during meta-training, enabling the model to achieve high detection accuracy on novel-class datasets with only a few optimization steps during meta-testing. Yan et al.[11] proposed a prediction head reshaping network (PRN) as an auxiliary branch on top of Faster R-CNN, leveraging meta-learning to infer class-specific attention vectors and thereby enhancing the representational capacity of region proposal features. Kang et al.[12] introduced a reweighting module into a single-stage detection network to automatically generate weighted features for novel-class detection, improving the generalization of meta-features. Zhang et al.[13] were the first to integrate a meta-learning framework into the end-to-end detection architecture DETR, proposing a correlation aggregation module that significantly reduces misclassification errors. Han et al.[14] extended the meta-learning baseline by introducing a variational autoencoder into few-shot object detection for the first time, presenting a meta-learning framework that incorporates both class-agnostic aggregation (CAA) and variational feature aggregation (VFA), which mitigates inter-class confusion while promoting better inter-class interaction. Lu et al.[16] proposed a conditional information coupling module that integrates query-branch information into the support branch, thereby enhancing query-aware representations within the support features. Conventional meta-learning approaches typically encode support features into class-level prototypes and fuse these prototypes with the query branch to facilitate prediction. However, such class-level prototypes often lack fine-grained details. Wang et al.[18] proposed a fine-grained feature aggregation that distills support images into fine-grained prototypes and assigns them to query images, effectively leveraging relationships among detailed features to help the model capture critical discriminative cues.

In summary, existing approaches primarily adopt the support-query episodic framework, with the goal of acquiring transferable prior knowledge during meta-training to enable fast generalization to novel categories at test time. Nevertheless, these methods suffer from several limitations: First, they compress the support set into a single class prototype, resulting in the loss of fine-grained and spatial structural information, which hinders the discrimination of visually similar categories. Second, the nested optimization strategy introduces high training complexity and is prone to overfitting under extremely limited sample conditions, leading to a noticeable drop in accuracy. More critically, current meta-learning frameworks do not explicitly disentangle causal semantic factors from confounding variables such as background clutter or domain shifts. Consequently, models are vulnerable to spurious correlations in novel scenarios, resulting in insufficient robustness.

2.1.2 Fine-tuning based methods

Two-stage fine-tuning methods do not require the design of complex proxy tasks; instead, they only require fine-tuning training on the target dataset. Compared to

meta-learning methods, fine-tuning training is simpler and easier to converge. For standard few-shot object detection, Yang et al.[19] proposed an efficient framework, PTF, without computational increment, and a knowledge inheritor, KI, proposed on this framework, which effectively promotes the transfer of base class knowledge and the adaptation speed of novel classes. Zhang et al.[20] used learnable tensors to encode general features and specific feature embeddings explicitly, and proposed a general and specific embedding learning method based on Transformer. To explore the impact of causal relationships on detection, the FR method[21] constructs a structural causal model containing causal generation factors and pseudo-generation factors, and performs semantic grouping and semantic decoupling through a front-door regulator. However, these fine-tuning methods do not simultaneously consider the recognition accuracy of both base classes and novel classes, and the performance of base classes tends to decrease during the fine-tuning process.

Unlike standard few-shot object detection, generalized few-shot object detection tasks aim to maintain the recognition accuracy of both base class and novel class. When facing test samples with both base class and novel class instances, generalized few-shot object detection modes have more practical application value. To resolve the contradiction between multi-stage and multi-task, Qiao et al. [10] employed gradient decoupling layers and prototype calibration blocks to decouple multiple detection modules, thereby significantly enhancing the baseline performance of few-shot object detection. Wu et al.[22] proposed a distillation framework based on a memory bank to explicitly learn the class-independent commonalities between base class and novel class. Wang et al.[6] proposed a TFA method, which adopts a two-stage paradigm. By freezing most layers in Faster RCNN, only the last layer is fine-tuned during the fine-tuning process, and the base class knowledge learned in the pre-training stage is transferred to the novel class in the fine-tuning stage. In addition, to address the problem of base class forgetting, the novel class is fine-tuned first, and then the weights of the novel class are used as initialization to fine-tune both the base class and the novel class simultaneously, to reduce the degree of base class forgetting as much as possible while adapting to the novel class. Since RCNN and RPN in Faster RCNN extract translation-invariant and translation-variable features, respectively, to resolve the contradiction between the two, and considering the confusion of hard negative samples with foreground samples, Bsdet[23] improved upon by proposing a binary similarity few-shot detector, which transforms the multi-class task into multiple binary classification tasks, and designed a novel feature enhancement method for hard negative samples. However, the hierarchical relationship between categories can effectively improve the performance of the model in the few-shot learning scenario. Existing fine-tuning methods ignore the complex hierarchical relationship between categories, making it difficult to learn the common knowledge between categories, resulting in confusion or missed detection in the recognition.

2.2 Semantic-guided FSOD with external knowledge

In recent years, some methods for few-shot object detection have utilized external knowledge sources, usually pre-trained language models or structured knowledge bases, to obtain semantic representations of object categories. This semantic information,

as a kind of "side information," is independent of the limited visual samples and provides guidance for the model's learning process[24]. To address the problem that limited training data cannot explore semantic information, Li et al.[8] integrated category text information into visual information through a knowledge distillation learning paradigm, thereby improving the ability of object detection models to obtain semantic side information. To better understand the relationship between text-based information and visual features and reduce dependence on specific knowledge sources, Du et al.[25] designed an image-text matching module to establish the association between text and visual features. Due to the limited sample size, the features learned by few-shot object detection models often deviate from the category center. To address this problem, Guo et al.[26] introduced domain knowledge composed of CLIP[27] and attribute knowledge graphs to obtain semantic and visual information to construct a discriminative prototype based on decoupling contrast, thereby making the category semantic boundaries clearer. Wang et al.[28] proposed a few-shot data augmentation method using semantic side-channel information, which improves the diversity of samples by fusing instances into different backgrounds, thereby expanding the training data. Zhao et al.[29] provided a distillation scheme based on a visual language model and designed an extended base-novel classifier to align with the distribution of the visual semantic model. However, existing semantic guidance methods ignore the knowledge forgetting problem of base classes. The model pays too much attention to the semantic information in novel samples, which may unconsciously cover or weaken the memory of existing knowledge of base classes, thereby affecting the accuracy of base class detection.

2.3 Hierarchical classification

Hierarchical classification is a method for classifying data based on the hierarchical relationships between categories. It can efficiently organize and process large-scale data, but it is also a challenging task. Its core lies in systematically organizing categories according to a predefined hierarchical structure. This task typically involves three key steps: first, constructing the hierarchical structure; second, selecting the hierarchical classification level; and finally, learning the hierarchical classifier.

