Towards Generalizable Few-Shot Object Detection via Enhanced Representation Learning

by

Yan Zhang

A thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the degree of
Master of Computer Science and Concentration Applied Artificial Intelligence

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work except where otherwise indicated. I am aware of the University of Ottawa regulations concerning plagiarism, including those regarding consequent disciplinary actions. Any use of the works of any other author, in any form, is properly acknowledged at their point of use.

Abstract

Few-shot object detection (FSOD), which aims to detect novel categories with minimal training examples, faces significant challenges in learning robust feature representations due to severe data scarcity. Additionally, FSOD models often struggle to distinguish objects from visually ambiguous backgrounds, restricting their generalization capability. We propose a novel FSOD framework designed to address these challenges through two key innovations. First, we introduce Wavelet-Semantic Fusion Attention (WSFA), which enhances semantic ViT features by incorporating frequency-domain information via discrete wavelet transform, providing complementary edge and texture cues through cross-modal attention. Second, we propose the Learnable Background Prototype (LBP) that explicitly models the background patterns, significantly improving foreground-background discrimination. These contributions are then integrated into a unified single-stage transformer-based detection framework with inter-class contrastive learning. Comprehensive experiments on standard FSOD benchmarks (PASCAL VOC and MS COCO) demonstrate that our method achieves stable improvements over strong baseline methods and outperforms existing state-of-the-art approaches. This work provides a practical solution for scenarios with limited annotated data, enhancing the applicability of object detection in real-world applications.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Robert Laganière, for his invaluable guidance, continuous support, and inspiring mentorship throughout this research journey. His patience, wisdom, and constructive feedback have not only significantly improved the quality of this work but also strengthened my analytical and critical thinking abilities. His belief in my potential motivated me to persevere through challenges and strive for excellence. I am deeply grateful for his lasting influence on both my academic development and personal growth.

I am also profoundly thankful to my mother for her endless love, understanding, and unwavering emotional support, which gave me the strength and encouragement to pursue my studies. My heartfelt thanks go to my friend Xuanqi Zhang for his companionship, valuable discussions, and constant encouragement, which made this academic journey more meaningful and enjoyable.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Few-shot object detection (FSOD) has emerged as a critical challenge in computer vision, primarily because traditional object detection models require extensive annotated data for each category [1]. Collecting and annotating thousands of images for each category is often impractical and costly, especially in specialized domains like medical imaging, robotics, and surveillance [2, 3, 4, 5, 6], where novel instances frequently arise and annotations are scarce. FSOD addresses this challenge by leveraging prior knowledge acquired from abundant base-class data to rapidly adapt and detect novel objects using only minimal annotated training examples, much like how humans can recognize a new concept from limited examples.

Nevertheless, achieving reliable detection with limited samples poses significant technical challenges. Unlike few-shot classification tasks that focus solely on identifying object categories, object detection additionally requires precise localization of objects within complex scenes that typically include multiple objects and cluttered backgrounds [7, 8]. Novel objects in a scene can be easily obfuscated by overlapping objects or visually ambiguous backgrounds, complicating object detection. Moreover, the challenge of achieving robust learning from limited samples is further compounded by the variability of object appear-

ance, texture and context [9]. The scarcity of training examples prevents models from learning robust feature representations that are invariant to appearance variations caused by viewpoint changes, scale differences, and illumination conditions.

Another fundamental challenge is effectively distinguishing novel objects from visually ambiguous backgrounds. In object detection, the background class typically dominants in an image, often causing detectors trained primarily on base classes to misclassify unfamiliar objects as background [10]. This issue is aggravated by limited exposure to novel objects during training, leading detectors to conservatively label uncertain regions as background. In complex or cluttered scenes where object boundaries are low-contrast, novel objects easily blend into the background, further obfuscating the detector's decision process [11]. Addressing this challenge requires explicitly modeling the background category, empowering the detector to effectively learn discriminative features robust that reliably distinguish objects from the background, even under limited training data.

Historically, most high-performing detection frameworks employed two-stage detection architectures, such as R-CNN family [12, 13, 14], where region proposals are first generated and subsequently classified. Few-shot detection methods, such as Meta R-CNN proposed by Yan et al. [7], extended this paradigm by introducing meta-learning techniques at the region-of-interest (RoI) level, thereby adapting the detector to novel classes with only a few examples. This was a pioneering approach that achieved promising results on few-shot detection benchmarks. Nonetheless, two-stage few-shot detectors have notable shortcomings. First, they rely on region proposals generated by a class-agnostic Region Proposal Network (RPN). If the RPN is trained mostly on base classes, its proposals for novel-class objects may be suboptimal, leading to a proposal quality gap between base and novel classes [15]. In other words, the detector might never even consider numerous novel objects if the proposal stage fails to highlight them. Second, many two-stage meta-learning detectors handle novel classes in a one-versus-rest strategy, generating adaptation vectors for each novel class individually. These approaches prevent processing multiple novel classes jointly during inference to effectively exploit possible relationships between the novel classes, reducing detection efficiency and accuracy when multiple novel classes must be simultaneously considered [16, 17].

Motivated by these limitations, recent works have pivoted toward single-stage and end-to-end FSOD architectures. Single-stage detectors, which directly predict bounding boxes and classes in one pass without intermediate proposal generation, offer a more unified pipeline that can be fully differentiable and potentially more flexible for meta-learning. Notably, Meta-DETR proposed by Zhang et al. [15], a recent approach that extends the Detection Transformer (DETR) framework [18] to few-shot scenarios, reformulates detection as a set prediction problem that seamlessly incorporates support-set information. The end-to-end paradigm can reason globally over both the query image and the support examples, eliminating the need for hand-crafted post-processing components such as non-maximum suppression (NMS) [19]. However, the performance of current approaches is still fundamentally bottlenecked by the quality of the feature representations fed into the transformer decoder. When the extracted visual features lack the discriminative power to distinguish a novel object from its surroundings, the subsequent attention mechanisms cannot compensate for this representational deficiency. Furthermore, while these end-to-end approaches successfully address the proposal-quality gap, they often rely on implicit background modeling through its self-attention mechanism, which can still be insufficient when novel objects are highly camouflaged or visually ambiguous. Consequently, advancing FSOD performance requires two critical innovations: developing a more powerful representation-learning mechanism that can capture robust features from minimal samples, and building an explicit background modeling mechanism that enhances the discriminability between novel instances and complex backgrounds.

1.2 Problem Overview

Despite recent progress, Few-Shot Object Detection remains a fundamentally challenging problem due to several intrinsic technical difficulties in learning robust representations from scarce data and distinguishing novel objects from complex backgrounds.

The primary challenge lies in extracting effective and discriminative feature representations for novel categories when training examples are extremely limited. Recent FSOD frameworks typically depend on high-level feature extractors, such as Convolutional Neural

Networks (CNNs) [20] or Vision Transformers (ViTs) [21], which are primarily pre-trained on base classes. Consequently, these feature extractors often struggle to effectively capture distinctive visual characteristics of novel objects, especially when such objects are sparsely represented or entirely absent during the initial training stage [22]. This limitation results in weak or non-discriminative feature responses for novel categories, causing detectors to exhibit significant overfitting to base-class patterns and failing to generalize effectively to new instances [23]. Thus, a critical hurdle is empowering the feature extractor to capture transferable visual cues (e.g., shape, texture, and edge details) that can distinguish novel categories, even from minimal annotated examples.

Furthermore, background confusion and visual clutter can significantly compromise the accuracy of novel object detection. In real-world scenarios, the background can be highly cluttered (e.g., leaves, branches, urban environments) and novel objects might have low-contrast boundaries, making them hard to isolate [24]. For instance, a novel-class animal blending into foliage may be missed because its features look too closely resemble the surrounding background patterns that the model has learned to ignore. This issue is aggravated by class imbalance: during training, models are exposed to significantly more background regions than novel objects, biasing them toward predicting "background" by default when facing uncertainty. As a result, explicit background modeling becomes essential to provide the detector with discriminative representations that can distinguish novel objects from diverse background patterns.

While existing frameworks have attempted to address these issues, they face structural limitations. Two-stage FSOD detectors (e.g., Meta R-CNN) rely on Region Proposal Networks (RPN) to isolate objects, but these RPNs often fail to generate high-quality proposals for unseen novel classes, creating a bottleneck that downstream classifiers cannot recover from [25]. Conversely, recent single-stage FSOD approaches (e.g., Meta-DETR) eliminate intermediate proposal generation stages entirely. The transformer's self-attention and cross-attention mechanisms allow it to consider the entire image context and all support class information jointly [15]. However, this holistic, image-level detection approach is inherently more challenging, demanding robust feature extraction and sophisticated attention mechanisms to accurately detect and discriminate novel objects against complex

backgrounds without explicit object proposals [26]. When visual features lack the discriminative power to separate a novel object from its surroundings, even sophisticated attention mechanisms struggle to compensate.

From the above analysis, it becomes clear that addressing FSOD's challenges requires improvements on multiple fronts. In particular, a better feature extraction mechanism and explicit background modeling are essential, especially for visually ambiguous and complex detection scenarios. A powerful backbone like a Vision Transformer (ViT) provides strong semantic features, but it may lose fine details that are crucial for distinguishing a novel object from a similar-looking background. Enhancing the backbone with additional cues, such as frequency-domain information, can supply critical details (object textures and edges) that enhance object discrimination against the background. Complementarily, explicitly modeling background patterns through a dedicated learnable prototype can profoundly mitigate background confusion by providing clear discriminative criteria to effectively distinguish candidate object regions from irrelevant background clutter. Such an integrated approach combining advanced feature refinement techniques and explicit background modeling strategies is crucial for reliably detecting novel objects in few-shot learning scenarios, thereby establishing a robust foundation for effective few-shot object detection frameworks.

1.3 Research Contributions

To address the aforementioned challenges, this thesis proposes an enhanced FSOD framework built upon a single-stage, transformer-based detection architecture. The key technical contributions of this work are summarized as follows:

1.3.1 Wavelet-Semantic Fusion Attention for Feature Refinement

We proposed a novel feature fusion mechanism that augments high-level semantic representations extracted from a Vision Transformer (ViT) backbone with rich frequency-domain cues. Specifically, our approach applied a discrete wavelet transform to the input image to

capture fine-grained frequency information, such as textures and edges, which is often suppressed or down-sampled in conventional feature extraction pipelines. These wavelet-based features are then adaptively fused with semantic features through a cross-modal attention mechanism that learns to emphasize frequency components most complementary to the semantic representation. This adaptive fusion preserves both fine-grained local detail and global semantic context, yielding more discriminative and robust feature representations for novel objects.

This contribution directly addresses the challenge of inadequate feature extraction for novel categories. The refined features enhance the detector's ability to recognize subtle structural and textural variations in novel objects, even from very limited samples. Empirical evaluations demonstrate that incorporating Wavelet-Semantic Fusion Attention module improves the detection of novel instances, particularly for objects with distinctive textures or fine details that are otherwise difficult to capture.

1.3.2 Learnable Background Prototype

In this thesis, *background* refers to any image region that does not correspond to a labeled foreground object category in the current detection setting. Under standard FSOD detection annotations, background includes (i) true scene background (e.g., sky, road, foliage) and (ii) unlabeled objects from categories outside the base/novel categories, which are treated as negatives [1].

In our work, we introduce a mechanism that explicitly models the background as a learnable prototype within the same representational space as foreground class prototypes. During training, the model learns to form a background embedding that captures the statistical characteristics of non-object regions by aggregating features from areas that do not belong to any foreground class. Inspired by open-set detection ideas, this background prototype acts as an anchor for the "none-of-the-above" category. At inference phase, the detector can compare candidate object features against this background prototype. If a candidate is more similar to the background prototype than to any known class prototype, the model can correctly suppress it as background.

In our pipeline, our learnable background prototype is integrated into the classification head of the detector, effectively providing the model with an explicit, data-driven notion of background appearance, addressing the prevalent issue of background confusion and false positives. By learning background representations jointly with foreground prototypes, the detector achieves more stable decision boundaries and robust suppression of background noise, even in cluttered scenes. Furthermore, because the background prototype is learned adaptively, it reduces class imbalance effects in few-shot learning by capturing consistent background statistics across both base and novel instances.

1.3.3 Unified End-to-End Framework with Inter-Class Contrastive Learning

We integrate the above two modules into a unified single-stage detection framework inspired by Meta-DETR, resulting in a fully end-to-end FSOD model. Our pipeline employs a transformer-based detector (building on Deformable DETR) that directly predicts object instances conditioned on the support set, thus eliminating proposal-generation dependencies. Within this pipeline, the Wavelet-Semantic Fusion Attention module augments the feature representations, while the Learnable Background Prototype is incorporated into the detector's output layer to refine classification scores. In addition, we utilizes an inter-class contrastive training strategy to further enhance category discriminability. Specifically, a contrastive loss encourages embeddings of the same class to cluster closely while pushing apart those of different classes, including the background prototype.

Overall, the integrated design preserves the advantages of Meta-DETR (no proposal stage, joint multi-class support, global reasoning) while enhancing it with our novel modules. The result is a unified pipeline that learns end-to-end to detect novel objects more reliably by (a) enriching feature representations and (b) explicitly modeling background and inter-class relationships.

1.4 Thesis Structure

The following chapters are organized as follows:

- Chapter 2 is the literature review. We first discuss general frameworks and core challenges associated with FSOD. Then, we systematically review existing FSOD approaches, categorizing them into two-stage and single-stage methodologies.
- Chapter 3 is the detailed methodology of our FSOD architecture. We begin by presenting an overview of our single-stage, transformer-based framework. Following this, we elaborate on the design and integration of our two primary contributions: Wavelet–Semantic Fusion Attention and the Learnable Background Prototype. For each component, We provide clear motivation, theoretical grounding, and implementation, demonstrating how they overcome identified limitations.
- Chapter 4 covers the details of the experimental configuration and results. We outline our chosen datasets (PASCAL VOC and MS COCO) and clearly describe the evaluation metrics employed. Subsequently, we specify the training procedures, hyperparameter settings, hardware configurations, and data preprocessing techniques. The main results section includes rigorous performance comparisons against state-of-the-art FSOD approaches, complemented by detailed ablation studies and qualitative analyses that validate the effectiveness of our proposed modules.
- Chapter 5 is the conclusion and future work. We reflect on the strengths and limitations of our proposed FSOD model, highlighting its practical significance and theoretical implications. Then, we discuss promising avenues for future research, suggesting possible extensions and enhancements that could further advance the field of few-shot object detection.

Chapter 2

Literature Review

2.1 Object Detection Frameworks

Modern object detection framework has undergone a remarkable transition from multi-stage, proposal-heavy systems [12, 13, 14] to single-stage, end-to-end architectures driven by Transformers [27, 18]. Understanding this trajectory is essential because FSOD and its variants adapt and extend these baselines rather than invent detection from scratch. Below, we summarize two-stage and single-stage detectors and highlight how their inductive biases, design choices, and post-processing requirements affect robustness in scarce-data settings.

2.1.1 Two-Stage Detection Frameworks

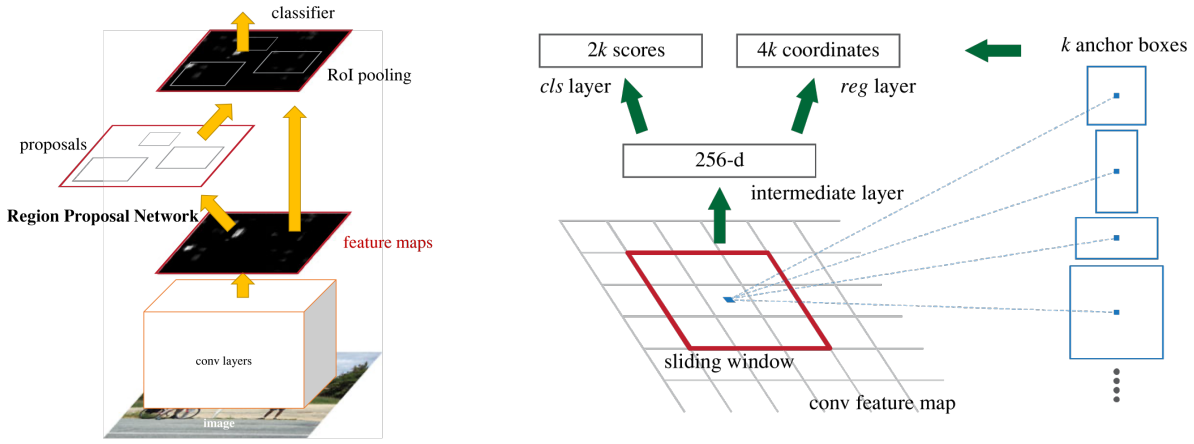
The two-stage paradigm traces back to the R-CNN family:

- **R-CNN** [12]: R-CNN introduces the concept of casting object detection as a region-level classification task on category-agnostic proposals. It uses an external algorithm (e.g., Selective Search [28]) to generate candidate region boxes, then warps each region to a fixed size and extracts deep CNN features for classification with a separate bounding-box regression to refine localization.

- **Fast R-CNN** [13]: Fast R-CNN streamlines the original R-CNN pipeline by performing a single pass of the CNN over the entire image and employing a Region of Interest (RoI) pooling layer to efficiently extract features for each proposal from the shared feature map. This design allowed for unified, end-to-end training of both classification and bounding-box regression, significantly enhancing detection speed compared to R-CNN's region-wise processing.

- **Faster R-CNN** [14]: To further improve efficiency, Faster R-CNN proposes the Region Proposal Network (RPN), a lightweight network that leverages the shared convolutional features of the detector backbone to directly generate object proposals (Figure 2.1). This integrated architecture eliminates the need for external proposal generation, rendering the two-stage detection pipeline fully end-to-end, thereby significantly enhancing both detection speed and accuracy.

These methods utilize a first-stage network to propose candidate regions and a second-stage network for region classification and refinement, often employing Feature Pyramid Networks (FPN) proposed by Lin et al. [29] to enhance multi-scale object detection capabilities.



(a) (b)

Figure 2.1: Overview of two-stage object detection architecture. (a) Faster R-CNN framework with Region Proposal Network (RPN) and detection head. (b) Detailed structure of RPN for generating region proposals. Reprinted from Ren et al. [14] © 2016 IEEE.

As demonstrated in Figure 2.1, the Region Proposal Network (RPN) plays a central role in modern two-stage frameworks such as Faster R-CNN, significantly shaping their efficiency and performance. Specifically, the RPN efficiently generates region proposals by sliding a small window over the shared convolutional feature maps produced by the detector backbone, simultaneously predicting objectness scores and bounding-box offsets relative to predefined anchor boxes varying in scales and aspect ratios [14]. While significantly accelerating the detection pipeline by reducing the proposal search space, the two-stage design necessitates per-proposal processing through RoI operations and separate classification heads, introducing computational overhead. More critically, the anchor-based design inherently biases proposals toward shapes and sizes frequent in the training distribution [30]. This limitation becomes particularly evident in few-shot scenarios, where novel classes often exhibit distinct appearance or size distributions, causing a mismatch with anchors tuned primarily on base classes. Consequently, this proposal quality gap fundamentally constrains two-stage detection on novel classes, as even robust downstream classifiers cannot recover objects missed during the proposal stage [22].

