Titre: Title:	From Temporal Coherence to Cross-Modal Intelligence: A Modular Framework for Video Object Detection
Auteur: Author:	Noreen Anwar
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Anwar, N. (2025). From Temporal Coherence to Cross-Modal Intelligence: A Modular Framework for Video Object Detection [Ph.D. thesis, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/70225/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/70225/
Directeurs de recherche: Advisors:	Guillaume-Alexandre Bilodeau, & Wassim Bouachir
Programme: Program:	Génie informatique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

From Temporal Coherence to Cross-Modal Intelligence: A Modular Framework for Video Object Detection

NOREEN ANWAR
Département de génie informatique et génie logiciel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*
Génie informatique

Novembre 2025

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Cette thèse intitulée :

From Temporal Coherence to Cross-Modal Intelligence: A Modular Framework for Video Object Detection

présentée par **Noreen ANWAR**
en vue de l'obtention du diplôme de *Philosophiæ Doctor*
a été dûment acceptée par le jury d'examen constitué de :

Tarek OULD-BACHIR, président
Guillaume-Alexandre BILODEAU, membre et directeur de recherche
Wassim BOUACHIR, membre et codirecteur de recherche
Lama SÉOUD, membre
Sousso KELOUWANI, membre externe

DEDICATION

To my husband and my parents,
You people are my backbone throughout this journey . . .

Acknowledgements

I would like to express my deepest gratitude to my research supervisor, Mr. Guillaume-Alexandre Bilodeau. His constant guidance, scientific expertise, and unwavering support throughout my doctoral studies have been instrumental in the completion of this thesis. I am sincerely thankful for his trust, availability, and encouragement at every stage of this journey.

I also extend my heartfelt thanks to my co-supervisor, Mr. Wassim Bouachir, for his valuable insights, constructive feedback, and kind mentorship. His contribution helped shape and refine the direction of this research.

I warmly acknowledge the LITIV laboratory for providing a rich and collaborative research environment. I am grateful to all current and past members of the lab with whom I had the pleasure of interacting, sharing ideas, and growing intellectually.

A special word of appreciation goes to my husband, whose patience, encouragement, and emotional support were a source of strength during the most challenging phases of this doctoral journey.

Finally, I would like to thank the members of the jury: **Prof. Tarek Ould-Bachir (President)**, **Prof. Lama Séoud**, **Prof. Sousso Kelouwani (external member from UQTR)**, for accepting to evaluate my work and for their thoughtful comments and suggestions that helped further strengthen this thesis.

RÉSUMÉ

La détection d'objets dans les vidéos est un problème fondamental en vision par ordinateur, avec des applications dont la conduite autonome, la surveillance intelligente, la robotique et l'interaction humain-machine. Malgré les avancées réalisées en détection d'objets sur des images statiques grâce aux réseaux de neurones profonds et aux architectures de type auto-attentive, l'extension de ces capacités aux séquences vidéo demeure un défi de taille. Cela s'explique par la diversité des perturbations visuelles et environnementales, telles que l'occultation d'objets, les mouvements rapides, la présence de distracteurs, les variations d'éclairage et les changements de pose des objets articulés. Dans cette thèse, la robustesse désigne la capacité d'un détecteur à conserver des prédictions précises et temporellement cohérentes malgré ces perturbations, et non seulement sur des images propres et bien cadrées.

Cette thèse propose de nouvelles stratégies et architectures pour améliorer la performance de la détection d'objets dans les vidéos en conditions réelles et non contraintes. L'objectif est de répondre aux principaux défis de la détection en exploitant la cohérence spatio-temporelle, la mémoire visuelle et la compréhension contextuelle multimodale. Nos travaux introduisent à la fois des méthodes de détection complètes et des mécanismes modulaires pouvant être intégrés dans de futurs pipelines de détection.