Major work on constructing the hierarchical structure falls into three main categories. First, it utilizes classifier scores for construction. Griffin et al.[30] proposed a method to automatically construct hierarchical classification trees using semantic or statistical correlations between categories to accelerate large-scale visual classification and improve generalization ability. Liu et al.[31] used a greedy partitioning strategy to construct a binary tree structure. Second, it focuses on feature selection for hierarchical classification. Freeman et al.[32] proposed a method for joint optimization of feature selection and hierarchical classifier structure design, aiming to build an efficient, accurate, and computationally inexpensive hierarchical classification system. Zhao et al.[33] proposed a hierarchical feature selection method combining data features and achieved structure-aware feature selection through recursive regularization. For learning hierarchical classifiers, Xie et al.[34] introduced a hierarchical deep learning framework of superclasses to improve fine-grained image classification; Guillaumin et al.[35] cleverly utilized the inherent semantic hierarchical structure and transferred localization

knowledge from semantically related "source class" to "target class", thereby realizing a hierarchical semantic implicit hierarchical reasoning mechanism. Recently, research has focused on integrating inherent semantic insights derived from category hierarchies into visual representations to steer the classification process. A representative approach is found in [36], which constructs a visual-semantic embedding framework to bridge the gap between linguistic knowledge and visual object identification. The literature[37] proposed mapping images to class embeddings to learn semantic discriminative features. Furthermore, some methods have introduced multi-task frameworks to efficiently solve deep hierarchical classification problems[38, 39].

For hierarchical classification tasks in few-shot object detection, Reference [40] first introduced hierarchical learning into the field of few-shot object detection, proposing a hierarchical contrastive learning method to constrain the feature space and effectively improve the generalization ability to scarce classes. However, this method uses a path product strategy, which is susceptible to early error propagation. Once a high-level node makes a mistake, even if the low-level features are correct, the entire path will be suppressed. Therefore, how to fully utilize the hierarchical semantic information between classes to enhance the model's generalization ability to novel classes while avoiding early error propagation remains a key problem that hierarchical classification urgently needs to solve in the field of few-shot object detection.

3 Method

In light of the fact that existing methods overlook the hierarchical relationships among categories—thereby preventing novel classes from effectively leveraging the abundant knowledge embedded in base classes—this paper proposes a base-to-novel hierarchical transfer strategy. This strategy progressively uncovers commonalities between base and novel classes and assigns a superclass label to each category for supervised learning. Furthermore, to address the limitation of prior semantic-guided approaches—whose textual descriptions struggle to generalize effectively across all categories—a fine-grained semantic encoding module is introduced. By integrating image encoding with a cross-modal semantic encoder, this module generates customized textual descriptions for each image, thereby capturing more generalizable semantic information. Accordingly, Section 3.1 first presents the task definition of few-shot object detection. Section 3.2 then elaborates on the model overview of the proposed method, followed by detailed descriptions in Sections 3.3 and 3.4 of the hierarchical learning procedure for superclass partitioning and the fine-grained semantic encoding module, respectively.

3.1 Task definition

In generalized few-shot object detection, the label space is partitioned into base classes (D_b, C_b) and novel classes (D_n, C_n), where $D = D_b \cup D_n, D_b \cap D_n = \emptyset; C = C_b \cup C_n, C_b \cap C_n = \emptyset$. The training process of generalized few-shot object detection models consists of two stages. In the first stage, the model is trained on a large-scale dataset of base classes to acquire robust foundational knowledge. This pre-training helps reduce the risk of catastrophic forgetting—specifically, the model's degradation

in base-class classification performance—and also alleviates overfitting to the scarce novel-class samples during subsequent fine-tuning. In the second stage, the model is fine-tuned on a balanced few-shot dataset that includes both base and novel classes.

Specifically, in a K-shot few-shot object detection task, each novel class is provided with exactly K annotated samples. To avoid overfitting to these scarce novel-class instances, the generalized few-shot detection framework performs fine-tuning on a balanced dataset in which every base class and every novel class is represented by exactly K samples. This balanced sampling strategy enables the model to rapidly adapt to and learn from the novel classes while simultaneously preserving its knowledge of the base classes.

3.2 Model overview

Few-shot object detection aims to enable models to recognize novel classes with only a limited number of samples. Existing few-shot object detection approaches typically rely on a two-stage fine-tuning paradigm to implicitly transfer knowledge from base classes to novel classes. However, these methods overlook the hierarchical semantic relationships between base and novel categories. To address this limitation, this paper proposes a framework that explicitly captures such hierarchical structure through superclass labeling and leverages a fine-grained semantic prototype module to generate detailed textual descriptions for guiding prototype construction. The overall architecture of the proposed model is illustrated in Figure 1. As shown in the figure,

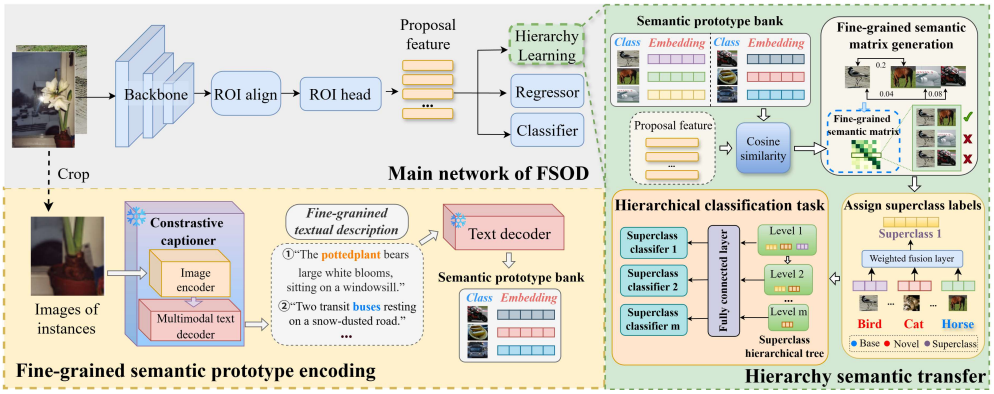

Fig. 1 Overall flowchart of the proposed method. The diagram includes the main network, hierarchical learning, and fine-grained semantic prototype encoding module. First, the input image is cropped and fed into the fine-grained semantic prototype encoding module to generate semantic prototypes. Second, the proposed region features from the backbone network are input into the hierarchical learning module to construct a hierarchical tree.

the proposed method—termed HST-FSOD integrates hierarchical semantic transfer with fine-grained semantic prototype encoding for few-shot object detection. Given an input image, two parallel processing streams are executed: The image is fed into a backbone network and a region proposal network (RPN) to extract region-of-interest

(RoI) features; Simultaneously, the same image is passed through a fine-grained prototype generator to produce a customized, fine-grained textual description. This textual description is then encoded by a semantic encoder to yield a semantic embedding, which is stored in a prototype module. Next, a similarity matrix is constructed by computing pairwise similarities among the semantic embeddings in the prototype module. Finally, this similarity matrix is input into a hierarchical learning module to perform superclass partitioning.

The training process consists of two stages: base-class training and fine-tuning. To maximize the effectiveness of the proposed components, different strategies are employed in each stage. During the base-class training stage, only the fine-grained prototype generator is utilized to produce fine-grained semantic descriptions and category-level prototypes for base classes, while the hierarchical learning module is deactivated. In the subsequent fine-tuning stage, both the hierarchical superclass learning strategy and the fine-grained prototype generator are activated jointly: the model carries out hierarchical superclass assignment and produces tailored fine-grained semantic descriptions, facilitating structured knowledge transfer from base to novel classes by leveraging explicit semantic relationships rather than relying on generic meta- or metric-learning paradigms.