In summary, two-stage methods excel at precise localization, benefiting from mature engineering components such as anchors, FPN, and NMS, leading to broad empirical success. However, these approaches introduce substantial complexity, including multiple heads, proposal heuristics, and post-processing steps that typically require dataset-specific tuning. In FSOD, although class-agnostic proposals can partially alleviate this challenge, the limitation remains: proposals trained primarily on base classes may inadequately serve novel classes, leading the second-stage network unable to detect novel objects missed during the proposal stage.

2.1.2 Single-Stage Detection Frameworks

Single-stage detectors remove the explicit proposal stage by directly predicting object class probabilities and bounding boxes over dense grids of the input image, resulting in a simpler, more efficient detection pipeline compared to two-stage approaches [31]. These methods can be broadly categorized into anchor-based and anchor-free methods, each with notable representative architectures outlined as follows:

Anchor-based Single-Stage Detectors:

- **YOLO** [32]: YOLO (You Only Look Once) popularizes the single-stage detection paradigm by framing detection as a unified regression task. It predicts bounding boxes and class probabilities directly from the entire image in one forward pass, achieving real-time performance by eliminating the need for a separate region proposal stage.
- **SSD** [33]: Single Shot MultiBox Detector (SSD) extends YOLO's single-stage paradigm by introducing a multi-scale feature representation. By predicting bounding boxes and class scores from multiple convolutional layers at different resolutions, SSD significantly improving detection accuracy, particularly for objects of varying sizes, without compromising inference speed.
- **RetinaNet** [34]: RetinaNet addresses the severe class imbalance problem inherent in dense single-stage detection by introducing focal loss, a modified cross-entropy loss that reduces the impact of easy background examples while emphasizing difficult, misclassified instances during training. Combined with a FPN for multi-scale feature extraction, RetinaNet substantially closes the accuracy gap between single-stage and two-stage detectors.

Anchor-free Single-Stage Detectors:

- **CenterNet** [35]: CenterNet detects each object as a triplet of key-points (top-left corner, bottom-right corner, and the center) rather than relying on predefined anchors. It incorporates cascade corner-pooling and center-pooling modules to enrich corner features and central region features respectively, thereby improving both detection precision and recall while maintaining real-time inference speed [36].
- **FCOS** [37]: Fully Convolutional One-Stage Object Detection (FCOS) reformulates object detection as a per-pixel classification and regression task, entirely eliminating reliance on anchor boxes and region proposals. It directly predicts objectness scores and bounding-box offsets at each location on feature maps, thus avoiding anchor-related hyperparameters and simplifying the pipeline while achieving competitive accuracy.

Transformer-based End-to-End Single-Stage Detector

Detection Transformer (DETR) reformulated object detection as a direct set prediction problem, eliminating the need for hand-crafted components such as anchor boxes and NMS. Instead, DETR predicts a fixed-size set of N bounding boxes and class labels in a single forward pass, and employs the Hungarian algorithm to establish optimal one-to-one matching between predictions and ground-truth objects, substantially reducing duplicate detections [18]. This set-based global loss ensures each ground-truth object corresponds to exactly one prediction, enabling end-to-end training without heuristic post-processing.

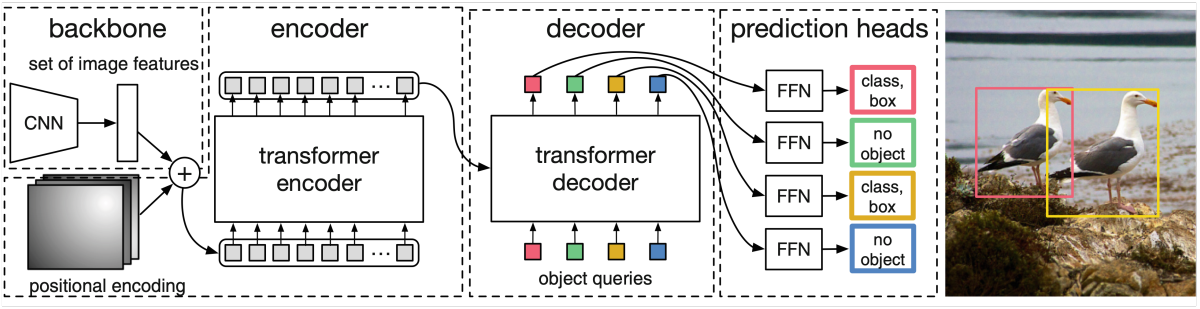


Figure 2.2: Overview of the Detection Transformer (DETR) architecture. Reprinted from Carion et al. [18].

As shown in Figure 2.2, DETR employs a Transformer encoder–decoder architecture that operates directly upon CNN feature maps. A CNN backbone first extracts high-level features, which are then flattened and supplemented with positional embeddings before entering a multi-head self-attention encoder. In parallel, a fixed number of learned object queries are input to the Transformer decoder, where each query attends to encoder features through cross-attention to gather contextual information about specific objects. The decoder outputs N embeddings (one per query), each processed by a shared feed-forward network to predict a class labels (or "no object") and bounding box coordinates. Leveraging its global self-attention mechanism, DETR captures object relationships and global image context, enabling it to effectively model complex scenarios, such as overlapping or grouped objects, within a unified, holistic and parallelizable detection pipeline.

Deformable DETR proposed by Zhu et al. [38] was introduced to address vanilla DETR's

slow convergence and limited small-object detection caused by the computationally expensive global attention mechanism. It replaces global attention with a multi-scale deformable attention module, which allows each query to focus on a small set of relevant feature locations across different scales. Specifically, each query attends only to a limited number of sampling points around learned reference positions, substantially reducing complexity from quadratic to linear and enabling efficient multi-scale feature aggregation. Consequently, Deformable DETR achieves significantly faster convergence and significantly improved small-object detection, making it more practical and effective for real-world applications.

The end-to-end nature of DETR makes it particularly well-suited for few-shot object detection. By eliminating anchor tuning, region proposals, and NMS, DETR provides a flexible framework that can naturally incorporate novel classes and support examples without task-specific modifications. Its object queries and cross-attention mechanisms enable direct reasoning over both support and query images, allowing the model to adapt to new categories with minimal training data. This streamlined design significantly reduces the need for extensive class-specific tuning, which is a critical advantage when only a handful of annotated examples are available for novel categories. This balance between structured guidance and end-to-end flexibility motivates the design considerations underlying FSOD methods, which we discuss next.

2.2 Feature Extraction in Object Detection

Modern object detectors, regardless of two-stage or single-stage paradigm, rely heavily on a robust feature extraction backbone [31]. Essentially, the backbone is the initial deep network that processes the input image and outputs rich visual feature maps, which downstream detection heads then utilize for object classification and localization. Consequently, the effectiveness of feature extraction directly impacts overall detection performance, as more powerful backbones produce richer and more robust feature representations, thereby achieving higher accuracy.

The evolution of object detection closely mirrors advancements in feature extraction architectures. Early detection frameworks, such as R-CNN, demonstrated a quantum

leap in accuracy by replacing hand-crafted features with CNN embeddings pre-trained on ImageNet [12, 39]. Subsequent improvements in backbone architectures led to consistent gains in detection benchmarks, making high-capacity CNNs the de facto choice for feature extraction [40, 7].

In FSOD, the role of the feature extraction module becomes even more prominent due to the scarcity of annotated examples for novel classes. FSOD approaches heavily rely on generalizable features previously learned from abundant base classes or external datasets rather than attempting to learn novel representations from scratch. Many state-of-the-art (SOTA) FSOD methods leverage a frozen backbone pre-trained on large datasets, adapting only the higher-level layers to novel categories [8]. For instance, the simple fine-tuning baseline TFA proposed by Wang et al. [22] freezes the entire backbone and region proposal network, updating only the classification head for novel categories. This strategy reflects a broader principle: if the backbone has learned semantically rich and invariant features, those can be directly reused to recognize novel instances.

Recent research indicates that representation learning under data scarcity remains a primary challenge in FSOD [1, 26]. Inadequate feature representations risk model overfitting to non-generalizable characteristics, impairing effective discrimination between novel classes and background regions. A key focus in FSOD research is improving feature quality via better backbones or additional training signals, such as the contrastive embedding losses employed by FSCE proposed by Sun et al. [41] to enhance class separation. Consequently, robust feature extraction is the foundation that enables few-shot detectors to generalize beyond their limited training examples.

Convolutional Neural Networks (CNNs) Convolutional neural networks (CNNs) have historically served as the foundational component in object detection frameworks [31]. CNNs generate hierarchical feature maps through layers of localized convolutions and pooling, naturally forming a multi-scale representation of the input image. This hierarchy aligns well with detection tasks requiring fine details captured in early layers. The inductive biases of CNNs (locality, translation invariance, and spatial hierarchy) make them data-efficient and robust, enabling effective training of object detectors even on

mid-sized datasets [42]. For instance, R-CNN leveraged ImageNet-pretrained AlexNet proposed by Krizhevsky et al. [43] to dramatically improve object detection performance, far surpassing previous models based on Histogram of Oriented Gradients (HOG) features [12]. Subsequent detectors adopted progressively deeper CNN backbones: Fast R-CNN initially utilized VGG-16 proposed by Simonyan et al. [44], but ResNet-50/101 proposed by He et al. [20] soon became standard as they significantly boosted detection accuracy without introducing proposal-stage overhead. Modern CNN backbones, such as DarkNet proposed by Redmon et al. [45] and EfficientNet proposed by Tan et al. [46], continued to improve feature representation, though accuracy gains diminished while computational costs continued to rise. Moreover, CNN backbones are commonly augmented with feature pyramid networks (e.g., FPN [29], PANet proposed by Liu et al. [47], etc.) for multi-scale feature representation. By combining feature maps from different convolutional stages, detectors can detect objects at various scales more effectively.

In summary, CNNs excel at learning generalizable visual features even from relatively limited datasets, a strength especially beneficial in FSOD tasks. However, CNN-based backbones exhibit notable limitations in FSOD: their localized receptive fields struggle to effectively capture long-range contextual relationships, and the convolution-generated features can lack sufficient expressiveness to distinguish nuanced differences among novel classes with limited annotated examples. These limitations have increasingly motivated researchers to explore transformer-based feature extractors, which promise enhanced global context modeling and richer feature representations, better addressing the demands of object detection scenarios.

Vision Transformers (ViTs) The Vision Transformer (ViT) architecture offers an alternative approach to feature extraction, using self-attention mechanisms to capture global relationships in an image, thereby addressing the limited receptive field constraints of CNNs [21]. Unlike CNNs that process local neighborhoods, ViTs operate on image patches and learn long-range dependencies through self-attention, achieving a global receptive field from the very first layer, as illustrated in Figure 2.3. This ability to capture distant dependencies proves advantageous for object detection, where contextual information from nearby objects

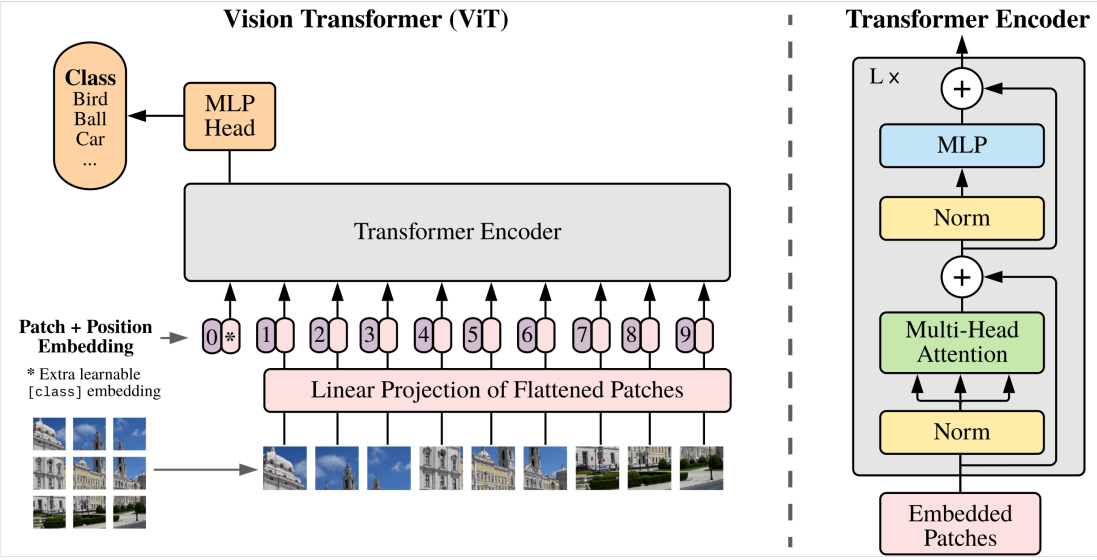


Figure 2.3: Overview of the Vision Transformer (ViT) architecture. Reprinted from Dosovitskiy et al. [21].

and background regions can significantly improve recognition accuracy. Early attempts to integrate ViTs into detection, such as DETR, kept a CNN backbone and used transformers only in the detector head [18]. However, more recent research has explored replacing the entire backbone with transformer encoders. For example, ViTDet proposed by Li et al. [48] explores plain, non-hierarchical ViT backbones for object detection, demonstrating that a pure transformer backbone pretrained on ImageNet-1K can compete with previous leading methods based on hierarchical ResNet backbones with minimal adaptations. The Swin Transformer proposed by Liu et al. [49] further introduces a hierarchical, multi-stage Vision Transformer with shifted-window self-attention and patch-merging, producing multi-resolution feature maps that form a feature pyramid analogous to CNN backbones. This hybrid of windowed self-attention and hierarchical design achieved state-of-the-art results on ImageNet-1K, COCO, and ADE20K. The key advantages of vision transformers lie in their capacity for modeling complex global interactions and rich feature representations, which can surpass the performance of traditional CNN features when sufficient training data is available.

However, transformers are known to be data-hungry and can overfit on small datasets due to their lack of built-in inductive biases [21]. Consequently, in low-data scenarios like FSOD, training a ViT backbone from scratch usually yields significantly worse performance than CNN-based methods. To effectively exploit the representational strengths of ViT in FSOD„ A few recent works have explored ViT-based backbones in few-shot detectors. For example, FCT proposed by Han et al. [50] replaces the conventional CNN with a pyramid ViT in a two-branch FSOD model, applying multi-level cross-attention cross-image self-attention between support and query features to improve the alignment of feature representations for novel category. These developments suggest that with proper pretraining and architectural tweaks, ViT-based backbones can outperform CNNs in detection even under few-shot conditions.

Self-Supervised Visual Foundation Models A significant advancement in feature extraction for object detection has emerged with the rise of visual foundation models. These models, such as contrastive vision-language models and self-supervised ViTs, aim to serve as universal feature extractors that can be adapted to a wide range of downstream tasks. In particular, self-supervised visual models like DINOv2 proposed by Oquab et al. [51] have shown remarkable effectiveness as backbones for object detection. DINOv2 is a ViT pretrained on a curated dataset of 142 million images through a combination of teacher-student self-distillation and an iBOT-inspired masked image modeling objective applied at the patch level, without any class labels [51, 52]. Its patch-level reconstruction mechanism forces the model to learn spatially fine-grained features, which is "friendly to downstream detection tasks" by making the representation localization-sensitive. The resulting feature representations are particularly effective for FSOD, where the model must generalize to novel categories while maintaining fine-grained spatial awareness from limited training samples. For example, FM-FSOD proposed by Han et al. [26] found that a frozen DINOv2 backbone yields significantly better FSOD performance than traditional ImageNet-pretrained CNNs or even other modern backbones like CLIP proposed by Radford et al. [53] or Masked Autoencoders (MAE) proposed by He et al. [54]. More recently, DE-ViT proposed by Zhang et al. [55] leverages a forzen DINOv2 ViT encoder and introduces a region-propagation network with prototype-based feature projection to produce robust

object representations for novel categories, which achieved state-of-the-art accuracy on multiple few-shot benchmarks, substantially outperforming previous methods. Meanwhile, segmentation-focused models like SAM2 proposed by Ravi et al. [56] provide powerful priors for locating objects in an image, which could be adapted to serve FSOD by generating proposal masks for novel categories. These approaches share the philosophy of *"learning general features once, and using them everywhere."* Foundation-model backbones eliminate the need to train detectors from scratch for each new task or class. Instead, they provide a pre-trained feature space where even rare objects can be recognized through simple adaptation, whether by fine-tuning a lightweight head or prompting a separate model.

Another line of research integrates contrastive vision-language models like CLIP into FSOD, which is trained to align images with textual descriptions, producing an embedding space where visual concepts have semantic meaning. In FSOD, this translates to embedding the support images or class names and computing their similarity to query region features. Some methods employ CLIP to re-score or filter proposals for novel categories by comparing image regions with class text embeddings [26]. This effectively leverages external knowledge to recognize objects: when an image region's features align with the text "bird" in CLIP's space, the region likely contains a bird, even if the detector's classifier was initially unsure. This approach reduces fine-tuning requirements by leveraging CLIP's pre-trained representations, which encode rich semantic knowledge. For example, CDMM-FSOD proposed by Shangguan et al. [57] addresses FSOD by leveraging rich textual semantics as an auxiliary modality to establish robust knowledge relationships between visual instances and language descriptions through multi-modal feature aggregation and bidirectional text semantic rectification. Recent studies find that insufficient utilization of modern foundation models is not enough for further improvements in FSOD [26]. The few-shot detector requires careful design to fully exploit the richer features.

In summary, the evolution of feature extraction in object detection reflects three distinct developmental trends, each prioritizing different aspects of representation learning:

1. **From Hand-crafted to Deep Inductive Biases (CNNs):** The initial wave moved from manual features to CNNs, leveraging local receptive fields and translation

invariance to efficiently capture hierarchical spatial patterns, though often limited by a lack of global context [12, 20, 29].

2. **From Local to Global Context Modeling (ViTs):** The second trend shifted toward Vision Transformers, replacing rigid inductive biases with self-attention mechanisms capable of modeling long-range dependencies and global shape bias, though at the cost of higher data requirements [18, 21, 48].

3. **From Task-Specific to Universal Foundation Models:** The most recent trend, exemplified by models like DINOv2, utilizes large-scale self-supervised pre-training to produce robust, "ready-to-use" visual features that generalize well to downstream tasks even with minimal supervision [53, 51].

Our work grounds its methodology firmly within the latest trajectory. We adopt a self-supervised Vision Transformer (DINOv2) as the backbone to leverage its powerful, generalizable semantic representations. However, we identify that while foundation models excel at high-level semantics, they may overlook the fine-grained, high-frequency details that are crucial for precise localization in few-shot scenarios. To bridge this gap, our proposed approach augments the foundation model's representation with explicit frequency-domain cues, thereby combining the robustness of the modern foundation model era with the detail-preserving strengths of signal processing.