Dans un premier temps, nous proposons STF (Spatio-Temporal Fusion), une approche de détection qui fusionne l'attention multi-images avec les caractéristiques visuelles extraites trame par trame, permettant une détection robuste malgré l'occultation, le flou de mouvement et les changements de point de vue. STF améliore la continuité temporelle en intégrant des indices issus des images voisines à l'aide de stratégies d'attention adaptatives. Des expériences approfondies menées sur les ensembles de données KITTI [1], Cityscapes [2] et VisDrone [3] montrent que STF améliore la précision et la stabilité temporelle dans des conditions bruitées et dynamiques.

Ensuite, nous avons développé LAQEM (Language-Augmented Query Evolution with Memory), qui intègre des informations sémantiques textuelles dans le processus de génération des requêtes visuelles. LAQEM utilise une mémoire dynamique pour conserver les informations visuelles sémantiques pertinentes au fil du temps, ce qui améliore la capacité du système à détecter des objets rares ou inconnus dans des scènes complexes. Cette méthode comble ainsi l'écart de signification entre les caractéristiques visuelles et les indices linguistiques, favorisant une détection plus généralisable et sémantiquement robuste.

Enfin, nous présentons DAMM (Dual Attention with Multimodal Memory), un modèle qui

réalise une fusion structurée entre le contexte visuel courant et une mémoire multimodale à long terme via un mécanisme d'attention croisée. DAMM gère efficacement les variations d'apparence et les scènes ambiguës en réutilisant les indices historiques et les représentations sémantiques. Il permet une inférence fiable même dans les cas de réapparition après une longue occultation ou une sortie du champ de vision.

Les modèles proposés sont évalués de manière exhaustive sur des jeux de données de référence, et démontrent des performances supérieures en termes de précision moyenne (mAP), de cohérence temporelle et d'efficacité d'inférence. Les résultats expérimentaux confirment que nos nouvelles méthodes fondées sur la mémoire et la multimodalité représentent une avancée significative par rapport aux techniques de pointe actuelles. Par ailleurs, nous montrons que la visibilité partielle suffit souvent à assurer la récupération des objets dans des scénarios complexes, ce qui illustre la robustesse de nos approches face à l'occultation, au flou de mouvement et à la déformation structurelle.

ABSTRACT

Video object detection is a fundamental problem in computer vision with applications in autonomous driving, intelligent surveillance, robotics, and human–machine interaction. Despite the progress made in static image detection using deep learning and transformer-based architectures, extending these capabilities to videos remains a significant challenge. This is due to a variety of environmental and visual perturbations, including object occlusion, rapid motion, distractors, dynamic lighting, and articulated pose changes. In this thesis, robustness refers to the ability of a detector to maintain accurate and temporally consistent predictions under such challenging conditions, rather than only on clean, well-framed images.

This thesis presents novel strategies and architectures to enhance video object detection performance in unconstrained, real-world conditions. The goal is to address the primary detection challenges by leveraging spatio-temporal consistency, visual memory, and multimodal contextual understanding. Our work introduces complete detection frameworks as well as modular mechanisms that can be integrated into future detection pipelines.

First, we propose STF (Spatio-Temporal Fusion), a detection framework that fuses multi-frame attention with frame-wise object features, enabling robust detection under occlusion, motion blur, and viewpoint changes. STF improves temporal continuity by integrating cues from neighbouring frames using adaptive attention strategies. Through extensive experiments on KITTI [1], Cityscapes [2], and VisDrone [3], STF demonstrates improved accuracy and temporal stability under noisy and dynamically changing conditions.

Second, we develop LAQEM (Language-Augmented Query Evolution with Memory), which incorporates text-based semantic priors into the visual query generation process. LAQEM uses dynamic memory to retain relevant visual–semantic information across frames, improving the system's ability to detect unseen or rare objects in challenging scenes. This framework bridges the gap between visual features and linguistic cues, enabling more generalizable and semantically robust detection.