3.3 Hierarchical superclass learning

The proposed method aims to use the similarity between text descriptions as nodes in a tree, and recursively construct a hierarchical tree based on the similarity values to assign a superclass label to each category. Supervised classification training is then performed using these superclass labels. The learning strategy can be broken down into three steps: generating a similarity matrix, constructing a hierarchical superclass tree, and supervised training for hierarchical classification.

Step 1: Generating the Similarity Matrix. As shown in the left image of Figure 3, the similarity matrix is calculated using proposed region features and semantic prototypes. First, we discuss proposed region features. The backbone network extracts the feature map of the input image, which is then processed by the region generation network to generate corresponding proposed region features. Proposed region features can be considered as features of each instance. Since each instance has a true class label, each region can be regarded as a true class instance. To compress each region feature into an image embedding that represents the class and facilitates matching with the subsequent semantic prototype dimension, the output proposed region features are mapped and compressed into several low-dimensional image embeddings. Since the model training in this paper uses batch training, each batch will contain multiple image embeddings of the same class. To construct image embeddings from local to global to achieve more robust image representation, this paper performs average pooling on multiple identical image embeddings, ultimately obtaining an image embedding matrix consistent with the number of classes, where each row represents the pooled image embedding. Next, the semantic prototype is generated by a fine-grained prototype generator, as detailed in Section 3.4. The final similarity matrix is calculated from the semantic prototype and the image embeddings. Specifically, both image embeddings I and semantic prototypes S are matrices of dimension $[C, D]$, where C represents

the number of categories and D represents the dimension of each image embedding or semantic prototype. The similarity matrix $M(i, j)$ can then be obtained using the following formula:

$$M(i, j) = I \cdot S^T, \tag{1}$$

where i and j represent the indices in the matrix. To ensure that the similarity output is stable in the $[0, 1]$ interval, which is beneficial to the convergence of model training, the result of each row of the similarity matrix is Softmax normalized so that the sum of all elements in each row is 1.

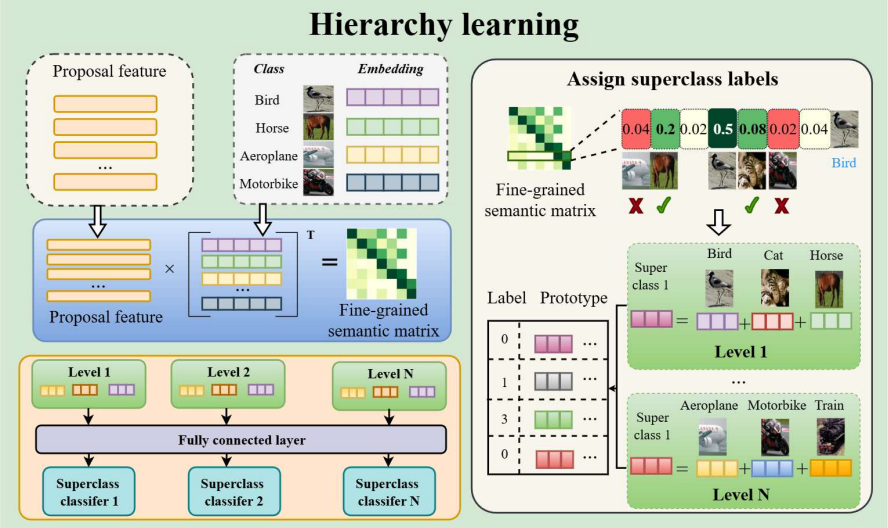

Fig. 2 Detailed diagram of hierarchical learning for superclasses. The left side of the figure shows the construction method of the similarity matrix and the hierarchical classification task, while the right side shows the merging process of the hierarchical trees.

Step 2: Construct a hierarchical superclass tree. As shown in Figure 3 (right), the superclass label for each category is obtained from the similarity matrix. For example, the i-th row represents the similarity between the image embeddings of all categories and the i-th base class semantic prototype. The K image embeddings with the highest scores are taken as subsets of the semantic prototype (for the optimal setting of hyperparameter K, see the parameter sensitivity experiment results). Since the score represents the similarity between categories, the sequence number of the semantic prototype is regarded as the superclass label of the subset of that category. Then, the K image embeddings with the highest similarity to each base class semantic prototype are calculated sequentially, and the corresponding semantic prototype category number is set as the superclass label. It is worth noting that since an image embedding may have high similarity with multiple semantic prototypes, this can lead to a category belonging to multiple superclasses. Therefore, to reduce this ambiguity, the same label will be reassigned to a superclass label based on the similarity; that is, the semantic prototype category number with the highest similarity will be used as the

superclass label of that category. In addition, to represent the category at each level, the merged superclass label will continue to iterate upwards. Each merged superclass embedding SP_j^L is obtained from its subordinate sub-category embeddings, as shown in the following formula:

$$SP_j^L = \sum_{i=1}^{K} s_i \cdot SP_i^{L-1}, \tag{2}$$

where L represents the level of the hierarchical tree to which the subclass belongs, and i and j represent the i-th subclass under the j-th superclass (a total of K subclasses). By recursively merging upwards layer by layer, a multi-level representation can be obtained. The specific construction process of the hierarchical tree is shown in Algorithm 1.

Algorithm 1 Process of building a hierarchical tree

Input: Similarity matrix $S \in R^{N_0 \times N_0}$, Semantic embedding matrix $E \in R^{N_0 \times d}$, Tree height $H \geq 1$
Output: Hierarchy $T = \{E(0), E(1), ..., E(H-1)\}$, where $E(0) = E$
Initialize: $T \leftarrow E$, $S(0) \leftarrow S$, $E(0) \leftarrow E$
for L **to** $H - 1$ **do**
  N $\leftarrow$ number of rows in $S(L-1)$ ▷ Determine the number of nodes that can be merged
  if $N < 3$ **then**
    break
  end if
  for $i = 0$ **to** $N - 1$ **do**
    r $\leftarrow S(L-1)[i, :]$ ▷ Fetch i-th row
    $j_1, j_2, j_3 \leftarrow$indices of top-3 largest values in r
    $v_1 \leftarrow E(L-1)[j_1, :]$
    $v_2 \leftarrow E(L-1)[j_2, :]$
    $v_3 \leftarrow E(L-1)[j_3, :]$
    $E(L)[i, :] \leftarrow v_1 + v_2 + v_3$ ▷ Fuse top-3 semantics
  end for
  $S(L) \leftarrow E(L) \times (E(L))^T$ ▷ Compute next-level similarity via matmul product
  Append $E(L)$ to T
end for
return T

To facilitate subsequent merging by mapping each category to its corresponding superclass label, this section introduces two concepts: a label mapping table and an embedding mapping table. The first is a label mapping table consisting of hierarchical label, [category label, superclass label, similarity], and the second is an embedding mapping table consisting of hierarchical label, [superclass label, superclass embedding]. During each training iteration, each category label and its corresponding superclass

label are first obtained and stored in the first label mapping table. Then, the corresponding superclass label embedding is calculated and stored in the second embedding mapping table.