2.3 Few-Shot Object Detection

Few-Shot Object Detection (FSOD) addresses the challenging problem of detecting objects from novel categories with only a limited number of annotated examples. Formally, we involve two sets of categories: a set of base classses $\mathcal{C}_{\text{base}}$ with abundant annotated instances, and a set of novel classes $\mathcal{C}_{\text{novel}}$ where only K annotated examples (shots) are available per class, i.e., $|\mathcal{C}_{\text{base}}| \gg |\mathcal{C}_{\text{novel}}|$ and $\mathcal{C}_{\text{base}} \cap \mathcal{C}_{\text{novel}} = \varnothing$, typically termed K-shot detection, where $K \in \{1, 2, 3, 5, 10, 30\}$.

The learning process typically follows a two-phase paradigm:

1. **Base Training:** The model is pretrained on abundant base-class data $\mathcal{C}_{\text{base}}$ to learn robust, generalizable feature representations.
2. **Few-shot Fine-tuning:** The model is adapted to recognize both $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ simultaneously to ensure the model consistently detect both previously encountered and newly introduced object categories. During this phase, the number of instances per category for both $\mathcal{C}_{\text{base}}$ and $\mathcal{C}_{\text{novel}}$ is limited to K.

The fundamental objective is to minimize the generalization error on $\mathcal{C}_{\text{novel}}$ given the extremely limited support set. This formulation presents a sharp contrast to traditional object detection, which assumes identical training and testing distributions with abundant data. In FSOD, the detector must overcome the distribution shift between base and novel categories while suppressing overfitting to the sparse support set.

As highlighted in recent literature, this data-scarce setting introduces four critical technical bottlenecks:

- **Representation learning under data scarcity**: Limited annotated data often causes models to overfit to irrelevant or spurious features and inadequately represent intra-class diversity, thereby weakening generalization and reducing discriminative capability against visually similar classes.
- **Proposal quality disparity**: RPNs primarily trained on numerous base-class data, frequently underperform in generating high-quality proposals for novel-class instances.
- **Background confusion**: Novel-class objects appearing in cluttered scenes are prone to misclassification as background or as visually similar base-class objects.
- **Optimization fragility**: FSOD methods show high sensitivity and variance concerning support-set selection, resulting in inconsistent and unstable performance.

To systematically explore and address these challenges, existing FSOD approaches can be generally categorized into two representative paradigms: Transfer-learning-based and

Meta-Learning-Based approaches. Each category is reviewed in detail in the subsequent sections.

2.3.1 Transfer-learning-based FSOD Approaches

Transfer-based FSOD begin with a detector pretrained on base classes with sufficient training data and then carefully fine-tune a small subset of parameters on novel classes [58]. This strategy prioritizes simplicity and leverages knowledge transfer, often achieving strong empirical performance under extreme data scarcity. In contrast to meta-learning pipelines which require complex episodic training, transfer-based methods involve a straightforward two-phase training (base pretraining followed by fine-tuning) and thus avoid meta-training overhead, as illustrated in Figure 2.4. When combined with strategies such as selective parameter freezing, balanced sampling between base and novel classes, and specialized losses to handle class imbalance, these methods have proven to be competitive baselines in few-shot detection scenarios.

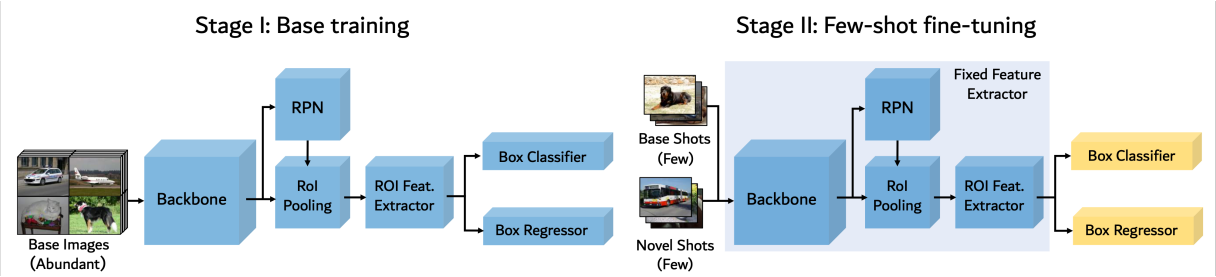


Figure 2.4: Overview of the Two-Stage Fine-tuning Approach (TFA) architecture. Reprinted from Wang et al. [22], licensed under CC BY 4.0.

- **TFA** [22]: TFA (Two-Stage Fine-tuning Approach) freezes the backbone and class-agnostic RPN (trained on base classes) and fine-tunes only the final classification and regression layers on a balanced mix of base and novel examples. A cosine-normalized classifier is used to stabilize gradients and suppress class imbalance during fine-tuning. This recipe established a strong baseline without any meta-learning episodes.

- **FSCE** [41]: Noting that misclassification (rather than localization error) is the main cause of novel-class failures, FSCE (Few-Shot object detection via Contrastive proposals Encoding) adds an auxiliary contrastive learning branch on RoI features. This branch pulls same-class proposal features closer and pushes different-class proposals apart in feature space, sharpening decision boundaries. Building upon TFA, FSCE slightly relaxes NMS and jointly fine-tunes the FPN, RPN, and detection heads, rather than maintaining these components in a frozen state. These modifications consistently improve novel class average precision (AP) over TFA by reducing inter-class confusion.
- **DeFRCN** [23]: Decoupled Faster R-CNN (DeFRCN) addresses the issue of conflicting gradients commonly observed in the standard two-stage fine-tuning. It introduces a Gradient Decoupled Layer to stabilize RPN training during few-shot adaptation, thereby alleviating interference between RPN and R-CNN stages. Additionally, DeFRCN incorporates a Prototypical Calibration Block that integrates metric-based class prototypes with conventional softmax predictions to enhance classification accuracy under limited data conditions. These enhancements improve training stability in extremely low-shot scenarios and mitigate catastrophic forgetting of previously learned base classes.

Empirically, transfer-based methods perform well when the pretrained backbone and RPNs already yield sufficiently object-centric proposals for novel classes, provided the class-imbalance between base and novel data is explicitly addressed during fine-tuning. However, they retain two inherent limitations from the two-stage detection paradigm: (i) a strong reliance on proposal quality leads to biases in RPN objectness scores learned primarily from base classes, potentially resulting in missed novel instances, and (ii) the RoI-centric fine-tuning approach fails to leverage contextual relationships among multiple novel-class objects, as each RoI is processed independently. These limitations become particularly pronounced under distribution shifts or in cluttered scenes, indicating that while transfer-based approaches efficiently reuse learned representations, they fall short of comprehensively addressing all challenges in FSOD.

2.3.2 Meta-learning-based FSOD Approaches

While transfer-based methods in FSOD rely on fine-tuning pretrained detectors with minimal supervision, meta-learning offers a fundamentally different solution: training models that can quickly adapt to new tasks by leveraging knowledge gained from previous tasks. Within the FSOD framework, meta-learning approaches simulate few-shot scenarios during the training phase, allowing the detector to acquire transferable knowledge priors that enable efficient generalization to novel object categories [58].

The central idea of meta-learning is to optimize a model not for a single task, but over a distribution of tasks. Each task is structured as an episode, typically composed of a support set (limited annotated examples for each class) and a query set (unseen instances of the same classes). The model learns to perform detection on the query set by conditioning on the support set, effectively "learning to learn" [59]. Over many such episodes, the model internalizes a learning strategy that generalizes well to novel categories.

Figure 2.5 presents a conceptual illustration of this episodic training paradigm. Throughout meta-training (outer loop), the model is exposed to many few-shot tasks, gradually adjusting its parameters to facilitate rapid adaptation (inner loop) when confronted with new categories under low-shot conditions.

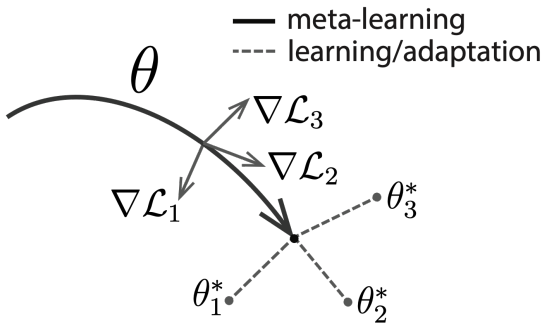


Figure 2.5: Conceptual illustration of episodic meta-learning. The outer-loop trajectory (solid) represents the meta-learning phase, optimizing the model parameters toward a meta-initialization (black dot) strategically positioned close to various task-specific optima. The inner-loop trajectories (dashed) denote rapid adaptation to novel tasks, achieved through a few gradient updates from the meta-initialization.

2.3.2.1 Two-Stage Meta-Learning FSOD Approaches

Two-stage meta-learning FSOD methods typically build upon the R-CNN-based framework by incorporating meta-adaptive components into both the proposal generation and classification stages [58]. These architectures generally comprise a backbone network coupled with a Region Proposal Network (RPN), followed by Region-of-Interest (RoI) pooling and category-specific prediction heads. Meta-learning modules are strategically integrated to modulate these stages conditioned on support examples, thereby enabling the detector to generalize to novel classes with limited samples.

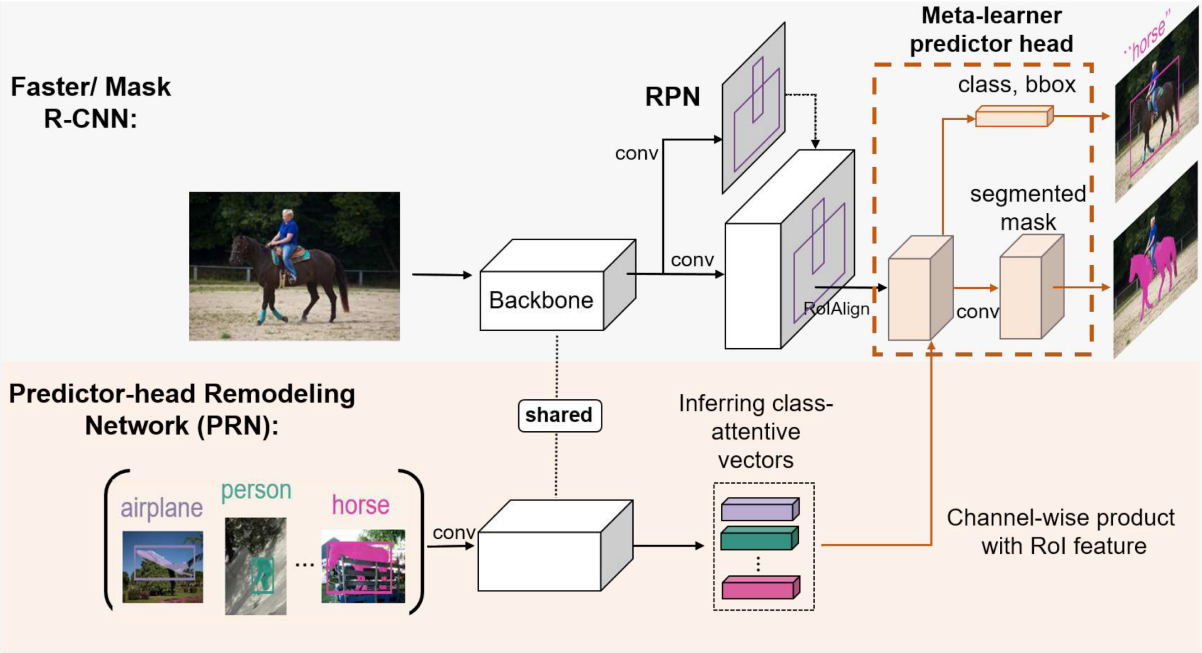


Figure 2.6: Overview of the Meta R-CNN architecture. Reprinted from Yan et al. [7] © 2019 IEEE.

Meta R-CNN Meta R-CNN represents a seminal work in integrating meta-learning into two-stage object detectors, and has since inspired numerous variants and extensions in the FSOD literature [7]. As illustrated in Figure 2.6, the framework employs the Predictor-head Remodeling Network as a support encoder to extract class-specific prototypes, which

subsequently modulate RoI features in the query image via a channel-wise attention mechanism. The model performs detection on a per-class basis by conditioning each query on a single support class, followed by independent binary classification. While this design retains compatibility with standard Faster R-CNN components and facilitates implementation, it introduces several limitations. First, the class-by-class processing paradigm incurs substantial computational overhead. Second, the model lacks the capacity to jointly reason about relationships across multiple novel classes, limiting its representational flexibility. Moreover, its reliance on a class-agnostic RPN trained exclusively on base categories limits its ability to generate proposals for novel-class instances that differ in scale or appearance.

Following Meta R-CNN, several works have advanced the two-stage meta-learning paradigm by targeting limitations such as weak proposal recall, spatial misalignment, and inefficient support-query fusion. These methods integrate deeper interaction modules and alternative architectures to improve FSOD performance.

- **Meta Faster R-CNN** [25]: This method enhances the RPN by introducing a prototype-matching mechanism that computes similarity between region proposals and support class prototypes, enabling class-aware objectness scoring. Additionally, it incorporates an attentive feature alignment module to reduce spatial misalignment between support and query features. This design improves novel-class recall and localization accuracy, particularly in cluttered scenes.

- **FCT** [50]: Fully Cross-Transformer (FCT) introduces a multi-level cross-attention framework that facilitates progressive fusion between support and query features at multiple backbone layers. This hierarchical interaction enhances semantic alignment and supports both coarse and fine-grained feature matching, leading to improved robustness in detecting small or occluded objects.

- **DE-ViT** [55]: DE-ViT proposes a ViT-based architecture with region-propagation mechanisms that construct dense, class-specific attention masks across the image. It further projects RoI features into a prototype-aligned subspace for final classification, eliminating the need for fine-tuning. This approach improves generalization and allows rapid deployment across diverse novel classes.

In summary, two-stage meta-learning FSOD methods have evolved from the pioneering Meta R-CNN toward more sophisticated support-query interaction mechanisms, addressing key challenges in proposal generation, feature alignment, and cross-class reasoning. While these approaches demonstrate strong localization accuracy and generalization capabilities, their multi-stage pipeline inherently incurs computational overhead, motivating the exploration of more efficient one-stage alternatives.

2.3.2.2 Single-Stage Meta-Learning FSOD Approaches

While two-stage frameworks offer modularity and fine-grained control, they also depend on proposal generation and heuristic post-processing mechanisms, which can prevent effective adaptation to novel categories. To address these limitations, an alternative line of research explores single-stage detectors that unify detection into an end-to-end pipeline, obviating the need for region proposals and anchor boxes [58]. By combining one-stage architectures with meta-learning principles, early works in this direction streamline both training and inference procedures, reduce proposal distribution bias, and directly leverage holistic image-level context for few-shot generalization.

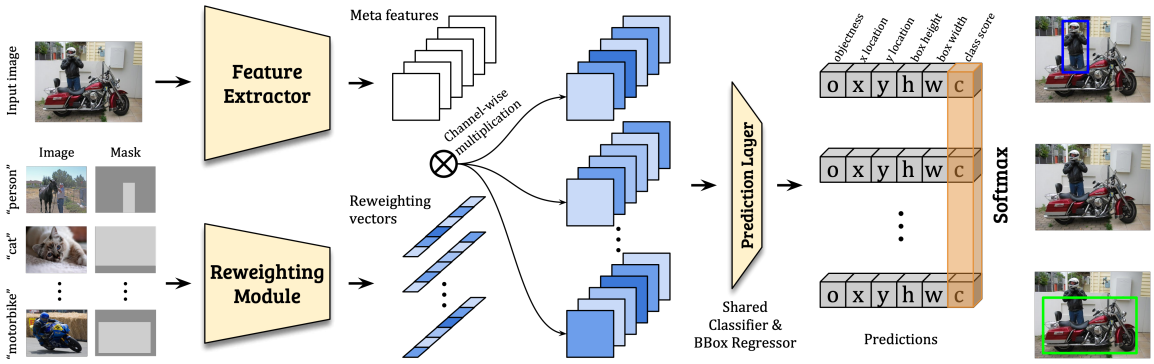


Figure 2.7: Overview of the FSRW architecture. Reprinted from Kang et al. [60] © 2019 IEEE.

As a seminal contribution, FSRW proposed by Kang et al. [60] established the paradigm of meta-learning-based feature adaptation within a proposal-free, single-stage Meta-Learning FSOD framework. As shown in Figure 2.7, it augments a single-stage detector with a

lightweight meta-learning module that dynamically reweights convolutional feature maps based on limited support examples. Specifically, a compact support encoder transforms limited annotated instances into a set of channel-wise weights, which are subsequently applied to modulate the query image's feature representation and amplify class-relevant signals. By embedding support information as feature modulation, the detector is adaptively biased toward novel-class features, enabling effective detection of unseen categories from limited training samples.

Despite their simplicity and effectiveness, earlier single-stage meta-learning methods faced limitations related to accuracy and adaptability. This motivated further innovations aimed at improving representation learning and feature interaction mechanisms, ultimately leading to more advanced architectures.

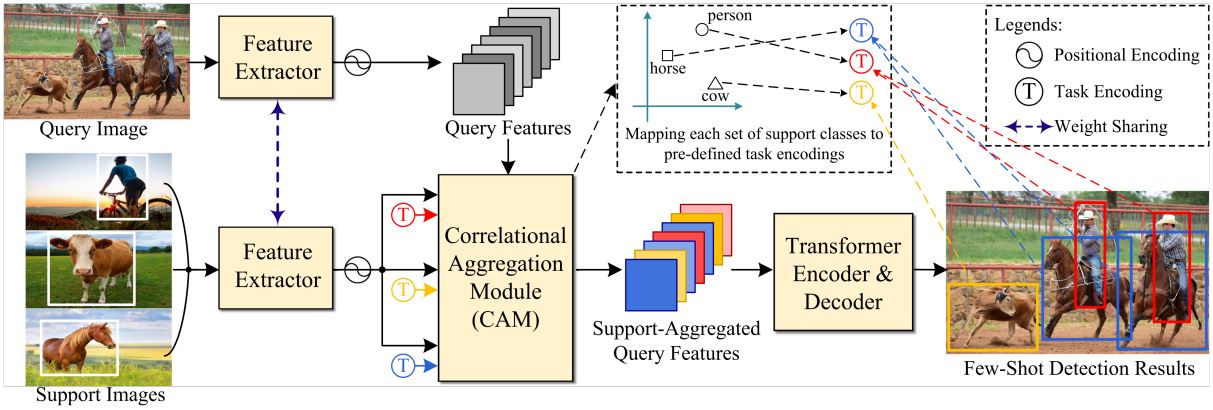


Figure 2.8: Meta-DETR is architecturally composed of four primary modules: (i) a CNN-based backbone for multi-scale feature extraction; (ii) a support encoder that aggregates visual features from few-shot exemplars into class-specific prototype representations; (iii) a Transformer encoder that captures global spatial relationships within the query image; and (iv) a Transformer decoder that integrates query image features with support-derived class prototypes via cross-attention mechanisms. Each object query interacts with both the query image and the class prototype to yield final detection predictions. Reprinted from Zhang et al. [15], licensed under CC BY 4.0..