Third, we present DAMM (Dual Attention with Multimodal Memory), a model that performs structured fusion between the current visual context and long-term multimodal memory using cross-attention. DAMM effectively handles appearance variations and ambiguous scenes by reusing historical cues and semantic embeddings. It supports reliable inference even in cases of reappearance after long occlusion or field-of-view exit.

The proposed models are extensively evaluated on standard video benchmarks, showing su-

perior performance in terms of mean Average Precision (mAP), temporal consistency, and inference efficiency. Experimental results confirm that our memory-aware and multimodal detection designs provide a significant advancement over state-of-the-art techniques in realistic settings. Furthermore, we show that partial visibility is sufficient for recovery in many challenging scenarios, demonstrating the strength of our approaches under occlusion, motion blur, and structural deformation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

AI	Artificial Intelligence
CNN	Convolutional Neural Network
DETR	DEtection TRansformer
DL	Deep Learning
IoU	Intersection over Union
LLM	Large Language Model
ML	Machine Learning
MOT	Multiple Object Tracking
NMS	Non-Maximum Suppression
NLP	Natural Language Processing
R-CNN	Region-based Convolutional Neural Network
ReLU	Rectified Linear Unit
STF	Spatio-Temporal Fusion
SAM	Segment Anything Model
LAQEM	Language-Augmented Queries with Evolving Memory
DAMM	Dual-Stream Attention with Multimodal Queries
UAV	Unmanned Aerial Vehicle
VLM	Vision-Language Model
ViT	Vision Transformer
VOD	Video Object Detection
mAP	mean Average Precision
GT	Ground Truth
CLIP	Contrastive Language–Image Pre-training
GPU	Graphics Processing Unit
FPS	Frames Per Second
SOTA	State of the Art

CHAPTER 1 INTRODUCTION

This thesis addresses the problem of robust object detection in videos captured under real-world unconstrained conditions. Video surveillance systems generate massive data streams every day. Manual analysis of recorded or live video is laborious and error-prone: studies show that a human operator's attention declines sharply after only 20 minutes of monitoring, and it is virtually impossible to accurately track more than ten camera feeds simultaneously [25]. To overcome these limitations, automated detection systems are increasingly employed to ensure continuous and reliable monitoring. However, even advanced automated systems can struggle in challenging conditions. For instance, in March 2024, a self-driving prototype misclassified a cyclist in low light, resulting in a near-miss at 35 km/h [26]. To equip autonomous systems, from self-driving cars to security robots, with continuous, reliable perception, robust video object detection is essential. In this thesis, we use the term *robust* to describe detectors that maintain accurate and temporally consistent predictions under occlusion, motion blur, illumination changes, viewpoint variation, camera motion, and clutter, rather than only in clean, well-framed conditions.

Object detection is a core task in computer vision that involves identifying and localizing instances of predefined classes within images or video frames, typically represented by bounding boxes. Specifically, road user detection, such as identifying vehicles, pedestrians, cyclists, and other dynamic entities, is a critical application in intelligent transportation systems and autonomous driving scenarios. While static image object detection methods analyze each frame independently, video object detection differs fundamentally as it must continuously track and localize moving objects across consecutive frames. Although both static and video detection share challenges like scale variations, viewpoint changes, and occlusions, video detection uniquely faces additional temporal complexities, including motion blur, abrupt appearance and disappearance of objects, and ensuring consistent object detection over time. Moreover, video detection must handle sparse visual cues due to rapid motion, structural occlusions lasting multiple frames, and ideally the requirement for low-latency inference in practical, real-world deployments. Despite recent advances in transformer-based detection architectures such as DETR [27] and Deformable DETR [28], as well as deep learning techniques, robust video-based road user detection remains challenging, motivating the development of specialized spatio-temporal methods.