Step 3: Supervised training of hierarchical classification. In Step 2, a multi-level classification tree was constructed. To utilize superclass labels to guide model training and improve the model's ability to perceive multiple levels, this section proposes a hierarchical classification head for superclasses to supervise training each embedding and optimize the model using the corresponding superclass labels. For example, a hierarchical tree has L levels, each with a hierarchical classification head. There are L hierarchical classification heads that need to output classification probabilities, consistent with regular classification heads. The probability output by each hierarchical classification head can be represented as the probability of the superclass label. The loss function used is the multivariate cross-entropy loss $\mathcal{L}_{cls}^{m}$. Since the number of samples in each level is not uniform, this can lead to an imbalance. Therefore, the total classification loss needs to be weighted by the loss of each level. The final loss can be expressed as follows:

$$\mathcal{L}_{cls}^{L} = \sum_{m=1}^{L} w_m \mathcal{L}_{cls}^{m}. \tag{3}$$

3.4 Fine-grained semantic prototype encoding module

As shown in the lower left of Figure 1, fine-grained semantic prototype encoding module is specifically designed for extracting fine-grained semantic prototypes. It mainly consists of three parts: a contrastive description generator, a text encoder, and a semantic prototype bank. The contrastive description generator is obtained from a pre-trained CoCa model[41]. The main function of the fine-grained semantic prototype generator is to generate fine-grained visual descriptions for each training image, semantically encode these descriptions, and finally store them in the semantic prototype bank. Its greatest advantage lies in its ability to generate personalized descriptions for each image, which significantly improves the generalization ability of semantic prototypes.

The semantic description generation process is as follows: First, the image is input to the proposed multimodal visual encoder. Then, the feature map output by the visual encoder is input to the multimodal text encoder. The text encoder outputs a fine-grained image description (e.g., given an image, the inference output is: "A bird with red feathers is standing on a branch"). This text is then compressed into an embedding representation to characterize the category prototype. Since multiple identical categories exist in the training set, all obtained semantic prototypes are average-pooled to obtain the final category prototype. It's important to note that the entire fine-grained semantic prototype encoding module runs before training and does not participate in model training. The parameters of the CoCa model and text encoder are completely frozen; the obtained category prototypes are used for subsequent base classes and fine-tuning training, but their content is not updated. Compared to some methods that directly generate class name descriptions or fine-grained descriptions of categories, the fine-grained semantic prototype encoding takes into account the

differences between semantics and vision, generating customized text descriptions for each image, thus obtaining more generalizable semantic information.

4 Experiments

To verify the effectiveness of the proposed method, experiments were conducted on two benchmark datasets, PASCAL VOC and MS-COCO, and ablation and visualization experiments were designed to validate the effectiveness of the proposed module. Furthermore, to maintain a fair comparison, this paper maintains the same novel class partitioning strategy as methods such as TFA[6] and compares it with state-of-the-art few-shot object detection methods. Since the proposed method is a generalized few-shot object detection method, this paper qualitatively and quantitatively analyzes the experimental results on both base class and novel class metrics.

4.1 Datasets and settings

Datasets and evaluation metrics: This paper trains and evaluates the model on two benchmark datasets, PASCAL VOC and MS-COCO. For the PASCAL VOC dataset, 5 classes are randomly selected from 20 categories as novel classes, and the remaining 15 categories are used as base classes. For the MS-COCO dataset, 20 categories identical to those in PASCAL VOC are randomly selected from 80 categories as novel classes, and 60 categories are used as base classes. Three different partitioning methods are used for both datasets, resulting in datasets containing samples of the same category but with different content. The experimental settings for sample size are consistent with the benchmark methods: for the PASCAL VOC dataset, the number of novel class samples $K \in \{1, 2, 3, 5, 10\}$ is selected; for the MS-COCO dataset, the number of novel class samples $K \in \{1, 2, 3, 5, 10, 30\}$ is selected. Because this paper focuses on the model's ability to identify both base and novel classes, nAP, nAPm, and bAP are selected as evaluation metrics for the two benchmark datasets.

Implementation details: The proposed method employs a two-stage fine-tuning training approach based on the DeFRCN framework, utilizing a ResNet-101 pre-trained on ImageNet as the backbone network, along with a GDL gradient decoupling module. All model training is performed on two RTX 4090 GPUs. During training, SGD is used for optimization, with a batch size of 16, momentum of 0.9, and weight decay of 5e-5. The learning rate is set to 0.02 during training and 0.01 during fine-tuning, decaying in the final 10% of the training.

4.2 Comparison methods

This paper selects classic and state-of-the-art methods from meta-learning and fine-tuning approaches as comparative methods, as shown in Table 1. The meta-learning methods include Meta-RCNN[11], FSRW[12], Meta-DETR[13], VFA[14], FM-FSOD[15], ICPE[16], MM-FSOD[17], and FPD[18]; the fine-tuning methods include TFA[6], DeFRCN[10], MFDC[22], FCT[42], PTF+KI[19], D&R[8], Norm-VAE[43], BSDet[23], SINDA[28], FR[21], MMKT[25], T-GSEL[44], MPAD[45], OPN[46], GSCL[47] and CELD[29]. This paper selects DeFRCN as baseline models, therefore

Table 1 Comparison methods and details

Method	Class	Periodical/Conference	Year	Single/Multiple runs
Meta-RCNN		ICCV	2019	Single run
FSRW		ICCV	2020	Single run
Meta-DETR		IEEE TIP	2022	Both
MM-FSOD		-	2022	Both
VFA	Meta-learning	AAAI	2023	Both
FM-FSOD		CVIU	2023	Both
ICPE		AAAI	2023	Single run
FPD		AAAI	2024	Both
TFA		ICML	2020	Both
DeFRCN		ICCV	2021	Both
MFDC		ECCV	2022	Both
FCT		CVPR	2022	Both
PTF+KI		IEEE TIP	2023	Single run
D&R		AAAI	2023	Both
Norm-VAE		CVPR	2023	Both
BSDet	Fine-tuning	IEEE TMM	2024	Single run
SINDA		CVPR	2024	Single run
FR		Neural Networks	2025	Single run
MMKT		Pattern Recognition	2025	Single run
T-GSEL		IJCV	2025	Single run
MPAD		ICLR	2025	Single run
OPN		ESWA	2025	Single run
GSCL		IEEE TPAMI	2025	Single run
CELD		Display	2026	Both

the experimental environment and parameter settings are consistent with the baseline models.

4.3 Comparison results with advanced methods

This paper compares the proposed method with current benchmark methods and state-of-the-art methods on two datasets: PASCAL VOC and MS-COCO. It also presents the model's metrics and comparative results on both datasets for single-round and multi-round training scenarios. Under most few-shot settings, the proposed method outperforms all current state-of-the-art methods.

4.3.1 Experimental results on PASCAL VOC

As shown in Tables 2 and 3, these tables present the comparative results of all methods on the PASCAL VOC dataset under single-round and multi-round training scenarios, covering test results for 1, 2, 3, 5, and 10 shots. Specifically, for the novel class, we used the mean accuracy (nAP50) over 10 rounds of testing when the IOU threshold is greater than 0.5 as the performance metric. It is worth noting that most comparative methods only provide results from a single round of experiments, leading to inconsistencies in the comparison methods. The data in the tables demonstrate that the proposed method significantly outperforms recent state-of-the-art methods, thereby quantitatively validating its effectiveness.