Meta-DETR represents a significant milestone in FSOD by integrating meta-learning principles into a single-stage, DETR–based Transformer framework, as illustrated in Figure 2.8. It reformulates few-shot detection as a unified set prediction problem, where object

queries dynamically attend to both query image features and class-aware support representations to generate detection outputs in a fully end-to-end manner [15]. By eliminating the reliance on proposal generation, anchor boxes, and heuristic post-processing mechanisms, the architecture streamlines both the training and inference pipelines.

To fully leverage the expressive capacity of this design, Meta-DETR introduces several architectural innovations that enable robust and scalable meta-adaptation:

- **Correlational Meta-Training**: Meta-DETR adopts a multi-class support training strategy wherein the Transformer decoder is conditioned on support exemplars from multiple classes concurrently, departing from the conventional single-class conditioning paradigm. By enabling the decoder to attend to support features across several classes simultaneously, the model learns to explicitly capture and exploit inter-class relationships during the meta-training phase. This design encourages the learning of discriminative representations that leverage both inter-class similarities and contrastive differences, thereby enhancing generalization to novel categories.

- **Elimination of Region Proposals**: Following the DETR design, Meta-DETR eliminates the reliance on RPN and obviates the need for anchor boxes or region-based processing pipelines. This design circumvents the proposal quality gap commonly observed in two-stage FSOD methods, where proposal generators trained exclusively on base classes fail to produce high-quality candidates for novel object categories. Instead, learnable object queries implicitly serve as content-aware region indicators, enabling the Transformer to directly localize and classify instances across the entire image in a holistic manner.

- **Set-Based Prediction**: Meta-DETR adopts DETR's one-to-one bipartite matching mechanism between predictions and ground-truth objects, which facilitates end-to-end set prediction while eliminating the dependence on heuristic post-processing operations such as NMS. This streamlined pipeline proves particularly advantageous for few-shot generalization, as conventional NMS with fixed confidence thresholds often fails to adapt to the diverse characteristics of novel classes with limited training examples.

- **Inter-Class Correlation Reasoning**: In contrast to prior methods that process each novel class independently through one-versus-all classification, Meta-DETR performs joint reasoning across all support classes within a unified forward pass. The decoder's multi-head attention mechanism enables each object query to simultaneously assess its compatibility with all class prototypes, thereby suppress confusion among visually similar novel categories and enhancing cross-class detection consistency.

2.3.2.3 Comparative Analysis: Synthesis and Choice of Paradigm

Beyond architectural differences outlined above, a meaningful comparison between two-stage and single-stage meta-learning paradigms hinges on three often overlooked factors: (i) score calibration and prediction consistency across episodic tasks, (ii) computational scaling with support classes and proposals, and (iii) robustness to label noise and outliers within support sets. Taken together, these factors reveal fundamental differences in inductive biases that substantially influence few-shot generalization capabilities, transcending the architectural comparisons previously discussed.

First, two-stage architectures inherits a score coupling problem: objectness scores (generated by the RPN) and classification confidences (produced by RoI heads) are learned from partially disjoint label distributions and subsequently combined through hand-crafted heuristics. During transfer from base to novel classes, the relative calibration between these score components can drift, degrading the global ranking of detections under strict evaluation thresholds. While techniques such as class-aware objectness modeling and cross-stage feature alignment offer partial mitigate this [25, 50], the final detection ranking still depends on the interaction of multiple independently-trained modules and manually-tuned thresholds. In contrast, single-stage meta-learning frameworks, such as Meta-DETR, generate a unified confidence per object query through a single training objective with one-to-one matching [15]. This architectural consistency reduces sensitivity to inter-module score drift and yields more stable precision.

Second, the two paradigms scale differently with the number of proposals and support classes. Two-stage methods must process a variable (often substantial) set of region proposals,

with many implementations employing class-wise conditioning that incurs near-linear computational overhead proportional to the number of novel categories during inference. In contrast, single-stage approaches operate within a more predictable computational budget: a fixed set of object queries attends jointly to both query image features and a consolidated bank of support prototypes, such that computational complexity is primarily determined by the number of queries rather than by proposal multiplicity or per-class passes [60, 15]. In practice, this yields more consistent inference throughput and facilitates capacity planning in few-shot scenarios requiring accommodation of varying numbers of novel classes.

Third, approaches vary significantly in their robustness to noisy or unrepresentative support examples. Two-stage pipelines that operate in a class-by-class and RoI-centric manner, typically construct decision boundaries closely aligned to specific support instances, making them vulnerable to outliers and annotation artifacts. Practices such as hard-negative mining and NMS further exacerbate this sensitivity. Conversely, single-stage methods aggregate support features before the decoding step. Techniques such as FSRW-style reweighting or prototype banks utilized in Meta-DETR allow these models to average signals across multiple examples and effectively reduce the impact of atypical support features via attention mechanisms [60, 15]. This support aggregation acts as an implicit denoising prior and is especially beneficial when K is extremely small or intra-class variance is high.

These paradigm differences also influence methodological ergonomics. Two-stage meta-learning exposes numerous interacting hyperparameters (proposal sampling, IoU thresholds, per-branch losses), making the training surface rugged and reproduction sensitive to implementation details [1]. Improvements to one component can inadvertently destabilize others. In contrast, single-stage set-prediction models consolidate supervision into a unified objective, reducing brittle tuning points and simplifying hyperparameter searches.

Beyond performance considerations, the choice of foundational architecture proves crucial for extensibility. Two-stage designs are deeply tied to RoI processing and proposal heuristics that complicate integration with modern pretraining strategies or multi-modal priors. However, DETR-style encoder–decoder structure provides clean and straightforward attachment points for richer support encoders, correlation-aware prototype aggregation [15].

In light of the above, this thesis adopts Meta-DETR as our foundational architecture not solely for its elimination of proposal reliance but also due to its methodological advantages.

2.4 Summary

This chapter reviewed the evolution of modern object detection from proposal-heavy two-stage pipelines to end-to-end set prediction with Transformers. Two-stage methods localize precisely but face proposal biases and multi-head calibration issues under data scarcity. Single-stage detectors simplify training through anchor-free designs and improved label assignment that suppress heuristic sensitivity, while DETRs further reframes detection as set prediction with one-to-one matching enabling global reasoning. We then examined feature extraction backbones, highlighting the trade-offs between CNN inductive biases and ViT global modeling, and summarized evidence that self-supervised visual foundation models provide localization-friendly representations for FSOD. Finally, we summarized key challenges in Few-Shot Object Detection (FSOD). We reviewed transfer-learning and meta-learning paradigms, highlighting the simplicity yet proposal bias of transfer-based methods and the adaptive generalization offered by meta-learning approaches. After comparing two-stage and single-stage meta-learning FSOD frameworks, we emphasized Meta-DETR's unified architecture, computational efficiency, and robustness to support-data variations.

While existing FSOD frameworks have made significant progress, challenges remain in effectively leveraging global context and suppressing sensitivity to support-data variations. Our research addresses these limitations by introducing enhancements to the Meta-DETR architecture, integrating richer support-query interaction mechanisms and robust feature extraction techniques into a unified FSOD framework, as detailed in the next chapter.

Chapter 3

Methodology

In this chapter, we present an enhanced few-shot object detection model that builds upon the End-to-End Detection Transformer (DETR) architecture through meta-learning principles. Our model is dedicated to addressing several limitations of existing few-shot object detection architectures, including how to enable the model to learn multi-modal object features and how to improve the model's ability to distinguish between objects and backgrounds, particularly in scenarios where object-background boundaries are highly ambiguous. As illustrated in model architecture, our approach consists of two components:

- **Wavelet-Semantic Fusion Attention for Feature Refinement**: To address the limitations in traditional few-shot object detection frameworks, we propose Wavelet-Semantic Fusion Attention, a cross-modal feature refinement strategy. Our approach utilizes a pretrained ViT backbone for semantic feature extraction, complemented by Wavelet-Semantic Fusion Attention to incorporate frequency-domain details like edges, corners, and textures. By adaptively integrating semantic and frequency-domain information, this attention mechanism notably enhances feature discriminability and robustness, especially beneficial in challenging FSOD scenarios.
- **Learnable Background Prototype** To stabilize the decision boundary between objects and background and prevents prototype contamination in few-shot scenarios,

we explicitly introduce a learnable background prototype that lives in the same space as class prototypes and participates in all support–query interactions. Rather than treating background as a residual category, the prototype competes with foreground classes during similarity computation so that background-like regions are attracted away from class prototypes.

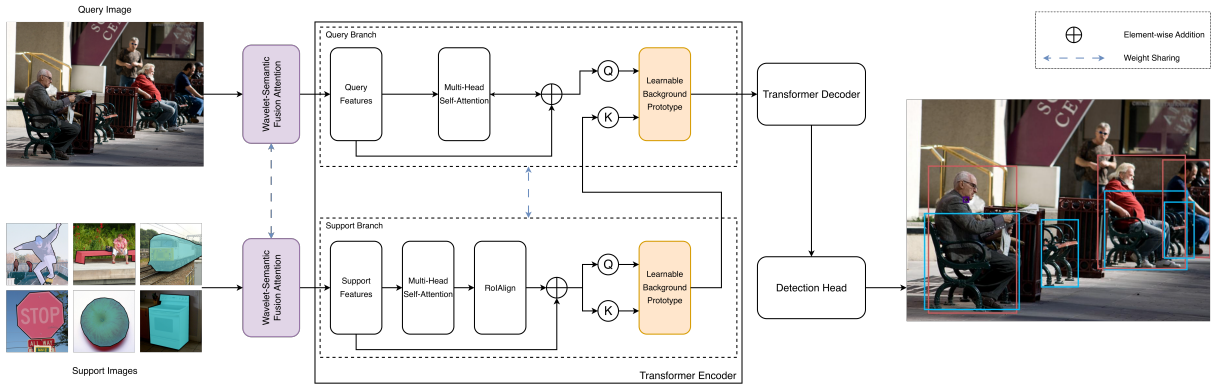


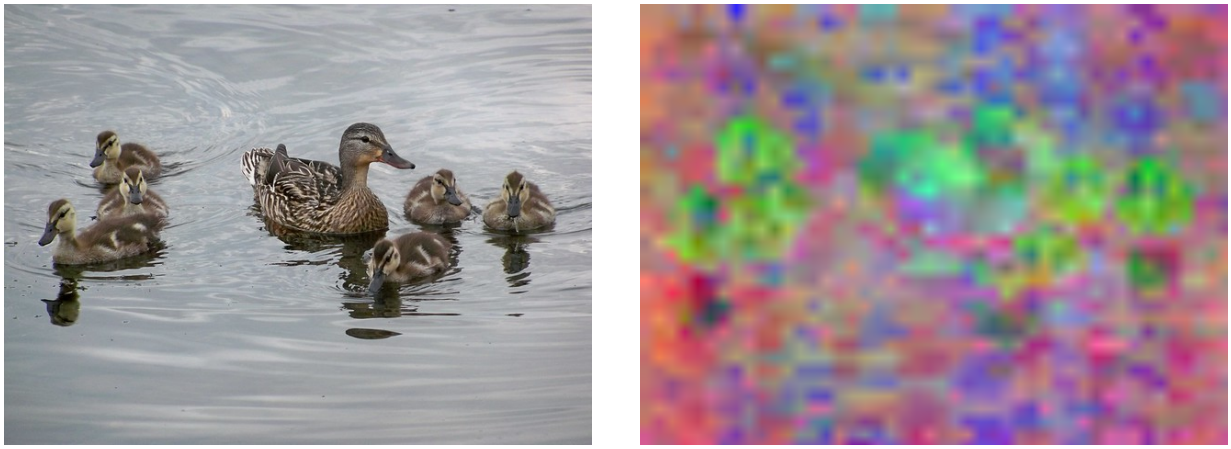
Figure 3.1: Overview of our proposed FSOD architecture.

To integrate the two proposed components into a unified detection framework, we adopt a meta-learning pipeline inspired by Meta-DETR [15]. Specifically, we replace its original semantic feature extractor with our proposed Wavelet-Semantic Fusion Attention module and further enhance the model by explicitly addressing feature confusion with the Learnable Background Prototype module.

As illustrated in Figure 3.1, both the query image and support examples first pass through the weight-shared Wavelet-Semantic Fusion Attention module to yield refined query and support features. These features are further refined by a lightweight multi-head self-attention. On the support side, robust class prototypes are constructed via RoI-Align-based attention aggregation, while the query side retains dense spatial representations. Meanwhile, the Learnable Background Prototype is introduced in both branches to explicitly suppress background interference and maintain a stable foreground–background boundary through explicit background supervision. To enhance the separability among class prototypes, an inter-class contrastive loss based on the InfoNCE criterion is then employed.

The remaining architecture follows a single-stage transformer encoder-decoder structure, which predicts object instances directly from the refined features and learned object queries. A class-agnostic detection head outputs class probabilities aligned to the support set sequence, seamlessly handling variable class counts through learnable background placeholders. This design preserves Meta-DETR's one-pass efficiency while ensuring robust alignment between support and query representations within a unified detection framework.

3.1 Wavelet-Semantic Fusion Attention for Feature Refinement



(a) (b)

Figure 3.2: Semantic features extracted using the DINOv2 ViT model. (a) Original image. (b) PCA visualization of extracted features, clearly capturing semantic distinctions, but lacking fine-grained textures and precise boundary details.

As demonstrated in Figure 3.2, pre-trained vision transformers such as DINOv2 have shown remarkable success in capturing high-level semantic representations of target objects through their attention-based feature extraction mechanisms. However, scenarios involving cluttered backgrounds, ambiguous object-background boundaries, or excessive background noise can severely impair the effective extraction of these representations. This impairment

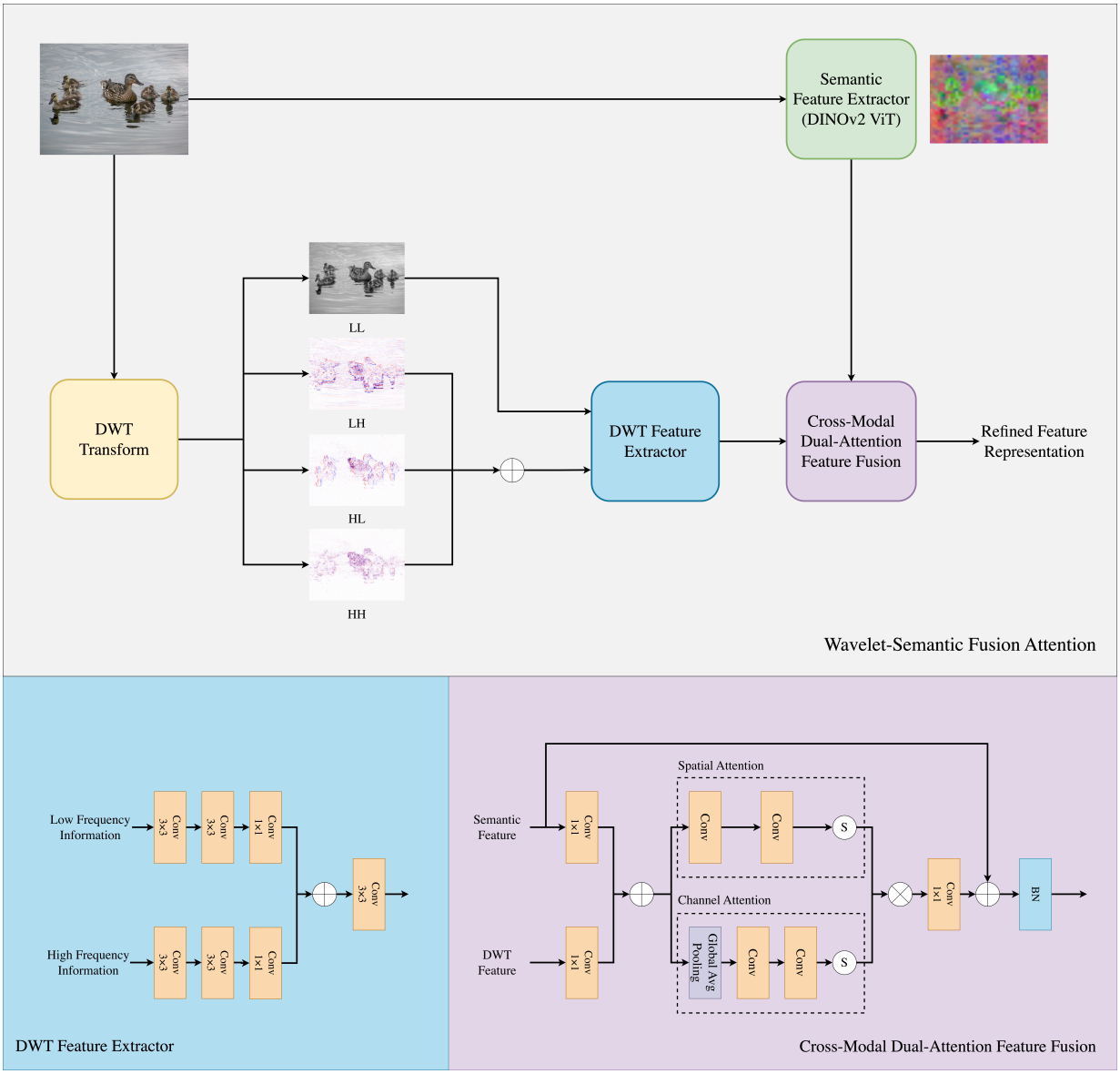


Figure 3.3: Overview of our proposed Wavelet-Semantic Fusion Attention module.

creates a fundamental bottleneck in few-shot object detection systems, where accurate feature representations are critical for learning from limited training data.

To address these limitations, we propose a wavelet-informed attention-based feature fusion module that synergistically combines semantic representations from DINOv2 with complementary frequency-domain features extracted through wavelet decomposition. As illustrated in Figure 3.3, our approach leverages the insight that semantic and frequency-domain features capture fundamentally different but complementary aspects of visual information. This fusion establishes a robust foundation that enhances feature discriminability in challenging scenarios while preserving the valuable semantic understanding from pre-trained backbones, thereby creating robust representations particularly suitable for few-shot object detection tasks.

Wavelet-based Frequency Analysis and Extraction

The discrete wavelet transform (DWT) [61] provides an elegant mathematical framework for multi-scale frequency-domain analysis, decomposing signals into both frequency and spatial domains. For a 2D image signal $f(x, y) \in \mathbb{R}^{H \times W}$, the 2D DWT is defined as:

$$W_\psi(j, m, n) = \frac{1}{\sqrt{2^j}} \sum_{x=0}^{H-1} \sum_{y=0}^{W-1} f(x, y) \psi^* \left(\frac{x - 2^j m}{2^j}, \frac{y - 2^j n}{2^j} \right) \tag{3.1}$$

where ψ represents the mother wavelet function, j denotes the scale parameter, (m, n) are the translation parameters, and ψ^* is the complex conjugate of ψ. In our method, we employ a Daubechies-4 (db4) wavelet basis, which offers an optimal trade-off between computational efficiency and frequency localization properties.