1.1 Problem Statement

Despite the recent progress in deep learning and vision transformer-based models for video object detection, achieving robust performance in unconstrained real-world environments remains a significant challenge. The complexity arises from multiple visual disturbances, object dynamics, and environmental factors that interfere with stable detection. This thesis specifically investigates the impact of such challenges and the shortcomings of existing appearance-based detection methods that lack temporal awareness and generalization mechanisms. In particular, we focus on failure modes where detection confidence and localization become unstable under occlusion, motion blur, illumination changes, distractors, and camera motion, and on how temporal and semantic cues can mitigate these effects. To study these aspects of robustness, we evaluate our methods on widely used public video benchmarks (e.g., KITTI, Cityscapes, MOT2017, VisDrone, UAVDT, UA-DETRAC) that are explicitly characterized by heavy occlusion, small objects, camera motion, and adverse weather or lighting, making them direct tests of the robustness factors targeted in this thesis. Below, we highlight the key issues of detection in videos.

Partial Occlusion. Partial occlusion occurs when a portion of the object of interest is visually obstructed by another element in the scene. This obstructing element can be another moving object, such as a vehicle or pedestrian, or a stationary component of the environment, such as buildings, poles, or trees. Partial occlusion often results in degraded visibility of the occluded object, making it challenging to extract reliable visual features and maintain consistent tracking over time. Detection models that cannot effectively distinguish between the occluding element (the entity causing obstruction) and the partially visible target object risk incorporating irrelevant visual features, leading to inaccuracies, model drift, and decreased detection precision.

Complete Occlusion and Exit from Field of View. In situations of complete occlusion or when the object temporarily exits the frame, the model loses all visual information. Traditional appearance models struggle to recover detection without explicit memory or contextual support. Although some methods rely on motion constraints, such as constant velocity, to interpolate object positions, such assumptions are frequently violated in unconstrained video scenarios.

Visual Distractors. Distractors refer to objects in the scene that resemble the target object in appearance or texture. These can lead to confusion in the feature space and cause the model to mistakenly detect or shift focus to irrelevant entities. This is especially problematic in crowded scenes or when background textures closely mimic the target appearance.

3D Pose Variation and Articulated Objects. A video represents a sequence of 2D projections of a 3D world. When an object undergoes 3D rotations or articulated deformation (e.g., a human turning their head or changing posture), its appearance can change drastically. Without mechanisms to handle such variations, standard detection models risk frequent failure or misclassification.

Illumination Changes. Lighting variability, either abrupt or gradual, can significantly affect object appearance in video frames. For instance, an object moving from a dim room to a sunlit window undergoes a dramatic shift in color tone and shadow profile. Detection models based on raw visual features may fail to maintain detection consistency under such conditions.

Camera Motion. While many video detection approaches assume static camera setups, this assumption often fails in practice. Scenarios involving mobile platforms (e.g., drones or handheld devices) introduce dynamic backgrounds and motion blur. Detection in such cases requires techniques that can distinguish object motion from camera-induced movement.

Real-Time Constraints. Beyond perceptual challenges, many applications demand fast inference speeds to ensure timely responses, particularly in scenarios such as autonomous driving or real-time surveillance. Achieving high accuracy under the previously mentioned constraints, while simultaneously maintaining computational efficiency, poses a challenging trade-off, especially for large-scale scenes or multi-object detection tasks.

In this thesis, we focus specifically on addressing these challenges by developing detection strategies that are temporally aware, contextually enriched, and robust to appearance shifts and occlusion. The goal is to propose efficient methods tailored to realistic dynamic environments, carefully balancing accuracy and computational efficiency to meet practical application requirements.

Conventional detectors rely heavily on spatial cues extracted from single frames, assuming visual consistency and independence between frames. However, these assumptions break down in unconstrained video. The projection of 3D dynamic scenes into 2D image sequences introduces ambiguity, loss of depth, and deformation. Moreover, standard detectors (e.g., R-CNN) [29] lacks temporal memory. ViT-based detectors [27, 28] handle spatial but not long-term temporal dependencies. They lack mechanisms for temporal memory, resulting in fragmented or inconsistent detections when objects reappear after occlusion or abrupt motion.