As shown in Table 2, the single-round test results on the PASCAL VOC dataset demonstrate that the proposed method achieves state-of-the-art performance across all three novel splits under various shot settings. Compared to existing fine-tuning and

meta-learning baselines, our approach consistently yields the highest average nAP50 of 60.3, surpassing the previous best method (MFDC, 59.0) by 1.3%. Notably, under Novel Split 1, our method sets new records in the 1-shot (64.5), 2-shot (66.4), and 3-shot (68.1) settings—outperforming the strongest competitor (CELD) by up to 0.7–2.2%. Additionally, in the 5-shot and 10-shot scenarios of Novel Split 2, our results (55.1 and 60.2) exceed all existing methods, with the 10-shot result marking the first time any approach has broken the 60.0 nAP50 barrier on this split. These consistent improvements across data partitions and sample regimes validate the effectiveness of our hierarchical semantic transfer and fine-grained semantic prototype encoding strategy in enhancing few-shot generalization.

Table 2 Single-round test results on the novel class in the PASCAL VOC dataset (nAP50), where bold denotes the highest score and underlined text indicates the second-highest performance.

Type	Method/Shot	Year	Novel Split1					Novel Split2					Novel Split3					AVG
			1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
Meta-learning	FSRW	ICCV'19	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9	29.0
	MetaR-CNN	ICCV'19	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1	31.3
	Meta-DETR	TPAMI'22	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6	50.4
	MM-FSOD	/	42.5	41.2	41.6	48.0	53.4	30.5	34.0	39.3	36.8	37.6	39.9	37.0	38.2	42.5	45.6	40.5
	VFA	AAAI'23	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9	56.7
	FM-FSOD	CVII'23	47.5	50.6	53.4	57.2	59.9	37.9	39.8	43.6	46.7	49.8	33.4	36.3	40.5	46.8	48.9	46.0
	ICPE	AAAI'23	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	50.9	53.1	55.3	60.6	60.1	54.1
	FPD	AAAI'24	48.1	62.2	64.0	67.6	**68.4**	29.8	43.2	47.7	52.0	53.9	44.9	53.8	58.1	61.6	62.9	54.6
Fine-Tuning	TFA/w/fc	ICML'20	36.8	29.1	43.6	55.7	57.0	18.2	29.0	33.4	35.5	39.0	27.7	33.6	42.5	48.7	50.2	39.4
	TFA/w/cos	ICML'20	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8	39.9
	DeFRCN	ICCV'21	56.3	62.1	63.2	67.8	65.8	34.9	35.7	48.2	54.2	53.3	50.6	57.1	55.4	59.8	63.0	55.4
	MFDC	ECCV'22	63.4	66.3	67.7	**69.4**	68.1	42.1	46.5	53.4	**55.3**	53.8	56.1	58.3	59.0	62.2	63.7	59.0
	FCT	CVPR'22	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7	50.3
	PTF+KI	TIP'23	57.0	62.3	63.3	66.2	67.6	42.8	44.9	50.5	52.3	52.2	50.8	56.9	58.5	62.1	63.1	56.5
	D&R	AAAI'23	60.4	64.0	65.2	64.7	66.3	47.9	46.8	48.1	52.7	53.1	55.7	57.9	57.6	60.6	61.9	55.9
	Norm-VAE	CVPR'23	62.1	64.9	67.8	69.2	67.5	39.9	46.8	54.4	54.2	53.6	**58.2**	60.3	61.0	**64.0**	**65.5**	58.9
	BSDet	TMM'24	58.6	60.4	62.5	67.7	67.6	42.2	**48.2**	50.6	52.3	55.6	49.5	59.3	61.9	63.6	62.5	57.3
	SINDA	CVPR'24	59.3	60.8	64.3	65.4	65.6	35.2	40.8	50.2	54.6	50.0	51.6	52.4	55.9	58.5	62.6	54.8
	FR	NN'25	61.2	65.8	67.1	67.6	67.9	36.4	46.5	49.5	53.8	54.2	53.5	56.4	57.2	60.3	62.1	56.8
	MMKT	PR'25	42.9	55.5	60.8	66.6	68.1	31.1	41.7	43.8	51.3	54.1	39.1	48.9	55.4	61.1	63.8	52.0
	T-GSEL	IJCV'25	50.4	63.6	61.9	68.6	67.3	31.3	32.9	43.6	47.9	53.9	41.2	49.8	54.1	62.1	61.9	52.6
	OPN	ESWA'25	53.7	58.8	60.2	65.0	66.9	38.7	41.8	46.4	50.1	53.6	47.1	53.3	56.1	60.0	61.6	53.9
	GSCL	TPAMI'25	58.6	61.9	65.2	68.8	67.7	38.8	46.7	52.8	55.1	54.1	56.5	58.1	59.6	61.0	63.1	57.9
	CELD	Display'26	63.8	63.3	66.0	68.3	67.1	**45.2**	46.6	52.7	54.1	52.0	57.8	60.1	59.9	62.2	64.1	58.9
	Ours	/	**64.5**	**66.4**	**68.1**	68.7	67.7	42.5	47.9	**54.6**	55.1	**60.2**	55.5	**61.8**	**64.0**	63.5	63.4	**60.3**

To mitigate the randomness of single-round evaluation, we conduct multi-round testing on PASCAL-VOC, including representative baselines that report such results. As shown in Table 3, our method achieves state-of-the-art performance across most settings. Notably, under Novel Split 1, we obtain 58.5, 57.2, 61.8, 64.3, and 68.7 nAP50 for 1–10 shots, outperforming the previous best (CELD) in 1-shot (+1.9%) and 10-shot (+0.1%), and significantly surpassing TFA by up to 23.2% in 1-shot. In the more challenging Novel Split 2, our method sets new records in 1-shot (45.0 vs. 42.4) and 5-shot (52.7 vs. 51.4). In Novel Split 3, we achieve 65.8 (5-shot) and 64.6 (10-shot), exceeding CELD by up to 3.2%. These results confirm the robustness and effectiveness of our approach under rigorous multi-trial evaluation.

Experimental results show that the proposed hierarchical semantic transfer and fine-grained semantic prototype encoding module modules effectively enhance knowledge transfer in few-shot detection. The hierarchical tree structure establishes multi-level semantic links between novel and base classes, enabling structured knowledge propagation across granularities. Meanwhile, the fine-grained semantic prototype

Table 3 Average results of multiple rounds of testing on the novel class in the PASCAL-VOC dataset (nAP50), where bold denotes the highest score and underlined text indicates the second-highest performance.