A key question raised regarding the WSFA module is the selection of Daubechies-4 over the computationally simpler Haar wavelet. Our choice is grounded in two fundamental signal processing properties: regularity and vanishing moments. Although Haar is computationally efficient, it is mathematically characterized by discontinuous step functions. When applied to natural images, these discontinuities introduce artificial 'blocking' artifacts. In deep learning contexts, such artifacts result in a highly irregular loss landscape, making it difficult

for network to converge during the few-shot fine-tuning phase. In contrast, db4 is continuous and smooth, enabling the network to better approximate the natural curves of objects without introducing artificial high-frequency noise. Second, and perhaps more importantly, is the property of vanishing moments. Haar has only one vanishing moment, meaning it reacts to linear gradients as if they were object edges, such as a shadow fading across a wall. By contrast, db4 possesses two vanishing moments, which allows it to mathematically 'ignore' linear illumination trends. This ensures that our WSFA module extracts features based on physical structure, not lighting conditions, providing the robust invariance required for generalization.

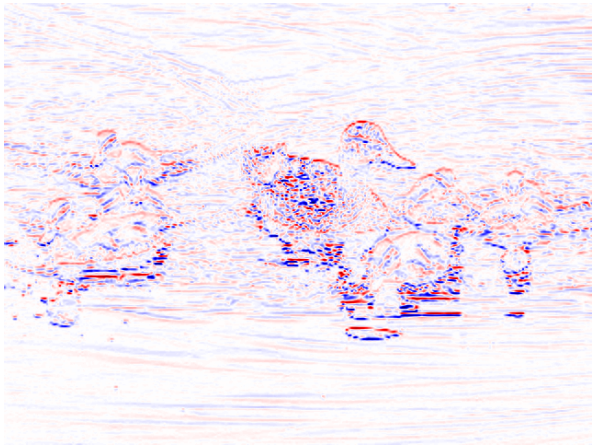
The 2D wavelet decomposition naturally separates an image into four distinct sub-bands: the LL component contains approximation coefficients that capture the fundamental structure of objects, while the high-frequency (LH, HL, and HH) components highlight vertical edges, horizontal edges, and diagonal details and corners, respectively, as shown in Figure 3.4. For an input RGB image $\boldsymbol{I} \in \mathbb{R}^{3\times H\times W}$, where the three channels represent color information, the wavelet transform is applied independently to each color channel. This decomposition yields:

$$\mathrm{WT}(\boldsymbol{I}) = \{\boldsymbol{I}_{\mathrm{LL}}, \boldsymbol{I}_{\mathrm{LH}}, \boldsymbol{I}_{\mathrm{HL}}, \boldsymbol{I}_{\mathrm{HH}}\} \tag{3.2}$$

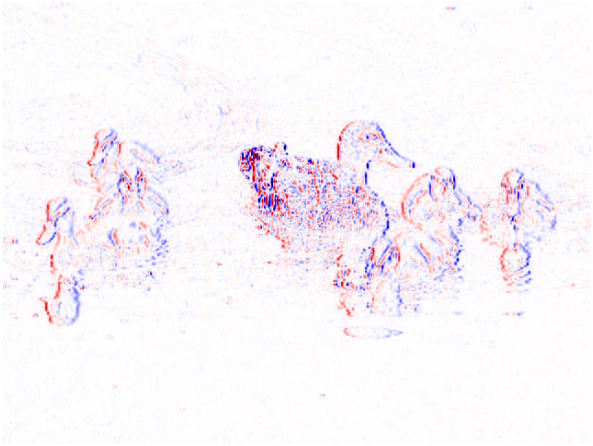
where each sub-band has dimensions $\mathbb{R}^{3\times \frac{H}{2}\times \frac{W}{2}}$, preserving the color channel information throughout the decomposition.

While multi-level wavelet decomposition theoretically offers richer multi-scale representations, our findings reveal that a single-level decomposition is sufficient for few-shot object detection tasks. Deeper levels increasingly capture coarse structure that is already encoded in the ViT semantic features, while the first level preserves fine edge/texture information that is most complementary. The adoption of single-level wavelet also reduces memory usage and improves throughput for the model. This counterintuitive result stems from the complementary nature of wavelet and semantic features. Deeper decomposition levels primarily capture increasingly coarse structural patterns, which is already well-represented in the semantic features extracted from DINOv2 ViT. In contrast, the first-level decomposition preserves fine-grained edge and texture details that are highly complementary

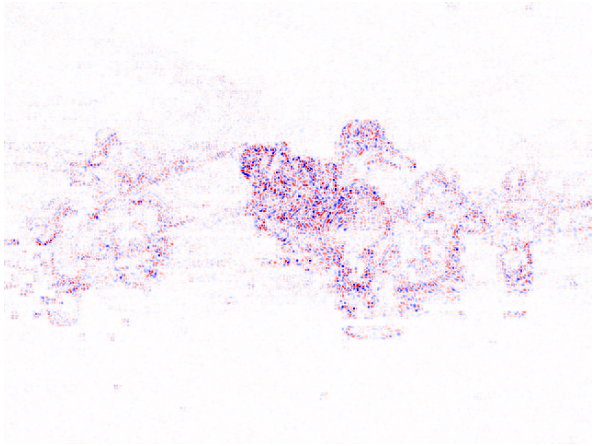
(a) LL Band



(b) LH Band



(c) HL Band



(d) HH Band

Figure 3.4: Frequency information decomposed by DWT from Figure 3.2a, showing approximation (LL) and directional details (LH, HL, HH).

to semantic representations, which also reduces memory and improves throughput. This finding highlights a critical design principle: effective feature fusion relies on strategically optimizing the complementarity between different feature modalities, rather than merely expanding the dimensionality or complexity of the feature representation.

Raw wavelet coefficients require extra transformation into learnable feature representations for effective neural network processing. We design a specialized dual-pathway convolutional architecture that processes low-frequency and high-frequency components independently, recognizing that these components encode fundamentally different types of frequency-domain information requiring distinct processing strategies.

The low-frequency pathway processes the approximation sub-band ($\boldsymbol{I}_{\mathrm{LL}}$), which preserves the original three-channel color structure. This pathway first applies spatial downsampling via a strided convolution, followed by processing through a sequence of convolutional layers with progressively increasing channel dimensions, enabling efficient extraction of coarse-level scene layout information, producing the low-frequency feature $\boldsymbol{F}_{\mathrm{low}}$. This design achieves two complementary objectives: improving computational efficiency and progressively abstracting the structural and color information encoded within the approximation coefficients.

The high-frequency pathway, in contrast, processes the concatenated LH, HL, and HH sub-bands, which collectively encode edge and detail information across three orientations while maintaining the original three color channels, resulting in:

$$\boldsymbol{I}_{\mathrm{high}} = [\boldsymbol{I}_{\mathrm{LH}} \oplus \boldsymbol{I}_{\mathrm{HL}} \oplus \boldsymbol{I}_{\mathrm{HH}}] \in \mathbb{R}^{9\times \frac{H}{2}\times \frac{W}{2}} \tag{3.3}$$

where $\oplus$ denotes concatenation along the channel dimension.

These concatenated high-frequency features are subsequently processed through a similar convolutional architecture mirroring the low-frequency pathway but adapted for the 9-channel input, enabling the network to extract discriminative patterns from the aggregated edge representations, yielding the high-frequency feature representation $\boldsymbol{F}_{\mathrm{high}}$.

This dual-pathway architecture leverages the inherent distinction between low-frequency structural semantics and high-frequency textural details, thereby empowering the network to acquire specialized feature representations optimized for each respective domain.

The processed frequency features from both pathways are then integrated into a unified wavelet feature representation $\boldsymbol{F}_{\text{wavelet}}$ by concatenating their outputs along the channel dimension, followed by a convolutional fusion layer:

$$\boldsymbol{F}_{\text{wavelet}} = \mathcal{C}_{\text{wavelet}}([\boldsymbol{F}_{\text{low}} \oplus \boldsymbol{F}_{\text{high}}]) \tag{3.4}$$

where $\mathcal{C}_{\text{wavelet}}$ is a 1×1 convolution followed by Batch Normalization and ReLU activation.

This final projection reduces the concatenated channels to the desired wavelet feature dimension d_w and enables the network to learn interactions between the processed low and high-frequency information, creating a comprehensive frequency-domain embedding.

Cross-Modal Dual-Attention Feature Fusion

The extracted wavelet features $\boldsymbol{F}_{\text{wavelet}}$ provide a comprehensive frequency-domain representation that is complementary to the semantic features obtained from pre-trained ViT. However, directly combining these two modalities through simple concatenation or addition would fail to capture their fundamentally distinct characteristics and varying importance across spatial locations. For instance, frequency-domain edge representations may be crucial at object boundaries but redundant in regions where semantic features already provide sufficient discrimination ability.

To address this challenge, we propose a convolutional attention-based mechanism that adaptively fuses frequency-domain and semantic information by selectively enhances semantic features with wavelet-derived information based on their contextual relevance. The core innovation lies in learning spatially-aware and channel-specific attention weights that determine the contribution of frequency-domain features at each location. This adaptive weighting is particularly crucial in few-shot object detection scenarios, where novel object categories often exhibit distinctive textural patterns that require flexible integration of frequency-domain cues to achieve robust discrimination against background clutter and semantically similar distractors.

Given semantic feature $\boldsymbol{F}_{\text{ViT}}$ extracted from the input image via DINOv2 ViT and frequency-domain feature $\boldsymbol{F}_{\text{wavelet}}$, we first establish a common representational space to

enable cross-modal interaction. Specifically, we employ learnable 1×1 convolutions to project both features into a shared latent space of dimension d_w, yielding the projected semantic features $\boldsymbol{F}'_{\text{ViT}}$ and wavelet features $\boldsymbol{F}'_{\text{wavelet}}$. This projection preserves the spatial structure of feature maps without flattening, thereby maintaining computational efficiency and spatial awareness. The aligned dimensionality enables the generation of meaningful attention weights that effectively correlate information across both modalities.

Subsequently, the projected feature maps are concatenated along the channel dimension:

$$\boldsymbol{F}_{\text{combined}} = [\boldsymbol{F}'_{\text{ViT}} \oplus \boldsymbol{F}'_{\text{wavelet}}] \tag{3.5}$$

The unified feature representation $\boldsymbol{F}_{\text{combined}}$ encodes both high-level semantic information and fine-grained frequency details, enabling the network to learn an adaptive, context-aware fusion strategy that leverages complementary information from both spatial and frequency domains.

Our attention mechanism operates through two complementary pathways: spatial attention and channel attention. The spatial-attention pathway generates a location-specific weight map that identify which regions would benefit most from frequency-domain enhancement. In few-shot detection scenarios, this is particularly important as novel objects may have distinctive edge patterns or textural signatures at specific locations that differ from the base classes. Given the unified feature representation $\boldsymbol{F}_{\text{combined}}$ as input, the spatial attention is computed through a lightweight convolutional bottleneck architecture, as shown in the following formulation:

$$\boldsymbol{A}_{\text{spatial}} = \sigma(\text{Conv}^{1}_{1\times1} \circ \text{ReLU} \circ \text{BN} \circ \text{Conv}^{d_w/4}_{1\times1}(\boldsymbol{F}_{\text{combined}})) \tag{3.6}$$

where σ denotes the sigmoid activation function, ReLU is Rectified Linear Unit activation function and BN represents Batch Normalization. The proposed bottleneck design employs channel reduction to achieve three objectives: minimizing computational burden, enforcing compact representation learning of spatial features, and acting as an inherent regularization strategy to prevent overfitting, which is particularly crucial in few-shot learning scenarios with limited training data.

The channel-attention pathway complements the spatial attention by modeling which semantic-frequency channels are most informative for refinement, independently of spatial location. Given the input $\boldsymbol{F}_{\text{combined}}$, the channel attention is computed as:

$$\boldsymbol{A}_{\text{channel}} = \sigma(\text{Conv}_{1\times1}^{d_w} \circ \text{ReLU} \circ \text{Conv}_{1\times1}^{d_w/4} \circ \text{GAP}(\boldsymbol{F}_{\text{combined}})) \tag{3.7}$$

where GAP represents Global Average Pooling. The resulting channel attention map $\boldsymbol{A}_{\text{channel}}$ is broadcast across the spatial dimensions to modulate the wavelet projection per channel, allowing the model to prioritize orientation-specific and texture-sensitive cues while suppressing channels that are uninformative in the current semantic context. Compared with the spatial-attention pathway that answers "where to attend", the channel-attention pathway addresses "what to attend" by operating in the channel domain to learn what types of frequency patterns should be emphasized.

The two attention pathways are combined multiplicatively to produce a unified wavelet-semantic dual-attention gating mask only applied to the projected wavelet stream. This asymmetric design choice was made deliberately for stability and interpretability. With the semantic feature $\boldsymbol{F}_{\text{ViT}}$ treated as an anchor that already captures global structure and category-level cues, the gated wavelet response injects fine and localized corrections only where evidence in the frequency domain is compelling. In particular, it enables selective incorporation of object boundaries, texture, and structural details that complement semantic representations, while relying primarily on semantic features in texture-less backgrounds and homogeneous regions where frequency-domain information provides minimal value.

The modulated wavelet features are then projected back to the original semantic feature dimension and integrated via a residual connection, followed by Batch Normalization for training stability. The residual pathway offers two critical benefits in our setting. First, it provides an identity fallback mechanism: in regions where wavelet evidence is weak or noisy, the module can bypass the wavelet branch and rely solely on the pre-trained DINOv2 backbone, preventing degradation. Second, it establishes a clear gradient path that allows the attention mechanism to learn when not to interfere with the semantic representations, which is crucial for few-shot scenarios where limited annotations may not cover the full variability of background patterns and imaging conditions. Additionally, a mild dropout on

the projected residual branch serves as an regularization mechanism to prevent overfitting to specific frequency patterns, encouraging the model to learn more robust and generalizable representations across diverse imaging scenarios.

Conceptually, our proposed module can be viewed as a learned, data-adaptive edge and texture prior that is conditioned on semantic context. Unlike fixed handcrafted priors, the wavelet decomposition provides orientation-selective and scale-aware frequency analysis through its distinct sub-bands that are robust against moderate photometric and geometric variations. The attention mechanism, in turn, learns to selectively activate these frequency-domain features in semantically ambiguous regions (e.g., occluded boundaries or background clutter), while suppressing them where they could be misleading. We found that a modest boost in edge fidelity and texture representation can substantially improve both localization accuracy and classification confidence. Our conditional fusion design also offers robustness to domain shift: When the target domain exhibits different texture statistics from pre-training data, the wavelet stream provides stable, low-level cues that does not rely on category-specific semantic context. The attention mechanism can emphasize only those frequency cues that are consistent with the current scene's semantics, thereby avoiding brittle overreliance on textures alone.

Fusion Loss and Optimization

Our proposed module is optimized jointly with the underlying detection objective. In addition to the primary task loss (classification and localization from the detector head), we introduce an auxiliary fusion loss $\mathcal{L}_{\text{fusion}}$ that explicitly regularizes the wavelet-semantic integration. This loss encourages the enhanced features to remain semantically aligned with the original representations while incorporating meaningful and structure-aware adjustments from the frequency domain.

Let $\boldsymbol{F} \in \mathbb{R}^{B\times C\times H\times W}$ denote the original semantic features extracted from DINOv2 for a mini-batch, and let $\tilde{\boldsymbol{F}} \in \mathbb{R}^{B\times C\times H\times W}$ be the refined features produced by our wavelet-semantic fusion attention with residual projection back into the semantic space. The fusion loss is designed following three principles reflected in our implementation.

Semantic Consistency: The first component is the semantic consistency loss $\mathcal{L}_{\text{consistency}}$, which penalizes large deviations from the original semantic representation based on mean squared error (MSE):

$$\mathcal{L}_{\text{consistency}} = \frac{1}{B \times H \times W} \sum_{b=1}^{B} \sum_{i=1}^{H} \sum_{j=1}^{W} \left\| \tilde{\boldsymbol{F}}_{b,:,i,j} - \text{sg}(\boldsymbol{F}_{b,:,i,j}) \right\|_2^2 \tag{3.8}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation, and $\| \cdot \|_2$ represents the ℓ_2 norm applied to the channel dimension.

By treating the original features $\boldsymbol{F}$ as a fixed teacher target through the stop-gradient operation, we block the semantic consistency loss from back-propagating gradients to the DINOv2 encoder. This term preserves the robust, generalizable representations learned from large-scale pre-training, which are particularly valuable in few-shot scenarios where fine-tuning on limited data could lead to catastrophic forgetting. Meanwhile, it establishes an asymmetric optimization paradigm similar to knowledge distillation, where the refined features $\tilde{\boldsymbol{F}}$ are encouraged to stay anchored to the semantic "teacher" while allowing the fusion module to learn meaningful and localized refinements. This anchoring mechanism prevents the refinement from overfitting to spurious frequency patterns in the limited training samples, thereby maintaining the semantic stability essential for generalizing to novel object categories.

Directional Alignment: The second component defines the directional consistency loss between original and refined features. Rather than allowing arbitrary changes in feature space, this term encourages the refinement to adjust primarily the magnitude along existing semantic directions, avoiding rotations into semantically unrelated subspaces. This is achieved through a cosine similarity-based penalty:

$$\mathcal{L}_{\text{direction}} = 1 - \frac{1}{B \times H \times W} \sum_{b=1}^{B} \sum_{i=1}^{H} \sum_{j=1}^{W} \frac{\tilde{\boldsymbol{F}}_{b,:,i,j} \cdot \boldsymbol{F}_{b,:,i,j}}{\|\tilde{\boldsymbol{F}}_{b,:,i,j}\|_2 \cdot \|\boldsymbol{F}_{b,:,i,j}\|_2 + \varepsilon} \tag{3.9}$$

where $\varepsilon > 0$ is a small constant (e.g., 10^{-8}) for numerical stability.

While $\mathcal{L}_{\text{consistency}}$ controls the magnitude of deviations, $\mathcal{L}_{\text{direction}}$ constrains their orienta-

tion, ensuring that enhancements amplify or attenuate existing semantic dimensions rather than introducing spurious directions. This directional constraint is particularly crucial in few-shot scenarios where limited training data makes the model vulnerable to overfitting to noise or task-irrelevant frequency patterns.