To overcome these limitations, this thesis explores architectures that explicitly integrate temporal context, spatial-temporal fusion, and memory-based reasoning. These strategies

are grounded in the theoretical insight that robust perception requires continuity, both in visual representation and in semantic understanding, across time. The next section outlines the research objectives derived from these challenges and theoretical motivations.

1.2 Objectives of the Research

The primary objective of this thesis is to develop spatio-temporal object detection models that capitalize on temporal continuity and semantic generalization using generic features extracted from large-scale pretrained models (e.g., VLMs). Instead of relying on assumptions about object motion or scene conditions, this work focuses on utilizing informative cues from previous frames and *multimodal* embedding representations that combine both visual and linguistic information to improve detection robustness and generalization in real-world video settings.

The specific objectives of this research are as follows:

1. To exploit temporal coherence by designing detection architectures that effectively integrate features from past frames, enabling continuity in object representation even in cases of occlusion, abrupt motion, or partial visibility;

2. To incorporate generic, pretrained visual-language features (e.g., from CLIP) that provide strong semantic priors, improving generalization to unseen categories and challenging conditions without explicit re-training;

3. To develop memory-based query selection mechanisms that identify and retain the most useful object representations over time, supporting consistent detection across long temporal sequences;

4. To evaluate the proposed models on diverse real-world video datasets, with a focus on assessing gains in robustness (as defined above), generalization, and detection stability under occlusion, motion blur, camera motion, and cluttered scenes compared to existing video object detection methods.

1.3 Contributions

The proposed framework for video object detection includes two principal components: (1) a dynamic object appearance modeling mechanism and (2) a transformer-based query refinement process for temporal reasoning and multimodal integration. The methods developed in this thesis are designed to fulfill the research objectives by effectively addressing challenges

such as appearance variability, occlusion, motion blur, and semantic generalization in unconstrained video settings. These contributions are presented in three research articles, each targeting a specific aspect of spatio-temporal perception progressively.

Spatio-Temporal Fusion for Robust Video Object Detection (STF): published at *CRV 2024*.

The first contribution presents a dual-branch attention mechanism to aggregate appearance features across multiple frames while preserving short- and mid-range temporal continuity. This work enhances robustness against motion blur and partial occlusion. The main contributions include:

- A dual-branch attention module that processes consecutive and sparse frames separately, then fuses them to reinforce temporal stability.
- A frame-wise adaptive weighting mechanism that prioritizes frames based on spatial alignment confidence.
- An improvement in temporal consistency and detection robustness without increasing inference time significantly, demonstrated on MOT2017, Cityscapes, and KITTMOT.

Language-Augmented Queries with Evolving Memory (LAQEM): Submitted to the Journal *Pattern Recognition*.

The second contribution proposes a transformer-based model that integrates language-based semantic guidance with a visual memory module for long-range reasoning. This model enables the detection and localization of unseen object categories and recovers object continuity through occlusion. The core contributions are:

- A multimodal query generator that combines visual and language features for improved object generalization.
- An evolving memory mechanism that retains relevant features across long temporal sequences.
- Strong generalization performance on the Visdrone, Cityscapes, UAVDT, and UA-DETRAC datasets and significant mAP gains on rare categories.

Dual-Stream Attention with Multimodal Queries (DAMM): Published at *BMVC 2025*

The third contribution develops a dual-stream transformer that fuses object-centric visual features with contextual semantic information through cross-modal attention layers. This model addresses scene ambiguity and cluttered object layouts. The primary contributions include:

- A structured attention design that separates spatial localization from semantic reasoning, improving clarity in dense environments.
- A multimodal fusion strategy that adapts query embeddings based on content and context using shared attention heads.
- Improved performance under small objects, crowded scenes, and high object similarity scenarios on Visdrone, Cityscapes, UAVDT, and UA-DETRAC.