Type	Method/shot	Year	Novel Split1					Novel Split2					Novel Split3				
			1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
Meta-learning	Meta-DETR	TPAMI'22	35.1	49.0	53.2	57.4	62.0	27.9	32.3	38.4	43.2	51.8	34.9	41.8	47.1	54.1	58.2
	VFA	AAAI'23	47.4	54.4	58.5	64.5	66.5	33.7	38.2	43.5	48.3	52.4	43.8	48.9	53.3	58.1	60.0
	FPD	AAAI'24	41.5	52.8	58.4	64.9	67.1	28.2	38.7	43.8	50.3	53.6	34.9	48.6	54.0	58.4	61.5
Fine-Tuning	TFA/w/fc	ICML'20	22.9	34.5	40.4	46.7	52.0	16.9	26.4	30.5	34.6	39.7	15.7	27.2	34.7	40.8	44.6
	TFA/w/cos	ICML'20	25.3	36.4	42.1	47.9	52.8	18.3	27.5	30.9	34.1	39.5	17.9	27.2	34.3	40.8	45.6
	DeFRCN	ICCV'21	40.2	53.6	58.2	63.6	66.5	29.5	39.7	43.4	48.1	52.8	35.0	38.3	52.9	57.7	60.8
	D&R	AAAI'23	41.0	51.7	55.7	61.8	65.4	30.7	39.0	42.5	46.6	51.7	37.9	47.1	51.7	56.8	59.5
	MFDC	ECCV'22	45.4	56.6	59.2	65.1	67.3	30.3	39.9	44.1	49.2	53.5	36.6	40.2	55.8	59.0	61.4
	CELD	Display'26	56.6	63.6	65.4	67.5	68.6	42.4	46.3	49.1	51.4	53.6	53.4	58.4	59.8	62.6	63.7
	Ours	/	**58.5**	**57.2**	**61.8**	64.3	**68.7**	**45.0**	**41.1**	**46.0**	**52.7**	**53.8**	38.5	47.7	**60.4**	**65.8**	**64.6**

encoding module generates image-specific textual descriptions and stores their embeddings in a prototype bank, providing discriminative semantic guidance that captures intra-class variation. Together, they replace coarse class-level semantics with a richer, alignment-aware representation, significantly boosting detection accuracy for novel classes.

Table 4 Test results on the base class (bAP50) in the PASCAL-VOC dataset, where bold represents the best result.

Method	**Novel-Split1**					**Novel-Split2**					**Novel-Split3**				
	1	**2**	**3**	**5**	**10**	**1**	**2**	**3**	**5**	**10**	**1**	**2**	**3**	**5**	**10**
FSRW	66.4	61.7	64.8	63.4	63.6	68.2	67.2	66.0	66.6	64.7	65.9	65.3	65.0	65.7	63.1
TFA/w/fc	77.1	75.8	76.0	75.1	75.4	73.8	74.9	75.4	75.3	76.1	77.8	77.7	75.6	75.8	76.2
TFA/w/cos	77.6	77.3	77.3	77.4	77.5	73.8	74.9	75.6	76.2	76.9	**78.7**	78.4	78.6	78.5	78.6
DeFRCN	78.2	77.3	77.4	78.3	78.3	79.5	78.6	77.4	78.3	78.0	**78.7**	78.2	78.6	78.4	78.1
MFDC	78.5	77.1	77.6	78.5	78.6	79.8	78.2	77.6	78.5	78.1	**78.7**	78.5	78.8	78.9	78.3
BSDet	77.5	77.9	75.1	77.2	77.5	78.5	78.6	76.4	78.3	78.4	77.5	77.7	76.3	77.5	78.1
Ours	**79.4**	**79.7**	**78.9**	**79.0**	**79.6**	**80.5**	**79.9**	**81.2**	**79.5**	**80.0**	78.5	**80.3**	**79.0**	**80.5**	**79.2**

Given that the proposed method belongs to a generalized few-shot object detection framework, experiments were conducted based on the benchmark model DeFRCN, with extensive evaluation on the base class detection performance. Five representative methods were rigorously selected for horizontal comparison. As shown in Table 4, compared with the benchmark method DeFRCN, the proposed method achieves consistent improvements in base class detection accuracy across all three novel splits and shot settings. The best result reaches 81.2 under Novel-Split2 (3-shot), representing a significant gain of 3.2% over DeFRCN (78.0). Moreover, the performance on base classes remains highly stable—deviating by less than 0.25% from the pre-fine-tuning metrics—demonstrating that our approach effectively mitigates catastrophic forgetting of base categories during novel class adaptation.

4.3.2 Experimental results on MS-COCO

The MS-COCO dataset, with its complex class system and high-frequency small-scale targets, forms a highly challenging benchmark scenario for few-shot object detection. The performance of our approach is evaluated against several leading benchmarks on the MS-COCO dataset, as detailed in Table 5. This comprehensive assessment reports

the detection accuracy for both base (bAP) and novel (nAP) categories across multiple experimental iterations.

Table 5 Novel class of metrics(nAP) on the MS-COCO dataset, where bold denotes the highest score and underlined text indicates the second-highest performance.

Method/shot	MS-COCO(nAP)					
	1	2	3	5	10	30
FSRW	-	-	-	-	5.6	9.1
MetaR-CNN	-	-	-	-	8.7	12.4
TFAw/fc	-	-	-	-	10.0	13.4
TFAw/cos	4.4	5.4	6.0	7.7	10.0	13.7
DeFRCN*	4.8	8.5	10.7	13.6	16.8	21.2
Meta-DETR	7.5	-	13.5	15.4	19.0	22.2
MM-FSOD	5.6	7.9	9.4	10.5	10.8	12.3
VFA	-	-	-	-	15.9	18.4
FCT	5.6	7.9	11.1	14.0	17.1	21.4
ICPE	-	-	-	-	19.3	23.1
MFDC	10.8	13.9	15.0	16.4	19.4	22.7
PTF-KI	-	-	-	-	13.0	16.8
D&R	6.1	9.5	11.5	13.9	16.4	20.0
Norm-VAE	9.5	13.7	14.3	15.9	18.7	22.5
FPD	-	-	-	-	15.9	19.3
BSDet	8.4	11.1	13.5	14.4	17.2	21.2
SINDA	10.2	14.5	15.8	16.9	19.1	23.1
FR	7.9	12.5	13.8	15.5	18.4	22.3
MMKT	-	-	-	-	19.2	**23.6**
T-GSEL	-	-	-	-	18.0	22.3
CELD	**10.9**	14.0	15.7	17.7	**19.6**	22.5
Ours	9.9	**15.3**	**16.8**	**18.5**	17.2	23.5

Table 6 Base class metrics on the MS-COCO dataset, where bold denotes the highest score and underlined text indicates the second-highest performance.

Method/shot	MS-COCO(bAPm)					
	1	2	3	5	10	30
TFAw/fc	31.5	31.4	31.5	31.8	32.0	33.8
TFAw/cos	31.9	31.9	32.0	32.3	32.4	34.2
DeFRCN	36.6	37.2	37.6	38.1	40.0	39.9
MFDC	36.9	**38.5**	37.3	37.6	40.5	39.8
BSDet	35.5	36.8	36.6	37.9	38.2	38.6
Ours	**37.2**	37.7	**38.3**	**38.4**	**41.1**	**40.2**

The table shows that the proposed method achieves the best performance in most sample size settings on the MS-COCO dataset. Specifically, our approach obtains the highest nAP scores of 15.3 (2-shot), 16.8 (3-shot), and 18.5 (5-shot), outperforming the previous state-of-the-art method CELD by 0.8, 1.1, and 0.8 points, respectively. Although MMKT achieves the top result at 30-shot (23.6), our method ranks second

with 23.5, demonstrating competitive scalability. Notably, compared to strong baselines such as DeFRCN and SINDA, the proposed method achieves a maximum relative improvement of approximately 12.5% in the novel class metric (e.g., 18.5 vs. 16.4 for DeFRCN at 5-shot). These results validate the efficacy and stability of the proposed framework across varying shot settings on a large-scale and challenging dataset.