Magnitude Regularization: The third component is a lightweight magnitude regularizer that prevents pathological global rescaling of feature activations. Instead of imposing strict per-channel or per-location constraints, we penalize discrepancies in the average absolute activation between the refined and original features:

$$\mathcal{L}_{\text{magnitude}} = \left| \frac{1}{B \times C \times H \times W} \sum_{b,c,i,j} \left| \tilde{\boldsymbol{F}}_{b,c,i,j} \right| - \frac{1}{B \times C \times H \times W} \sum_{b,c,i,j} \left| \boldsymbol{F}_{b,c,i,j} \right| \right| \tag{3.10}$$

This deliberately coarse regularizer serves as a safeguard against rare failure modes where the attention mechanism might saturate, leading to either uncontrolled amplification or excessive attenuation of feature responses. By constraining only the global activation budget rather than individual spatial or channel-wise statistics, we preserve the fusion module's ability to selectively boost high-frequency cues at critical locations (e.g., object boundaries and textural regions), while preventing systematic drift in overall magnitude of feature activations. Strict per-channel regularization could overly constrain the model's ability to adapt to the diverse frequency characteristics of novel categories. In contrast, our global budget constraint provides stability without sacrificing the localized, adaptive refinements essential for discriminating new objects from background clutter.

The three components are combined into a unified fusion loss, which is computed only during training phase and incurs no additional cost at inference:

$$\mathcal{L}_{\text{fusion}} = \mathcal{L}_{\text{consistency}} + \beta \cdot \mathcal{L}_{\text{direction}} + \lambda \cdot \mathcal{L}_{\text{magnitude}} \tag{3.11}$$

where the hyperparameters β and λ are set to 0.1 and 0.001, respectively.

The overall training objective augments the detector's task loss with the fusion loss, ensuring semantic consistency during refinement while selectively incorporating wavelet

cues:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{det}} + \alpha_{\text{query}} \cdot \mathcal{L}_{\text{fusion}}^{(q)} + \alpha_{\text{support}} \cdot \mathcal{L}_{\text{fusion}}^{(s)} \tag{3.12}$$

where $\mathcal{L}_{\text{det}}$ is the standard detection objective (classification and box regression) computed on query images; $\mathcal{L}_{\text{fusion}}^{(q)}$ and $\mathcal{L}_{\text{fusion}}^{(s)}$ denote the auxiliary fusion losses applied to query and support streams respectively. Specifically, we set the balancing weights to $\alpha_{\text{query}} = 0.1$ for query images and $\alpha_{\text{support}} = 0.2$ for support images. By enforcing stronger regularization on support embeddings, we stabilizes class prototypes and encourage them to remain close to their pre-trained semantic directions while still admitting localized wavelet-informed refinements at object boundaries and textured regions.

The module is trained end-to-end with the detector in an episodic few-shot setting. The DINOv2 encoder remains frozen throughout training, serving as a stable semantic anchor. A linear warm-up schedule is applied to α_{query} and α_{support} over the first N_{warmup} training iterations to allow the cross-modal dual-attention mechanism to calibrate before fusion constraints take full effect. For small batch sizes (e.g., $B \leq 4$), we freeze batch normalization statistics in the fusion module after an initial calibration phase to maintain training stability.

The fused wavelet-semantic features subsequently serve as enhanced input representations for both query and support images. In the few-shot detection pipeline, the refined support features contribute to constructing robust class prototypes, while query features facilitate precise detection via subsequent classification and bounding box regression heads.

In summary, the proposed Wavelet-Semantic Fusion Attention module integrates complementary high-frequency wavelet details with semantic representations from DINOv2 via an adaptive cross-modal dual-attention mechanism. The introduced fusion loss further ensures semantic consistency and structural robustness of the refined features, effectively addressing challenges in few-shot object detection scenarios characterized by limited data and complex backgrounds.

3.2 Learnable Background Prototype

Traditional few-shot object detection inherently faces a significant challenge in distinguishing target objects from complex and varied backgrounds, particularly when object and background regions share similar visual characteristics. This challenge is profoundly magnified by the scarcity of training samples, which fail to adequately represent the extensive variability found in real-world backgrounds. For instance, in underwater imagery, murky water and floating particles create ambiguous regions where background textures closely resemble the surfaces of objects, severely complicating precise object-background discrimination. Similarly, in industrial defect detection tasks, the intricate surface patterns of defect-free materials often visually mimic actual defects, further increasing difficulties in accurately isolating target objects from background noise. These scenarios necessitate explicit background feature modeling to ensure that irrelevant or misleading background information is effectively suppressed during inference.

Existing few-shot detectors commonly regard the background as implicit negative samples, assuming regions that do not match any class prototype naturally belong to the background. However, such indirect modeling approaches become inadequate when background regions exhibit visual features semantically similar to novel object classes, or when background appearances significantly deviate from patterns encountered during training. Moreover, in attention-based detection frameworks, background patches lacking explicit representation can inadvertently influence the construction of class prototypes through feature similarity computations. This unintended interference can corrupt class representations and substantially increase false-positive detections.

To overcome these limitations, we propose the Learnable Background Prototype (LBP), a module designed to represent the background as a learnable entity participating directly in feature interactions, rather than relying solely on implicit rejection. The core idea behind the LBP module is to position the background prototype in the same representational space as object class prototypes. During similarity-based feature aggregation, the LBP module actively competes with class prototypes by attracting features corresponding to background-like regions, thereby preventing these regions from contaminating the object

class representations.

Our design incorporates two complementary mechanisms: Foreground Feature Refinement (FFR) and Explicit Background Supervision (EBS). The FFR mechanism utilizes the learnable background prototype to refine both support class embeddings and query patch representations through an attention-based competition process. By integrating the background prototype into the similarity computation while excluding it from direct feature aggregation via a gradient-isolated soft-scaling strategy, FFR enables foreground features to dominate the competition against background responses. This ensures that aggregated feature representations remain uncontaminated and focused on object-relevant information. In contrast, the EBS mechanism provides explicit supervision to the background prototype by expanding the detector's output space to include a dedicated background dimension. Through this explicit supervisory signal, EBS guides the background prototype to capture the statistical structure of non-object regions and learn the decision boundary that distinguishes background from target objects. Importantly, this process preserves the bipartite matching for foreground predictions while maintaining stable gradient flow for background learning.

Background Prototype Construction

At the core of the Learnable Background Prototype (LBP) module lies a learnable background embedding $\boldsymbol{e}_{\text{bg}}$, designed to explicitly represent the distinctive background prototype. Unlike hand-crafted background features or heuristically selected negative samples, $\boldsymbol{e}_{\text{bg}}$ is optimized in an end-to-end manner within the episodic meta-learning framework, enabling it to adaptively capture background characteristics specific to the target scenarios. This embedding is initialized randomly and refined iteratively through gradients originating from two primary sources: attention-based competition during feature interaction and explicit supervisory signals from the detection objective.

To effectively integrate learnable background embedding $\boldsymbol{e}_{\text{bg}}$ within the few-shot detection pipeline, we introduce background placeholders into the support set sequence. Specifically, each training episode samples K target classes from the dataset, structuring

the support set into a fixed-length sequence comprising N positions. When the number of sampled classes K is less than the sequence length N, the remaining $N-K$ positions are systematically filled by background placeholders rather than arbitrary padding. Crucially, all background placeholders share the same learnable background embedding $\boldsymbol{e}_{\text{bg}}$, ensuring that the background embedding uniformly participates in sequence-based attention computations alongside class prototypes. This deliberate and uniform utilization of learnable background embedding $\boldsymbol{e}_{\text{bg}}$ allows background features to consistently compete with foreground features during similarity assessments, significantly enhancing discriminative capacity.

This placeholder strategy naturally accommodates variability in the number of target classes encountered during fine-tuning and inference. In real-world few-shot detection scenarios where the exact category counts are often uncertain or variable, the sequence-based arrangement seamlessly adapts to arbitrary numbers of classes by uniformly assigning the shared background embedding $\boldsymbol{e}_{\text{bg}}$ to all unoccupied sequence positions.

Importantly, the learnable background embedding $\boldsymbol{e}_{\text{bg}}$ is globally shared across episodes and is consistently employed within both the support and query encoder branches. This universal sharing encourages the embedding to capture generalized, domain-invariant background characteristics rather than episode-specific patterns. During training, episodes with similar backgrounds provide stable gradients that reinforce the embedding's representational consistency, while diverse episodes prevent it from collapsing into overly specialized modes. Through this unified mechanism, the LBP develops a robust and transferable background embedding that generalizes effectively across diverse environmental conditions.

Foreground Feature Refinement (FFR)

Building upon the design principles introduced above, we next describe the Foreground Feature Refinement (FFR) mechanism, enabling the learnable background prototype to refine both support class embeddings and query patch representations by regulating their interaction with background cues. The FFR mechanism uniformly employs a dual-pathway attention approach within both support and query encoder branches to amplify foreground features while explicitly suppressing background interference.

The support encoder processes a batch of images corresponding to K target classes, yielding the initial ordered sequence $\mathbf{S} = \{\boldsymbol{e}_0, [\text{bg}], \boldsymbol{e}_1, \boldsymbol{e}_2, \ldots\}$, where each $\boldsymbol{e}_n$ denotes the feature embedding for the n-th class, and [bg] represents a background placeholder.

In both branches, the mechanism involves two parallel transformation pathways applied to input embeddings. In the first pathway, each feature embedding $\boldsymbol{e}_n$ undergoes a linear projection via weight matrix $\mathbf{W}_1$, resulting in $\boldsymbol{e}'_n = \mathbf{W}_1\boldsymbol{e}_n$. Here, the background placeholder [bg] is uniformly replaced by the learnable background embedding $\boldsymbol{e}_{\text{bg}}$, forming the augmented embedding sequence:

$$\mathbf{S}' = \{\boldsymbol{e}'_0, \boldsymbol{e}_{\text{bg}}, \boldsymbol{e}'_1, \boldsymbol{e}'_2, \ldots\} \tag{3.13}$$

In the second pathway, embeddings are transformed via another linear projection via weight matrix $\mathbf{W}_2$, generating $\boldsymbol{e}''_n = \mathbf{W}_2\boldsymbol{e}_n$. Crucially, the background placeholder is replaced by a soft-scaled, gradient-isolated background embedding in this pathway, effectively preventing background contamination:

$$\tilde{\boldsymbol{e}}_{\text{bg}} \triangleq \gamma \cdot \text{sg}(\mathbf{W}_2\boldsymbol{e}_{\text{bg}}) \tag{3.14}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operation and γ is a learnable scalar parameter initialized to a small value, which allows the model to adaptively learn the appropriate background strength directly from data. The embedding sequence thus becomes:

$$\mathbf{S}'' = \{\boldsymbol{e}''_0, \tilde{\boldsymbol{e}}_{\text{bg}}, \boldsymbol{e}''_1, \boldsymbol{e}''_2, \ldots\} \tag{3.15}$$

In the **Support Branch**, attention similarities are computed through Scaled Dot-Product Attention (SDPA) proposed by Vaswani et al. [27] across embeddings in the augmented sequence $\mathbf{S}'$:

$$\mathbf{A}_s = \text{softmax}\left(\frac{\mathbf{Q}_{(s)}(\mathbf{K}_{(s)})^\intercal}{\sqrt{d}}\right) \tag{3.16}$$

where d is the dimension of the feature embeddings and both queries $\mathbf{Q}_{(s)}$ and keys $\mathbf{K}_{(s)}$ are the augmented embedding sequence $\mathbf{S}'$, enabling each embedding to symmetrically interact

and mutually assess similarity. Each row of $\mathbf{A}_s$ indicates the attention distribution of a feature embedding (or background embedding) over all embeddings, guiding the aggregation of refined support features:

$$\mathbf{F}_{\text{out}}^{(s)} = \mathbf{A}_s \mathbf{S}'' \tag{3.17}$$

The background embedding $\boldsymbol{e}_{\text{bg}}$ competes to attract ambiguous background-like features. Due to the soft-scaling strategy combined with a gradient-isolated operation, background features are effectively suppressed, preventing them from dominating the aggregated object representations, thus maintaining high-quality foreground embeddings.

The query encoder employs a parallel FFR mechanism with a slightly modified input structure. Specifically, the module accepts two inputs: the support set sequence $\mathbf{S}_{\text{refined}}$ refined by the support encoder branch, and query patch features $\mathbf{X}_q$. The objective is to refine query patches by selectively aggregating evidence from class prototypes, effectively suppressing background signals.

In the **Query Branch**, each query patch embedding is first projected into the shared prototype space via a linear transformation $\mathbf{W}_3$, yielding $\mathbf{X}'_q = \mathbf{W}_3 \mathbf{X}_q$. The support set sequence $\mathbf{S}_{\text{refined}}$ is processed identically as in the support branch, producing two variants: $\mathbf{S}'_{\text{refined}}$ with the background placeholder replaced by $\boldsymbol{e}_{\text{bg}}$, and $\mathbf{S}''_{\text{refined}}$ with the background placeholder replaced by $\tilde{\boldsymbol{e}}_{\text{bg}}$. The attention matrix between the query patches and the augmented support set sequence $\mathbf{S}'_{\text{refined}}$ is then computed through SDPA as:

$$\mathbf{A}_q = \text{softmax}\left(\frac{\mathbf{X}'_q (\mathbf{S}'_{\text{refined}})^\intercal}{\sqrt{d}}\right) \tag{3.18}$$

The refined class-specific features are then aggregated by:

$$\mathbf{F}_{\text{out}}^{(q)} = \mathbf{A}_{\text{q}} \mathbf{S}''_{\text{refined}} \tag{3.19}$$

The soft-scaled, gradient-isolated background embedding employed in both branches ensures that background-related embeddings participate in the attention competition without propagating gradients through their value pathways, thus protecting the quality

of aggregated foreground representations. Consequently, even query patches visually resembling backgrounds predominantly attend to the learnable background embedding $\boldsymbol{e}_{\text{bg}}$, effectively emphasizing foreground categories and suppressing background interference.

Finally, to retain global contextual information, refined class-specific features $\mathbf{F}_{\text{out}}^{(q)}$ are concatenated with the original query patch features $\mathbf{X}_q$, followed by processing through a lightweight fusion network:

$$\mathbf{F}_{\text{final}} = \text{FFN} \circ \text{Conv}(\mathbf{X}_q \oplus \mathbf{F}_{\text{out}}^{(q)}) \tag{3.20}$$

where FFN provides nonlinear transformations, and Conv adjusts channel dimensions. This integration strategy ensures that refined query embeddings maintain essential spatial and semantic cues while enhancing object-background discriminability, delivering robust inputs for subsequent detection stages.

Explicit Background Supervision (EBS)

While the attention-based competition implicitly shapes the learnable background embedding $\boldsymbol{e}_{\text{bg}}$ through similarity-driven gradients, we further reinforce its training by introducing an Explicit Background Supervision (EBS) mechanism at the detection head. This explicit supervision directly links $\boldsymbol{e}_{\text{bg}}$ to the classification output space, ensuring that it actively captures features that the detector should reject as background.

Specifically, we introduce background placeholders into the support set sequence to explicitly represent background regions. When the number of target classes K is smaller than the sequence length N, the remaining positions are systematically filled by these background placeholders, each associated with the shared embedding $\boldsymbol{e}_{\text{bg}}$. During the forward pass, each placeholder position generates a corresponding background prediction probability in the classification output, explicitly representing the detector's confidence that a query patch belongs to background rather than any foreground class.

During training, these background prediction probabilities are directly supervised in parallel with foreground class predictions. Proposals matched to ground-truth objects

are encouraged to suppress their background prediction probabilities, emphasizing clear separation between object and background. Conversely, unmatched proposals, representing false positives or ambiguous background regions, receive strong supervisory signals to activate the background predictions. This explicit supervision generates direct and robust gradients to the learnable background embedding $\boldsymbol{e}_{\text{bg}}$, guiding it towards capturing consistent and generalizable background features.

Moreover, the EBS mechanism synergistically complements the previously introduced Wavelet-Semantic Fusion Attention module. The fusion process enhances fine-grained details such as edges and textures, which could inadvertently amplify irrelevant background features if unconstrained. Through explicit background predictions aligned with the background placeholders, the EBS mechanism constrains these enhancements, ensuring they strengthen foreground object boundaries rather than aggravating background confusion.

Chapter 4

Experiment

4.1 Datasets and Metrics

4.1.1 Datasets

FSOD experiments are typically conducted on standard benchmark datasets that have been adapted for the few-shot scenario [1]. In this work, we evaluate our approach on the PASCAL VOC [62] and MS COCO [63] datasets. We follow the same dataset splits and experimental setup commonly used in FSOD research to ensure a fair comparison, including identical base/novel class partitions and number of shots per class.

PASCAL VOC Dataset The PASCAL Visual Object Classes (VOC) dataset is a widely used benchmark for object detection, originally comprising 20 object categories including person, vehicle, animal, and household objects [62]. Standard detection models are typically trained on the combined PASCAL VOC 2007 and 2012 train/val sets (approximately 16k images) and evaluated on the PASCAL VOC 2007 test set (4952 images), with all instances annotated with bounding boxes across all 20 classes. In the few-shot setting, initially proposed by Meta R-CNN [7] and subsequently adopted and expanded by later studies [15, 55, 1], the 20 object classes of PASCAL VOC are partitioned into a set of base classes and

novel classes. Following previous works [7, 15], three different splits of base/novel class configurations (often called Split 1, Split 2, and Split 3) are utilized to provide diverse evaluation scenarios [1]. In each split, a distinct subset of 5 classes is selected as novel, with the remaining 15 classes serving as base classes, as detailed in Table 4.1.

Novel Split	Novel Categories
1	bird, bus, cow, motorbike, sofa
2	aeroplane, bottle, cow, horse, sofa
3	boat, cat, motorbike, sheep, sofa

Table 4.1: Novel class splits for FSOD on the PASCAL VOC dataset.

Microsoft COCO Dataset The Microsoft COCO (Common Objects in Context) dataset is a large-scale object detection benchmark with 80 object categories ranging from everyday objects to animals and people. It contains 118k training images and 5k validation images, with a much greater diversity and number of object instances compared to PASCAL VOC [63]. Adapting MS COCO for FSOD requires defining base and novel classes in a similar manner. Following common practice [7, 15, 1], we perform a fixed split of MS COCO's 80 classes into 60 base classes and 20 novel classes, where the novel classes are chosen to be those that overlap with the PASCAL VOC categories. This choice provides challenging novel classes (e.g., person, car, dog, etc.,), while enabling direct comparison with FSOD methods evaluated on PASCAL VOC. The remaining 60 classes (non-VOC categories) are treated as base classes.

4.1.2 Evaluation Metrics

We evaluate detection performance using the standard metrics employed in PASCAL VOC and MS COCO benchmarks, ensuring direct comparability with existing FSOD methods. In essence, we measure Average Precision (AP) for object detection and summarize results as mean Average Precision (mAP) over classes.