Together, these three contributions form a cohesive framework for robust video object detection. They are designed to address challenges incrementally: from stabilizing detection across frames (STF), to recovering information via semantic memory (LAQEM), and finally to generalizing across ambiguous inputs using cross-modal representations (DAMM).

1.4 Structure of the Thesis

This dissertation is organized as follows. Chapter 1 introduces the research problem, motivation, and contributions. Chapter 2 reviews the literature on object detection architectures, including two-stage and one-stage detectors, transformer-based models, and spatiotemporal approaches, as well as advances in vision–language models. Chapter 3 provides an overview of the proposed contributions. Chapters 4 through 6 present the three main research articles: Chapter 4 details the Spatio-Temporal Fusion (STF) module, Chapter 5 introduces the Language-Augmented Queries with Evolving Memory (LAQEM) framework, and Chapter 6 describes the Dual-Stream Attention with Multi-Modal Queries (DAMM) architecture for transportation applications. Chapter 7 discusses complementary methodological aspects and additional experimental results. Chapter 8 provides a general discussion on modular fusion, query strategies, and multimodal detection. Finally, Chapter 9 concludes the thesis by summarizing the contributions, highlighting limitations, and outlining directions for future research.

CHAPTER 2 LITERATURE REVIEW

Object detection is a core task in computer vision, enabling systems to locate and classify instances of objects within visual data. Applications span across autonomous vehicles, robotics, video surveillance, and human-computer interaction. Over the past four decades, object detection has undergone significant transformations, primarily driven by the rise of deep learning. This chapter presents a chronological overview of object detection methods, tracing their evolution from handcrafted feature-based techniques to contemporary transformer and vision-language models. We highlight milestones particularly relevant to both image-based and video-based detection.