4.4 Visualization results

This section presents the overall qualitative results of the proposed method in detecting novel target categories, visualized in Figure 3. As shown in the figure, the baseline method yields significant false positives, particularly in cases of category confusion (e.g., bus and car). In contrast, the proposed method, having learned the relationships between categories, significantly reduces both false positives and false negatives. For example, in the image of "sheep," the proposed method improves category confidence by establishing category relationships, successfully identifying the sheep in the upper right corner and effectively enhancing the model's detection accuracy.

Furthermore, the proposed method enhances the model's generalization ability through fine-grained descriptive information, enabling accurate detection even when faced with the same target from different perspectives and with varying features (e.g., the "person" in the image varies considerably across different images and from different perspectives). These qualitative results clearly demonstrate that the proposed method effectively learns the relationships between categories and significantly improves the model's generalization ability through fine-grained descriptions.

4.5 Ablation experiment

To further verify the reliability of the proposed hierarchical semantic transfer module and fine-grained semantic prototype encoding module in improving detection performance, this section designs three sets of ablation experiments and validates them on the PASCAL VOC dataset. These three sets of experiments include a baseline model, a baseline model with the hierarchical semantic transfer module added, and a baseline model with both the hierarchical semantic transfer module and the fine-grained semantic prototype encoding module added.

Table 7 Comparison of detection accuracy of models with/without hierarchical transfer and fine-grained semantic prototype encoding modules on the PASCAL VOC dataset (the best result is in bold).

Model	**Novel Split1**				
	1	**2**	**3**	**5**	**10**
Baseline	56.3	62.1	63.2	67.8	65.8
Baseline w/ HL	63.6	66.0	66.9	67.7	66.3
Baseline w/ HL&FGP(ours)	**64.5**	**66.4**	**68.0**	**68.7**	**67.7**

Notably, since the similarity matrix generated by the fine-grained semantic encoding module is used for superclass learning (i.e., the two modules work collaboratively), in the second set of experiments, the similarity matrix used for superclass learning



（a）Detection results of the baseline model (DeFRCN)



（b）Detection results of our method

Fig. 3 Comparison of qualitative results of the baseline method(DeFRCN) and the proposed method on the VOC dataset.

is calculated from the proposed region features (saving the proposed region features for each category). The experimental results show that the proposed method's metrics are improved to some extent after introducing the hierarchical semantic transfer module, indicating that this module can effectively mine the relationship between the base class and the novel class, thereby improving the recognition performance of the novel class.

Furthermore, after introducing the fine-grained semantic prototype encoding module, the generated similarity matrix becomes more accurate, enabling the joint mining of inter-class relationships with greater precision using the hierarchical tree, thereby significantly improving the overall recognition accuracy.

4.6 Feature visualization experiment

Figure 4 shows the visualization results of the model's features on the PASCAL VOC dataset. These features are composed of ROI features output by the region generation

network from the selected test set. Since these features are later used for classification tasks, they can clearly represent the coupling relationship between classes. The first row in the figure represents the detection results of the baseline model, and the second row represents the detection results of the proposed method. As shown in the leftmost column of the comparison chart, the figure shows a visualization comparison of the features of the baseline model and the proposed method after fine-tuning. Compared with the baseline model, the features output by the proposed method are more tightly packed within classes and have a higher degree of separation between classes. This suggests that the proposed method can better establish the discrimination space of base class features and mitigate the base class forgetting phenomenon caused by fine-tuning to some extent.

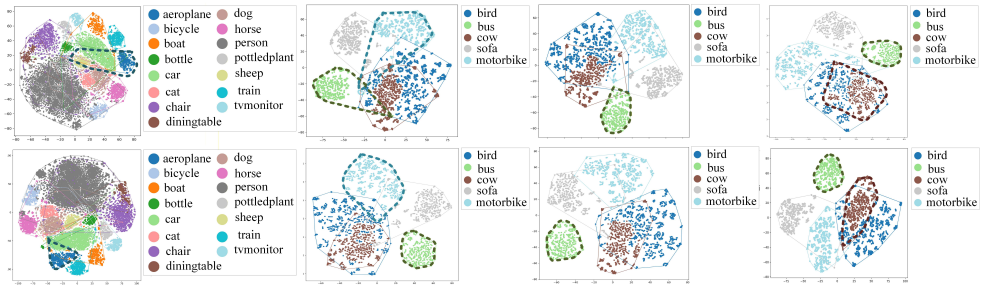

Fig. 4 Visualization of T-SNE features. The first row and second row of the figure represent the feature visualization results of the baseline model and the proposed method, respectively. The first column shows the visualization of 15 base class features, while the second to fourth columns show the visualization of 5 novel class features under 1-shot, 5-shot, and 10-shot sample sizes, respectively.

Columns 2 to 4 of the figure show a visual comparison of the novel class features of the baseline model and the proposed method under sample sizes of 1-shot, 5-shot, and 10-shot. As can be seen from the figure, the proposed method can significantly alleviate inter-class confusion of novel classes, especially for "cow" and "bird" in the third column. The features of the baseline model cause severe overlap between these two classes, and the inter-class variance is large, which will adversely affect subsequent classification. In contrast, the features inferred by the proposed method are more compact within classes and have higher inter-class discriminative power. This indicates that after introducing the two modules, the model effectively learns the hierarchical relationships between classes and the key semantic information for distinguishing classes, thereby improving the accuracy of novel class recognition.

5 Conclusion

This paper proposes a few-shot object detection (FSOD) framework that integrates base-novel hierarchical semantic transfer with fine-grained semantic prototype encoding. The hierarchical strategy employs a visual-text similarity matrix to recursively

cluster categories into a multi-layered superclass tree, effectively facilitating knowledge migration by modeling shared semantic granularities. Simultaneously, a fine-grained encoding module—consisting of a contrastive description generator and a semantic prototype bank—is introduced to generate discriminative, image-specific descriptors, thereby overcoming the rigidity of static textual priors. Experimental results on PASCAL VOC and MS-COCO demonstrate that the method significantly enhances detection accuracy for both base and novel classes. However, since the hierarchical tree is constructed offline based on pre-calculated similarities, it lacks the flexibility to dynamically adapt category relationships during the fine-tuning phase. Such a static hierarchy may introduce erroneous inductive biases when encountering semantic ambiguity or significant base-novel semantic shifts; therefore, future work will focus on developing dynamic relational modeling to improve the model's robustness against domain discrepancies.

Acknowledgements. The authors would like to thank their colleagues from the Machine learning group for discussions on this paper. This work was supported by the Super Computing Platform of Beijing University of Posts and Telecommunications.

Declarations

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.

Funding Declaration

No funding was received.