Object detection performance is commonly evaluated by the Intersection over Union (IoU), defined as the ratio of the intersection area to the union area between the predicted box B_{pred} and the ground truth box B_{gt}:

$$\text{IoU} = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|} \in [0, 1] \tag{4.1}$$

An IoU of 1 indicates perfect alignment, whereas 0 indicates no overlap. During evaluation, a predicted box is considered a true positive (TP) when its IoU with a ground-truth box of the same class exceeds the given threshold (e.g., 0.5) and the ground truth remains unmatched by higher-scoring predictions. Each ground-truth object can only be matched to one prediction. If multiple predictions overlap the same object, only the one with highest confidence is designated as a true positive, while the others will be counted as false positives (FP). Conversely, any ground-truth object that remains unmatched by any prediction is counted as a false negative (FN). Additionally, predictions with IoU below the threshold or mismatched class labels are also counted as false positives. Precision is the proportion of positive predictions that are actually correct, and recall is the proportion of actual positives that are detected, defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4.2}$$

While precision measures prediction reliability, it does not account for missed detections. Thus, recall is introduced as a complementary metric, representing the proportion of actual positives that are correctly detected:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4.3}$$

In object detection, precision and recall can be balanced by adjusting the confidence threshold applied to the detector's predictions. Specifically, a higher confidence threshold results in fewer detections, leading to increased precision but reduced recall. Conversely, a lower threshold generates more detections, enhancing recall but potentially decreasing precision. By varying the confidence threshold, we obtain a Precision–Recall (PR) curve for each class, plotting precision against recall. Average Precision (AP) is defined as the

area under the PR curve for a given class. Intuitively, AP summarizes the detector's precision-recall performance with a single number, integrating over all recall levels. In practice, the area under the PR curve is computed via a discretized summation.

For a detector evaluated on multiple classes, we report the mean Average Precision (mAP), which is simply the mean of the AP values across all target classes. The IoU threshold has a significant impact on the AP value, so detection benchmarks often report AP at various IoU thresholds to evaluate performance under different localization requirements. The mean Average Precision at IoU threshold 0.5 is denoted as mAP@0.5. This threshold is relatively permissive, allowing predictions with partial overlap to qualify as true positives. To assess more precise localization, detectors are also evaluated at higher IoU thresholds such as 0.75, denoted as mAP@0.75. A higher threshold like 0.75 demands much tighter overlap between the predicted box and the ground truth, so AP@75 is typically lower than AP@50 for a given model, but it reflects the model's ability to localize objects accurately. In the MS COCO protocol, mAP@[0.5:0.95] is computed by averaging AP across 10 IoU thresholds from 0.5 to 0.95 at intervals of 0.05. In effect, this metric averages the AP over a range of IoU criteria, providing a single score that rewards detectors capable of both detecting objects (even with some localization error) and precisely localizing them.

Overall, by adhering to these dataset configurations and evaluation metrics, our experimental methodology aligns with previous studies. This ensures that the datasets and metrics for our experiments are consistent, rigorous, and allow for meaningful comparison against state-of-the-art results in the field. The following sections will detail our experimental results under this setup.

4.2 Main Results

4.2.1 Implementation Details

We employ a two-stage training strategy that is commonly adopted in previous FSOD approaches [7, 15, 64, 22]. In the first stage (base training), the model is pretrained on a large set of base classes with abundant annotated examples, thereby learning generalizable

feature representations for object detection. In the second stage (few-shot fine-tuning), we train the model on both base and novel classes with limited training samples per class. The K-shot detection setting restricts training to K annotated examples per novel category during fine-tuning phase. This two-stage strategy allows the detector to transfer knowledge gained from the base classes to novel classes, effectively addressing the data scarcity in the novel set. By maintaining a balanced training set of base and novel data in the fine-tuning stage, the model avoids severe class imbalance and catastrophic forgetting of base-class knowledge, thus ensuring it can detect both previously seen and new object categories in a unified manner.

For data preprocessing, we follow standard object detection protocols. During training, we apply standard augmentations including random horizontal flips and color jittering ($p = 0.33$). Input images are randomly resized with the shorter side sampled from 480 to 800 pixels at 32-pixel intervals, while preserving aspect ratio (longer side capped at 1152 pixels). During inference, images are resized to a fixed scale of 800 pixels on the shorter side (capped at 1152 on the longer side). All images are normalized using ImageNet mean and standard deviation.

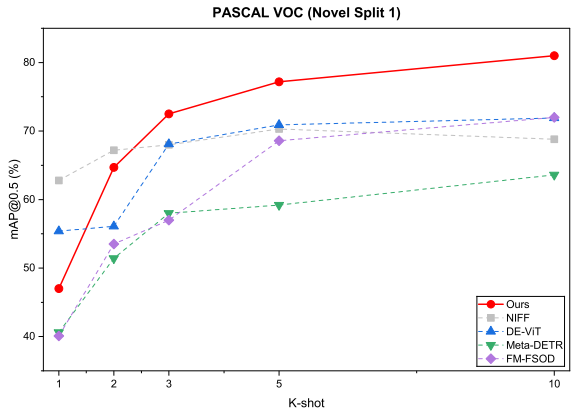
All experiments performed on identical hardware with four NVIDIA RTX 4090 GPUs (24 GB memory each) running Ubuntu 24.04 LTS with the PyTorch framework. We adopt distributed data parallel training across 4 GPUs to handle the computational demands of meta-learning episodes and the DETR transformer architecture. For the specific hyperparameter configuration, we optimize the model using the AdamW optimizer [65] with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}, and set the batch size to 8 images. Following prior work [15], the base training stage runs for 50 epochs on PASCAL VOC base classes and 25 epochs on MS COCO base classes. The learning rate is multiplied by 0.1 at epoch 45 for PASCAL VOC and epoch 20 for MS COCO to refine the model as training stabilizes. Throughout training, we also employ common best practices such as gradient clipping and learning rate warm-up to ensure stable optimization. These implementation details and hyperparameter choices are aligned with those in recent FSOD works to ensure a fair comparison.

Method	Novel Split 1					Novel Split 2					Novel Split 3					Avg
	1	2	3	5	10	1	2	3	5	10	1	2	3	5	10	
Meta R-CNN [7]	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	31.1
TFA [22]	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
FSCE [41]	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5	46.6
DeFRCN [23]	53.6	57.5	61.5	64.1	60.8	30.1	38.1	47.0	53.3	47.9	48.4	50.9	52.3	54.9	57.4	51.8
FCT [50]	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	50.9
FSDetView [64]	24.2	35.3	42.2	49.1	57.4	21.6	24.6	31.9	37.0	45.7	21.2	30.0	37.2	43.8	49.6	36.7
NIFF [66]	**62.8**	**67.2**	68.0	70.3	68.8	38.4	42.9	54.0	56.4	54.0	56.4	62.1	61.2	64.1	63.9	59.3
DiGeo [67]	37.9	39.4	48.5	58.6	61.5	26.6	28.9	41.9	42.1	49.1	30.4	40.1	46.9	52.7	54.7	43.9
FM-FSOD [26]	40.1	53.5	57.0	68.6	72.0	33.1	36.3	48.8	54.8	64.7	39.2	50.2	55.7	63.4	68.1	53.7
DE-ViT [55]	55.4	56.1	68.1	70.9	71.9	**43.0**	39.3	58.1	61.6	63.1	**58.2**	**64.0**	61.3	64.2	67.3	60.2
FSRW [60]	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	28.4
Meta-DETR [15]	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	50.2
Ours	47.0	64.7	**72.5**	**77.2**	**81.0**	42.6	**57.3**	**66.4**	**68.1**	**75.6**	40.1	58.9	**66.7**	**69.8**	**75.8**	**64.2**

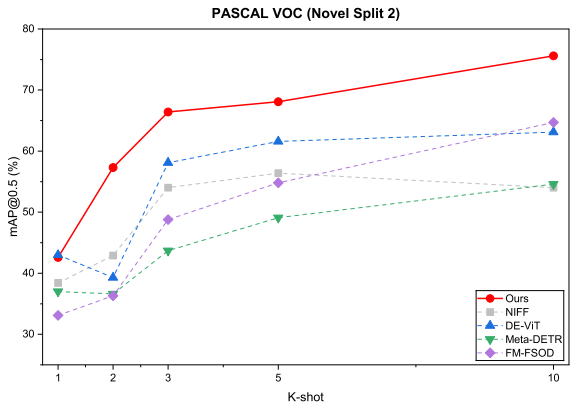
Table 4.2: Few-shot detection performance (mAP@0.5) on the PASCAL VOC dataset. Methods are listed chronologically according to their publication dates. We compare our method against state-of-the-art approaches across three novel class splits, each evaluated with 1, 2, 3, 5, and 10-shot scenarios. The rightmost column reports the average performance across all splits and shot settings. Results in **Bold** indicate the best performance per configuration, and results with underline denote the second-best performance. Our method achieves superior results across most FSOD approaches.

Method	Backbone	10-shot			30-shot		
		mAP@[0.5:0.95]	mAP@0.5	mAP@0.75	mAP@[0.5:0.95]	mAP@0.5	mAP@0.75
Meta R-CNN [7]	ResNet-101	8.7	19.1	6.6	12.4	25.3	10.8
TFA [22]	ResNet-101	9.1	17.1	8.8	12.1	22.0	12.0
FSCE [41]	ResNet-101	11.1	23.0	9.8	15.3	29.0	14.2
DeFRCN [23]	ResNet-101	16.8	-	-	21.1	-	-
FCT [50]	PVTv2	17.1	30.2	17.0	21.4	35.5	22.1
FSDetView [64]	ResNet-101	10.3	25.1	6.1	14.2	31.4	10.3
NIFF [66]	ResNet-101	18.8	-	-	20.9	-	-
DiGeo [67]	ResNet-101	10.3	18.7	9.9	14.2	26.2	14.8
DE-ViT [55]	ViT-L/14	**34.0**	**52.9**	**37.0**	34.0	53.0	**37.2**
FSRW [60]	DarkNet-19	5.6	12.3	4.6	9.1	19.0	7.6
Meta-DETR [15]	ResNet-101	19.0	30.5	19.7	22.2	35.0	22.8
Ours	ViT-L/14	32.1	51.2	33.7	**35.5**	**56.7**	**37.1**

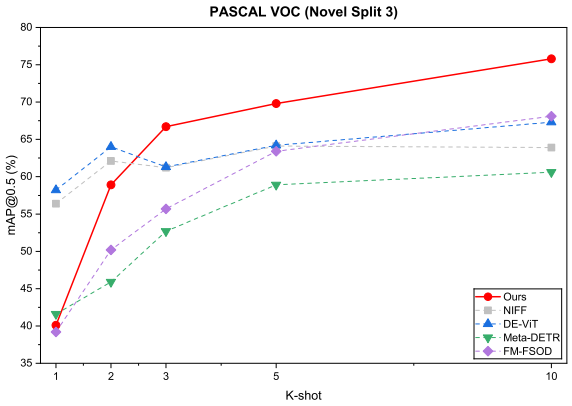
Table 4.3: Few-shot detection performance on novel categories of the COCO dataset. Methods are listed chronologically according to their publication dates. We compare our method with state-of-the-art approaches under 10-shot and 30-shot scenarios, reporting results using mAP@[0.5:0.95], mAP@0.5, and mAP@0.75 metrics. The backbone architectures used by each method are also provided for reference. Results in **Bold** indicates the best performance per evaluation setting. Our method consistently achieves competitive results compared to recent methods across all configurations.



(a)



(b)



(c)

Figure 4.1: Quantitative comparison of Few-shot detection performance (mAP@0.5) on the PASCAL VOC dataset under three standard novel class splits, including (a) Novel Split 1, (b) Novel Split 2, and (c) Novel Split 3. Our method is compared against representative state-of-the-art FSOD approaches, demonstrating consistent performance gains across different splits and shot settings.

4.2.2 Performance Comparison

(a) Ground Truth (b) Ours

Figure 4.2: Qualitative results of our proposed method on the MS COCO dataset (10-shot scenario). Detected objects are highlighted with bounding boxes. All highlighted objects belong to novel categories (*car*, *bus*, *person*).

Figure 4.2 presents qualitative comparisons between the ground truth (GT) annotations and the predictions from our proposed method. The Figure 4.2a shows the GT annotations accurately identify the primary bus, as well as the smaller, distant car and multiple pedestrians located in the background. In the Figure 4.2b, our method pretrained on MS COCO under 10-shot setting successfully detects the primary objects, including the bus, the distant car, and the pedestrian in the foreground, though with slight bounding box deviations compared to the GT annotations. Notably, our approach identifies an additional distant car, partially obscured by surrounding objects, which was absent in the GT annotations, demonstrating robust generalization capabilities even under limited training conditions.

However, under the constraints of FSOD with only 10 annotated examples per novel class, several detection challenges emerge: (1) multiple pedestrians, nearly blending into the background, remain undetected due to their small scale, insufficient visual distinction from the surroundings, and the limited discriminative cues provided by their clothing color; (2) the bus driver is also missed, likely because the windshield reflects textures from the sky, introducing visual interference that complicates clear identification and reduces

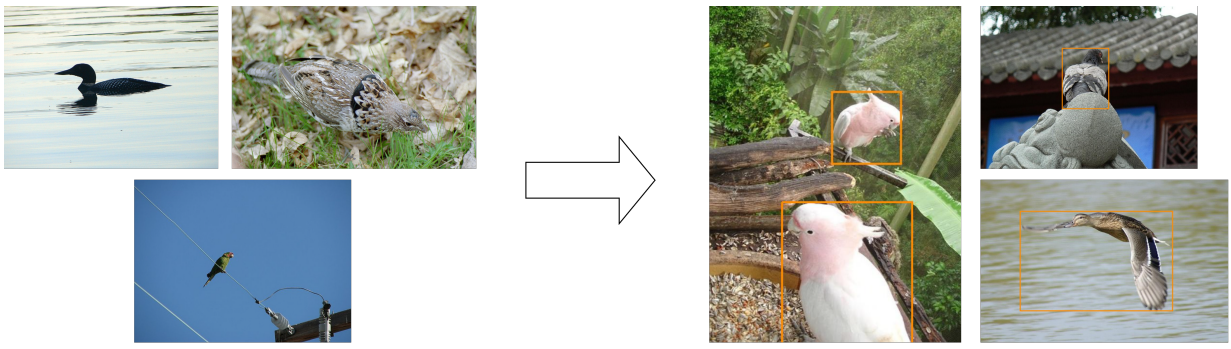


Figure 4.3: Qualitative results of our proposed method on the PASCAL VOC dataset (Novel Split 1, 3-shot scenario). The left three images represent the support examples for the novel class "*bird*", previously unseen by the model during base training. The images on the right demonstrate our model's detection results on query images, highlighting its capability to accurately detect novel instances after learning from only three support examples. Predicted bounding boxes are shown in orange.

the distinction between the driver and background regions. The identified shortcomings highlight common challenges in few-shot scenarios and underscore the importance of enhancing support set diversity and discriminative feature representation to better handle subtle visual distinctions.

These qualitative observations, along with additional visualization on PASCAL VOC presented in Figure 4.3, motivate a deeper quantitative analysis. We compare our method with prior work on PASCAL VOC and MS COCO under the standard few-shot protocols. Results are summarized in Figure 4.1 and Tables 4.2 and 4.3. Overall, our proposed method significantly enhances the performance of the Meta-DETR baseline, with the improvements becoming more evident as the number of shots increases.

On PASCAL VOC (mAP@0.5), our approach achieves an average performance of 64.2 mAP across all novel splits and shot configurations, yielding a significant +14.0 improvement over the baseline. This performance advantage expands with more training examples, smoothly increasing from the 1/2-shot scenarios to the 3/5/10-shot settings, and exceeding 80 mAP in the most data-rich configuration. Compared to DE-ViT, the state-

of-the-art two-stage R-CNN-based approach, our approach exhibits a slight performance gap in the extreme low-shot scenarios (1-2 shots), but achieves consistent gains from the 3-shot regime onward across all splits, indicating that our method scales more effectively once adequate intra-class variation is captured.

On MS COCO, similar performance trends are observed across all IoU evaluation thresholds. Our method substantially surpasses the baseline across both evaluation settings, achieving gains of +13.1 mAP@[0.5:0.95] in the 10-shot setting, and +13.3 in the 30-shot setting. Importantly, these performance gains persist under stricter localization requirements (mAP@0.75), demonstrating robust detection quality. Against the two-stage SOTA approach, we observe slightly lower mAP@[0.5:0.95] at 10-shot but achieve superior performance at 30-shot, while achieving near-equivalent performance at high-IoU thresholds.

The performance crossover at 30-shot on MS COCO highlights fundamental differences in how our method and two-stage R-CNN-based methods, such as the SOTA method DE-ViT, leverage additional support data. This advantage arises from three primary factors. First, our end-to-end optimization jointly refines feature extraction, prototype learning, and detection heads, whereas DE-ViT separately optimizes its RPN and detection heads, potentially leading to suboptimal feature utilization. Second, unlike DE-ViT which faces performance saturates between 10 and 30 shots (34.0 $\rightarrow$ 34.0 mAP@[0.5:0.95]), our approach continues to improve substantially (32.1 $\rightarrow$ 35.5, gain of +3.4), demonstrating superior sample efficiency in capturing intra-class variation and reducing prototype noise. Third, our set prediction mechanism naturally aggregates information across support examples via attention-based refinement, allowing it to construct more robust class representations that better handle the diverse object scales and contexts present in COCO's challenging scenarios. Notably, our larger improvement at mAP@0.5 (53.0 $\rightarrow$ 56.7, +3.7) compared to DE-ViT (37.1 $\rightarrow$ 37.2, +0.1) indicates our advantage primarily stems from improved classification confidence rather than localization precision, which benefits from the higher quality of learned prototypes.

Split	Baseline	WSFA	LBP	Novel mAP@0.5					Avg.	Δ
				1	2	3	5	10		
1	✓			39.8	50.2	56.2	57.6	61.5	53.0	(0)
	✓	✓		44.3	61.7	70.1	75.4	79.4	66.2	+13.2
	✓	✓	✓	47.0	64.7	72.5	77.2	81.0	68.5	+2.4
2	✓			35.4	35.2	42.5	47.8	53.2	42.8	(0)
	✓	✓		40.1	54.4	64.8	65.6	73.8	59.7	+16.9
	✓	✓	✓	42.6	57.3	66.4	68.1	75.6	62.0	+2.3
3	✓			40.5	44.6	51.2	57.4	59.2	50.6	(0)
	✓	✓		40.8	56.3	64.5	67.8	73.7	60.6	+10.0
	✓	✓	✓	40.1	58.9	66.7	69.8	75.8	62.3	+1.7

Table 4.4: Ablation study of few-shot detection performance (mAP@0.5) on the PASCAL VOC dataset.

4.3 Ablation Study

To evaluate the FSOD performance of our proposed method, we conduct ablation studies on the two proposed components of our method including the Wavelet–Semantic Fusion Attention (WSFA) and the Learnable Background Prototype (LBP). This analysis quantifies their individual and combined contributions to our method. Unless specified otherwise, results represent novel-class mAP@0.5 on PASCAL VOC under the standard evaluation protocols as mentioned before. Comprehensive ablation results are summarized in Table 4.4.