References

[1] Wang, S., Xia, C., Lv, F., Shi, Y.: Rt-detrv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1628–1636 (2025). IEEE

[2] Yang, L., Jiang, H., Song, Q., Guo, J.: A survey on long-tailed visual recognition. International Journal of Computer Vision **130**(7), 1837–1872 (2022)

[3] Tran, P.V.: SimLTD: Simple Supervised and Semi-Supervised Long-Tailed Object Detection (2025)

[4] Xin, Z., Chen, S., Wu, T., Shao, Y., Ding, W., You, X.: Few-shot object detection: Research advances and challenges. Information Fusion **107**, 102307 (2024)

[5] Köhler, M., Eisenbach, M., Gross, H.-M.: Few-shot object detection: A comprehensive survey. IEEE Transactions on Neural Networks and Learning Systems **35**(9), 11958–11978 (2023)

[6] Wang, X., Huang, T.E., Darrell, T., Gonzalez, J.E., Yu, F.: Frustratingly Simple Few-Shot Object Detection (2020)

[7] Pei, W., Wu, S., Mei, D., Chen, F., Tian, J., Lu, G.: Few-shot object detection by knowledge distillation using bag-of-visual-words representations. In: European Conference on Computer Vision, pp. 283–299 (2022). Springer

[8] Li, J., Zhang, Y., Qiang, W., Si, L., Jiao, C., Hu, X., Zheng, C., Sun, F.: Disentangle and remerge: interventional knowledge distillation for few-shot object detection from a conditional causal perspective. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1323–1333 (2023)

[9] 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)

[10] Qiao, L., Zhao, Y., Li, Z., *et al.*: Defrcn: Decoupled faster r-cnn for few-shot object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8681–8690 (2021)

[11] 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)

[12] 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)

[13] 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)

[14] Han, J., Ren, Y., Ding, J., Yan, K., Xia, G.-S.: Few-shot object detection via variational feature aggregation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 755–763 (2023)

[15] Li, Y., Feng, W., Lyu, S., Zhao, Q.: Feature reconstruction and metric based network for few-shot object detection. Computer Vision and Image Understanding **227**, 103600 (2023)

[16] Lu, X., Diao, W., Mao, Y., Li, J., Wang, P., Sun, X., Fu, K.: Breaking immutable: Information-coupled prototype elaboration for few-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1844–1852 (2023)

[17] Han, G., Chen, L., Ma, J., Huang, S., Chellappa, R., Chang, S.-F.: Multi-modal few-shot object detection with meta-learning-based cross-modal prompting. arXiv

preprint arXiv:2204.07841 (2022)

[18] Wang, Z., Yang, B., Yue, H., Ma, Z.: Fine-grained prototypes distillation for few-shot object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 5859–5866 (2024)

[19] Yang, Z., Zhang, C., Li, R., Xu, Y., Lin, G.: Efficient few-shot object detection via knowledge inheritance. IEEE Transactions on Image Processing **32**, 321–334 (2022)

[20] Zhang, X., Chen, Z., Zhang, J., *et al.*: Learning general and specific embedding with transformer for few-shot object detection. International Journal of Computer Vision **133**(2), 968–984 (2025)

[21] Zhang, Y., Li, J., Ji, Q., Li, K., Liu, L., Zheng, C., Qiang, W.: Intervening on few-shot object detection based on the front-door criterion. Neural Networks **185**, 107251 (2025)

[22] Wu, S., Pei, W., Mei, D., Chen, F., Tian, J., Lu, G.: Multi-faceted distillation of base-novel commonality for few-shot object detection. In: European Conference on Computer Vision, pp. 578–594 (2022). Springer

[23] Lu, Y., Chen, X., Wu, Z., Tan, M., Yu, J.: Binary similarity few-shot object detection with modeling of hard negative samples. IEEE Transactions on Multimedia **26**, 4805–4818 (2023)

[24] Jonschkowski, R., Höfer, S., Brock, O.: Patterns for learning with side information. arXiv preprint arXiv:1511.06429 (2015)

[25] Du, Y., Liu, F., Jiao, L., Li, S., Hao, Z., Li, P., Wang, J., Wang, H., Liu, X.: Text generation and multi-modal knowledge transfer for few-shot object detection. Pattern Recognition **161**, 111283 (2025)

[26] Guo, Y., Ma, L., Luo, X., Xie, S.: Dp-ddcl: A discriminative prototype with dual decoupled contrast learning method for few-shot object detection. Knowledge-Based Systems **297**, 111964 (2024)

[27] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). Proceedings of Machine Learning Research

[28] Wang, Y., Zou, X., Yan, L., Zhong, S., Zhou, J.: Snida: Unlocking few-shot object detection with non-linear semantic decoupling augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12544–12553 (2024)

[29] Zhao, T., Qiu, H., Wang, L., Dai, Y., Wu, Q., Li, H.: Class extension logits distillation for few-shot object detection. Displays, 103338 (2026)

[30] Griffin, G., Perona, P.: Learning and using taxonomies for fast visual categorization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE, Anchorage, AK, USA (2008)

[31] Liu, B., Sadeghi, F., Tappen, M., Porikli, F.: Probabilistic label trees for efficient large scale image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 843–850. IEEE, Portland, OR, USA (2013)

[32] Freeman, C., Kulić, D., Basir, O.: Feature-selected tree-based classification. IEEE Transactions on Cybernetics **43**(6), 1990–2004 (2013)

[33] Zhao, H., Zhu, P., Wang, P., Ye, J.: Hierarchical feature selection with recursive regularization. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 3483–3489. IJCAI Organization, Melbourne, Australia (2017)

[34] Xie, S., Yang, T., Wang, X., Lin, Y.: Hyper-class augmented and regularized deep learning for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2645–2654 (2015)

[35] Guillaumin, M., Ferrari, V.: Large-scale knowledge transfer for object localization in imagenet. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3202–3209. IEEE, Providence, RI, USA (2012)

[36] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., al.: Devise: A deep visual-semantic embedding model. In: Advances in Neural Information Processing Systems, vol. 26, pp. 2121–2129. Curran Associates, Inc., Red Hook, NY, USA (2013)

[37] Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 638–647. IEEE, Waikoloa, HI, USA (2019)

[38] Bertinetto, L., Mueller, R., Tertikas, K., Vedaldi, A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 12506–12515. IEEE, Seattle, WA, USA (2020)

[39] Wu, H., Merler, M., Uceda-Sosa, R., Oh, S.J., Perona, P.: Learning to make better mistakes: Semantics-aware visual food recognition. In: Proceedings of the 24th ACM International Conference on Multimedia, pp. 172–176. ACM, Amsterdam, Netherlands (2016)

[40] Zhang, L., Wang, Y., Zhou, J., Li, H., Li, X.: Hierarchical few-shot object detection: Problem, benchmark and method. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2002–2011. ACM, Lisbon, Portugal (2022)

[41] Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: Contrastive Captioners are Image-Text Foundation Models (2022)

[42] Han, G., Ma, J., Huang, S., Chen, L., Chang, S.-F.: Few-shot object detection with fully cross-transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5321–5330 (2022)

[43] Xu, J., Le, H., Samaras, D.: Generating features with increased crop-related diversity for few-shot object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 19713–19722 (2023)

[44] Zhang, X., Chen, Z., Zhang, J., *et al.*: Learning general and specific embedding with transformer for few-shot object detection. International Journal of Computer Vision **133**(2), 968–984 (2025)

[45] Vu, A.-K.N., Truong, Q.-T., Nguyen, V.-T., Ngo, T.D., Do, T.-T., Nguyen, T.V.: Multi-Perspective Data Augmentation for Few-shot Object Detection (2025)

[46] Wang, B., Yu, D.: Orthogonal progressive network for few-shot object detection. Expert Systems with Applications **264**, 125905 (2025)

[47] Chen, R., Zhang, H., Li, J., Liu, L., Huang, Z., Cao, X.: Generalized semantic contrastive learning via embedding side information for few-shot object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(8), 6496–6514 (2025)