Our baseline detector follows a Meta-DETR style set prediction architecture without WSFA or LBP, achieving split-wise averages of 53.0, 42.8, and 50.6 mAP on Novel Splits 1–3, respectively, for an overall average of 48.8 mAP. Then, incorporating WSFA into this baseline yields substantial and consistent improvements of +13.2, +16.9, and +10.0 mAP across the three splits. Notably, these improvements scale significantly with the number of available shots: improvements are modest at 1-shot (+3.2 mAP on average), performance increases sharply from 2 shots onward, achieving gains of between +14 and +18 mAP. This scaling pattern validates WSFA's design rationale by integrating frequency-domain cues with

high-level semantic features to enhance support–query matching and effectively leverage intra-class variability provided by additional support examples. The largest improvements occur consistently within the 2–10 shots, with Novel Split 2 demonstrating particularly strong gains, averaging around +20 mAP across these shot settings. These benefits are particularly evident for texture-sensitive novel classes, such as *bird*, *cat*, *sheep*, and *motorbike*, which strongly depend on fine-grained visual details.

Then, integrating the LBP module with WSFA provides an additional and consistent boost of approximately +2 mAP across all splits. These improvements peak in the lower-shot range (1–3 shots) and slightly diminish at 5-10 shots. This behaviour suggests that LBP primarily addresses early-stage false positives by introducing a dedicated, learnable background prototype to prevent background regions from incorrectly matching foreground classes. Explicitly modeling the background enhances precision, particularly at moderate recall levels, complementing the substantial recall gains from WSFA. One minor exception is observed at 1-shot on Novel Split 3, where LBP temporarily decreases performance, but this effect completely disappears at higher shot settings. We believe this occurs due to the severe data scarcity in the 1-shot scenario, where both foreground prototypes and the newly introduced background prototype remain inadequately trained, leading to ambiguous boundaries within the feature space. This ambiguity can cause genuine foreground instances to be misclassified as background, thus lowering recall. As more training samples become available, prototype representations stabilize, naturally resolving this issue.

Taken together, the full model combining WSFA and LBP module achieves 64.3 mAP overall, compared to 48.8 for the baseline and 62.2 for WSFA alone, representing a cumulative gain of +15.5 mAP over the baseline. The shot-wise breakdown further emphasizes the scaling behavior. Relative to the baseline, the full model improves by +4.7, +17.0, +18.6, +17.4, and +19.5 mAP at 1, 2, 3, 5, and 10 shots, respectively. This pattern demonstrates that performance increases steadily with additional shots where more support examples allow WSFA to capture richer intra-class variations and LBP to more reliably separate foreground from background.

In summary, the ablation study clearly validates the individual effectiveness and complementary nature of the proposed WSFA and LBP modules in our FSOD approach.

Specifically, integrating WSFA consistently results in substantial performance improvements across all few-shot scenarios, particularly by leveraging frequency-domain cues to capture richer intra-class variability. Further incorporating the LBP module leads to additional consistent gains, primarily by reducing early-stage false positives through explicit background modeling. Notably, the performance gains exhibit clear scalability with increased training shots, highlighting the proposed model's robust capacity for learning discriminative object representations, even from limited data.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In this work, we explored advanced methodologies to overcome critical challenges in FSOD, specifically addressing the difficulty in generating discriminative features from limited supervision and reducing confusion between foreground objects and visually similar background areas. We propose a single-stage, transformer-based FSOD framework that integrates two key contributions: Wavelet–Semantic Fusion Attention and Learnable Background Prototype. Wavelet–Semantic Fusion Attention is a cross-modal mechanism that refines support-query interactions by combining high-level semantic information obtained from a large Vision Transformer with detailed frequency-domain structure captured by wavelet transforms. Meanwhile, the Learnable Background Prototype explicitly models background as a first-class entity, improving separation between objects and challenging negative regions. These enhancements maintain the streamlined, end-to-end training of DETR-style set prediction while strengthening both representation quality and decision boundaries under few-shot scenarios.

Our empirical evaluations validate these contributions. Experiments on PASCAL VOC across all standard splits and shot settings demonstrate substantial improvements in novel-class detection accuracy. On the more complex MS COCO dataset, our method consistently

outperforms previous works and remains competitive with state-of-the-art two-stage models. Nevertheless, like any method, our approach has limitations. The reliance on a large ViT backbone provides strong transferable features but increases computational cost and memory footprint. Additionally, the gains are most pronounced at looser localization thresholds, indicating that while features become more discriminative, bounding-box regression remains challenging in the lowest-shot settings. Finally, our training protocol still requires a reasonably sized base set, while performance degrades significantly in extremely low-data scenarios.

Despite the identified limitations, this work represents an important step toward developing more adaptable and data-efficient FSOD architecture. The capability to recognize novel object categories from only a few examples is essential for applying vision systems in domains where exhaustive annotation is impractical, such as medical imaging, rare species monitoring, or rapidly changing industrial inspection tasks. Our results demonstrate that careful architectural design, which integrates multi-modal feature fusion and explicit background modeling, can substantially enhance detection performance within a unified framework. We hope these insights will motivate further exploration into flexible, generalizable detection methods. Ultimately, bridging the gap between data-intensive deep learning and human-like rapid adaptation from limited examples remains a central challenge in computer vision.

5.2 Future Work

Looking ahead, we plan to extend our framework to address existing limitations and explore new research directions.

Improving localization performance in extreme few-shot scenarios is a key priority. Incorporating structure-aware box refinement techniques, such as geometry-constrained decoders or cascade-style detection heads, could significantly boost accuracy at stricter IoU thresholds. Another important direction is to enhance the adaptivity and efficiency of the frequency-semantic fusion mechanism by exploring dynamic wavelet selection or learned

spectral dictionaries, potentially yielding richer and more robust feature representations from minimal examples.

Additionally, extending our approach beyond static images offers significant potential. Specifically, investigating the efficacy of frequency-semantic interactions and explicit background modeling in video-based FSOD, 3D scene understanding, and instance segmentation represents a promising research direction. Cross-modal evaluation will elucidate the generalizability of our approach and provide insights for designing robust, data-efficient vision systems.

Through these future directions, we hope to push the boundaries of FSOD towards greater flexibility, robustness, and real-world applicability.

References

[1] Zhimeng Xin, Shiming Chen, Tianxu Wu, Yuanjie Shao, Weiping Ding, and Xinge You. "Few-shot object detection: Research advances and challenges". In: *Information Fusion* 107 (2024), page 102307. DOI: 10.1016/j.inffus.2024.102307.

[2] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. "Self-supervision with Superpixels: Training Few-Shot Medical Image Segmentation Without Annotation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2020, pages 762–780. DOI: 10.1007/978-3-030-58526-6_45.

[3] Song Tang, Shaxu Yan, Xiaozhi Qi, Jianxin Gao, Mao Ye, Jianwei Zhang, and Xiatian Zhu. "Few-shot medical image segmentation with high-fidelity prototypes". In: *Medical Image Analysis* 100 (2025), page 103412. DOI: 10.1016/j.media.2024.103412.

[4] Zihan Wang, Bowen Li, Chen Wang, and Sebastian Scherer. "AirShot: Efficient Few-Shot Detection for Autonomous Exploration". In: *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2024, pages 11654–11661. DOI: 10.1109/IROS58592.2024.10801738.

[5] Erjun Sun, Di Zhou, Yan Tian, Zhaocheng Xu, and Xun Wang. "Transformer-based few-shot object detection in traffic scenarios". In: *Applied Intelligence* 54.1 (2024), pages 947–958. DOI: 10.1007/s10489-023-05245-5.

[6] Jamuna S Murthy, Dhanashekar Kandaswamy, and Wen-Cheng Lai. "Towards Secure Video Surveillance: A Few-Shot Spatiotemporal Perception Transformer for Unseen Behavioral Anomalies". In: *2025 IEEE International Conference on Advanced Visual and Signal-Based Systems (AVSS)*. 2025, pages 1–6. DOI: 10.1109/AVSS65446.2025.11149904.

[7] Xiaopeng Yan, Ziliang Chen, Anni Xu, Xiaoxi Wang, Xiaodan Liang, and Liang Lin. "Meta R-CNN: Towards General Solver for Instance-Level Low-Shot Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pages 9577–9586. DOI: 10.1109/ICCV.2019.00967.

[8] Anish Madan, Neehar Peri, Shu Kong, and Deva Ramanan. "Revisiting Few-Shot Object Detection with Vision-Language Models". In: *Advances in Neural Information Processing Systems*. Volume 37. Curran Associates, Inc., 2024, pages 19547–19560. DOI: 10.52202/079017-0617.

[9] Jian Zhang, Chenglong Zhao, Bingbing Ni, Minghao Xu, and Xiaokang Yang. "Variational Few-Shot Learning". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pages 1685–1694. DOI: 10.1109/ICCV.2019.00177.

[10] Shan Zhang, Yao Ni, Jinhao Du, Yuan Xue, Philip Torr, Piotr Koniusz, and Anton van den Hengel. "Open-World Objectness Modeling Unifies Novel Object Detection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2025, pages 30332–30342. DOI: 10.1109/CVPR52734.2025.02824.

[11] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. "Camouflaged Object Detection With Feature Decomposition and Edge Reconstruction". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pages 22046–22055. DOI: 10.1109/CVPR52729.2023.02111.

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2014, pages 580–587. DOI: 10.1109/CVPR.2014.81.

[13] Ross Girshick. "Fast R-CNN". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015, pages 1440–1448. DOI: 10.1109/ICCV.2015.169.

[14] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.6 (2017), pages 1137–1149. DOI: 10.1109/TPAMI.2016.2577031.

[15] Gongjie Zhang, Zhipeng Luo, Kaiwen Cui, Shijian Lu, and Eric P. Xing. "Meta-DETR: Image-Level Few-Shot Detection With Inter-Class Correlation Exploitation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.11 (2023), pages 12832–12843. DOI: 10.1109/TPAMI.2022.3195735.

[16] Guangxing Han, Yicheng He, Shiyuan Huang, Jiawei Ma, and Shih-Fu Chang. "Query Adaptive Few-Shot Object Detection With Heterogeneous Graph Convolutional Networks". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pages 3263–3272. DOI: 10.1109/ICCV48922.2021.00325.

[17] Yan Ren, Yanling Li, and Adams Wai-Kin Kong. "Adaptive Multi-task Learning for Few-Shot Object Detection". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2025, pages 297–314. DOI: 10.1007/978-3-031-72667-5_17.

[18] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-End Object Detection with Transformers". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2020, pages 213–229. DOI: 10.1007/978-3-030-58452-8_13.

[19] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S. Davis. "Soft-NMS – Improving Object Detection With One Line of Code". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pages 5561–5569. DOI: 10.1109/ICCV.2017.593.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pages 770–778. DOI: 10.1109/CVPR.2016.90.

[21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV].

[22] Xin Wang, Thomas Huang, Joseph Gonzalez, Trevor Darrell, and Fisher Yu. "Frustratingly Simple Few-Shot Object Detection". In: *Proceedings of the 37th International Conference on Machine Learning*. Volume 119. Proceedings of Machine Learning Research. PMLR, 2020, pages 9919–9928.

[23] Limeng Qiao, Yuxuan Zhao, Zhiyuan Li, Xi Qiu, Jianan Wu, and Chi Zhang. "DeFRCN: Decoupled Faster R-CNN for Few-Shot Object Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pages 8681–8690. DOI: 10.1109/ICCV48922.2021.00856.

[24] Zican Zha, Hao Tang, Yunlian Sun, and Jinhui Tang. "Boosting Few-Shot Fine-Grained Recognition With Background Suppression and Foreground Alignment". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.8 (2023), pages 3947–3961. DOI: 10.1109/TCSVT.2023.3236636.

[25] Guangxing Han, Shiyuan Huang, Jiawei Ma, Yicheng He, and Shih-Fu Chang. "Meta Faster R-CNN: Towards Accurate Few-Shot Object Detection with Attentive Feature Alignment". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.1 (2022), pages 780–789. DOI: 10.1609/AAAI.v36i1.19959.

[26] Guangxing Han and Ser-Nam Lim. "Few-Shot Object Detection with Foundation Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pages 28608–28618. DOI: 10.1109/CVPR52733.2024.02703.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Volume 30. Curran Associates, Inc., 2017.

[28] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. "Selective Search for Object Recognition". In: *International Journal of Computer Vision* 104.2 (2013), pages 154–171. DOI: 10.1007/s11263-013-0620-5.

[29] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. "Feature Pyramid Networks for Object Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pages 2117–2125. DOI: 10.1109/CVPR.2017.106.

[30] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z. Li. "Bridging the Gap Between Anchor-Based and Anchor-Free Detection via Adaptive Training Sample Selection". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pages 9759–9768. DOI: 10.1109/CVPR42600.2020.00978.

[31] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. "Object Detection in 20 Years: A Survey". In: *Proceedings of the IEEE* 111.3 (2023), pages 257–276. DOI: 10.1109/JPROC.2023.3238524.

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pages 779–788. DOI: 10.1109/CVPR.2016.91.

[33] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. "SSD: Single Shot MultiBox Detector". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2016, pages 21–37. DOI: 10.1007/978-3-319-46448-0_2.

[34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. "Focal Loss for Dense Object Detection". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017, pages 2980–2988. DOI: 10.1109/ICCV.2017.324.

[35] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. "CenterNet: Keypoint Triplets for Object Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pages 6569–6578. DOI: 10.1109/ICCV.2019.00667.

[36] Hei Law and Jia Deng. "CornerNet: Detecting Objects as Paired Keypoints". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2018, pages 765–781. DOI: 10.1007/978-3-030-01264-9_45.

[37] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. "FCOS: Fully Convolutional One-Stage Object Detection". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pages 9627–9636. DOI: 10.1109/ICCV.2019.00972.

[38] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. "Deformable DETR: Deformable Transformers for End-to-End Object Detection". In: *International Conference on Learning Representations*. 2020.

[39] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "ImageNet: A large-scale hierarchical image database". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2009, pages 248–255. DOI: 10.1109/CVPR.2009.5206848.

[40] Zhaowei Cai and Nuno Vasconcelos. “Cascade R-CNN: Delving Into High Quality Object Detection”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pages 6154–6162. DOI: 10.1109/CVPR.2018.00644.

[41] Bo Sun, Banghuai Li, Shengcai Cai, Ye Yuan, and Chi Zhang. “FSCE: Few-Shot Object Detection via Contrastive Proposal Encoding”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pages 7352–7362. DOI: 10.1109/CVPR46437.2021.00727.

[42] Zihao Wang and Lei Wu. “Theoretical Analysis of the Inductive Biases in Deep Convolutional Networks”. In: *Advances in Neural Information Processing Systems*. Volume 36. Curran Associates, Inc., 2023, pages 74289–74338.

[43] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “ImageNet Classification with Deep Convolutional Neural Networks”. In: *Advances in Neural Information Processing Systems*. Volume 25. Curran Associates, Inc., 2012.

[44] Karen Simonyan and Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2015. arXiv: 1409.1556 [cs.CV].

[45] Joseph Redmon and Ali Farhadi. “YOLO9000: Better, Faster, Stronger”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pages 7263–7271. DOI: 10.1109/CVPR.2017.690.

[46] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning*. PMLR, 2019, pages 6105–6114.

[47] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. “Path Aggregation Network for Instance Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2018, pages 8759–8768. DOI: 10.1109/CVPR.2018.00913.

[48] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. "Exploring Plain Vision Transformer Backbones for Object Detection". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer Nature Switzerland, 2022, pages 280–296. DOI: 10.1007/978-3-031-20077-9_17.

[49] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pages 10012–10022. DOI: 10.1109/ICCV48922.2021.00986.

[50] Guangxing Han, Jiawei Ma, Shiyuan Huang, Long Chen, and Shih-Fu Chang. "Few-Shot Object Detection With Fully Cross-Transformer". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pages 5321–5330. DOI: 10.1109/CVPR52688.2022.00525.

[51] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. *DINOv2: Learning Robust Visual Features without Supervision*. 2024. arXiv: 2304.07193 [cs.CV].

[52] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. "Image BERT Pre-training with Online Tokenizer". In: *International Conference on Learning Representations*. 2021.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021, pages 8748–8763.

[54] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. "Masked Autoencoders Are Scalable Vision Learners". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pages 16000–16009. DOI: 10.1109/CVPR52688.2022.01553.

[55] Xinyu Zhang, Yuhan Liu, Yuting Wang, and Abdeslam Boularias. "Detect Everything with Few Examples". In: *Proceedings of The 8th Conference on Robot Learning*. PMLR, 2025, pages 3986–4004.

[56] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. *SAM 2: Segment Anything in Images and Videos*. 2024. arXiv: 2408.00714 [cs.CV].

[57] Zeyu Shangguan, Daniel Seita, and Mohammad Rostami. "Cross-Domain Multi-Modal Few-Shot Object Detection via Rich Text". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2025, pages 6570–6580. DOI: 10.1109/WACV61041.2025.00640.

[58] Yaqing Wang, Quanming Yao, James T. Kwok, and Lionel M. Ni. "Generalizing from a Few Examples: A Survey on Few-shot Learning". In: *ACM Computing Surveys* 53.3 (2021), pages 1–34. DOI: 10.1145/3386252.

[59] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2017, pages 1126–1135.

[60] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. "Few-Shot Object Detection via Feature Reweighting". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pages 8420–8429. DOI: 10.1109/ICCV.2019.00851.

[61] Ingrid Daubechies. “Orthonormal bases of compactly supported wavelets”. In: *Communications on Pure and Applied Mathematics* 41.7 (1988), pages 909–996. DOI: 10.1002/CPA.3160410705.

[62] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. “The Pascal Visual Object Classes (VOC) Challenge”. In: *International Journal of Computer Vision* 88.2 (2010), pages 303–338. DOI: 10.1007/s11263-009-0275-4.

[63] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. “Microsoft COCO: Common Objects in Context”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing, 2014, pages 740–755. DOI: 10.1007/978-3-319-10602-1_48.

[64] Yang Xiao, Vincent Lepetit, and Renaud Marlet. “Few-Shot Object Detection and Viewpoint Estimation for Objects in the Wild”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.3 (2023), pages 3090–3106. DOI: 10.1109/TPAMI.2022.3174072.

[65] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG].

[66] Karim Guirguis, Johannes Meier, George Eskandar, Matthias Kayser, Bin Yang, and Jürgen Beyerer. “NIFF: Alleviating Forgetting in Generalized Few-Shot Object Detection via Neural Instance Feature Forging”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pages 24193–24202. DOI: 10.1109/CVPR52729.2023.02317.

[67] Jiawei Ma, Yulei Niu, Jincheng Xu, Shiyuan Huang, Guangxing Han, and Shih-Fu Chang. “DiGeo: Discriminative Geometry-Aware Learning for Generalized Few-Shot Object Detection”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pages 3208–3218. DOI: 10.1109/CVPR52729.2023.00313.