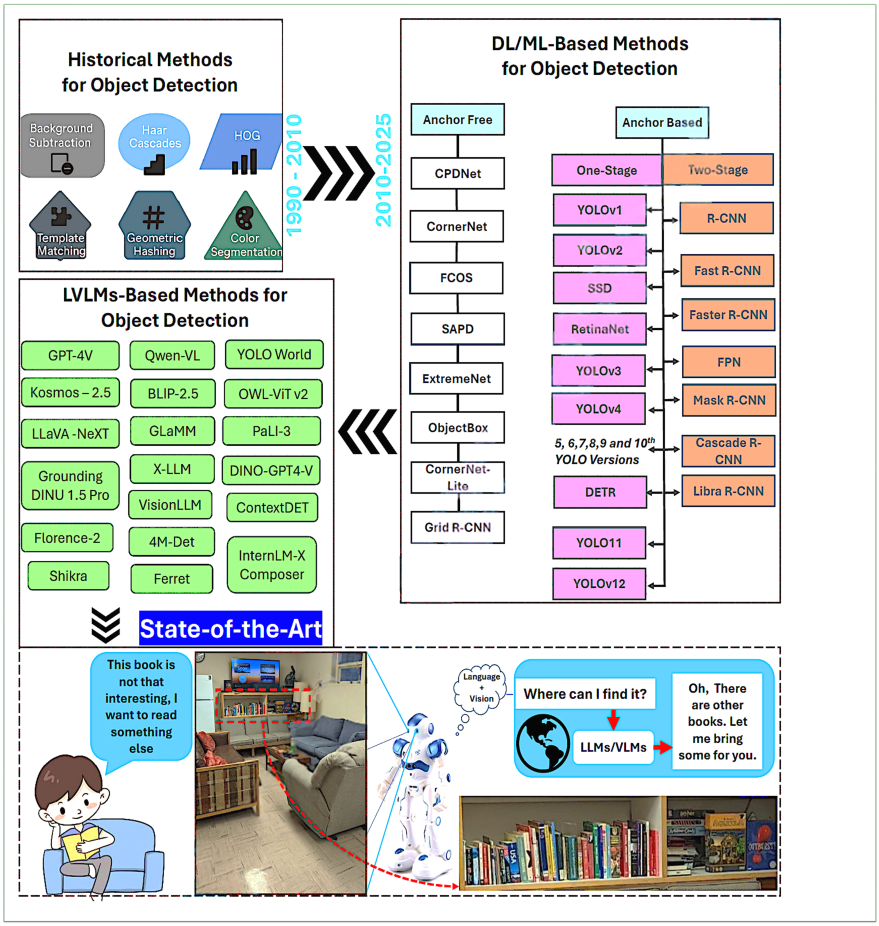

Figure 2.1 Historical evolution of object detection techniques in computer vision, adapted from Sapkota and Karkee [4].

Figure 2.1 presents a historical overview of early object detection methods, which were primarily based on hand-crafted features. Background subtraction methods, such as Gaussian

Mixture Models (GMM) [30], modeled each pixel as a mixture of Gaussians to distinguish foreground motion from a static background. Haar cascades [31] used a sequence of simple rectangular features combined with a boosting algorithm for efficient real-time object detection, notably faces. The Histogram of Oriented Gradients (HOG) [32] extracted gradient orientation histograms from local image patches to represent object shapes, enabling robust detection of pedestrians and rigid structures. While foundational, these methods were sensitive to scale, occlusion, and lighting changes, limiting their effectiveness in complex, real-world settings and motivating the transition toward more complex learning-based models.

From 2010 onward, deep convolutional neural networks (CNNs) revolutionized object detection. The introduction of anchor-based frameworks, such as Faster R-CNN [33], YOLO [34], and RetinaNet [9], enabled real-time detection with high accuracy. These models used predefined bounding boxes and multi-scale feature maps to learn object locations and categories.

In parallel, anchor-free methods such as FCOS [35] and CenterNet [14] emerged as alternatives to anchor-based detectors by predicting object centers and bounding boxes directly from feature maps, rather than relying on predefined anchor boxes. This simplified the detection pipeline and improved localization precision, particularly for objects of varying scales and aspect ratios. However, both anchor-based and anchor-free CNN models often struggled to capture global context, especially in complex or cluttered scenes.

A major paradigm shift occurred with the introduction of DETR [17], which brought transformers, originally developed for NLP, into visual object detection. Unlike previous detectors that relied heavily on convolutional architectures, DETR used self-attention to model global dependencies directly, enabling end-to-end detection without relying on handcrafted components like anchor boxes or post-processing heuristics. Its improved version, Deformable DETR [28], introduced sparse attention mechanisms to better handle small objects and accelerate convergence.

More recently, object detectors that integrate Vision-Language Models (VLMs), such as DetGPT [36], ContextDET [37], and VOLTRON [38], have emerged. These models enhance object understanding by conditioning detection on language prompts or contextual queries, enabling few-shot and zero-shot capabilities. Such integration is particularly promising for open-vocabulary detection and temporal reasoning in video understanding.

This review establishes a foundation for understanding how modern object detection systems have evolved and highlights the transition toward multimodal, context-aware, and temporally consistent architectures. In the following sections, we will go deeper into specific categories of detection systems, including those designed for images and those adapted for video data,

with a particular focus on temporal modelling and memory mechanisms.

2.1 Architectures of Object Detectors

Deep learning-based object detectors generally consist of three main components: a backbone network for feature extraction, a neck module for feature fusion, and a detection head for predicting object classes and bounding boxes. The backbone network extracts hierarchical feature representations from an input image, leveraging multi-stage convolutional layers to capture both low-level and high-level information.

The neck module facilitates multi-scale feature fusion by aggregating information from different stages of the backbone. This module often incorporates top-down and bottom-up pathways, allowing for enhanced spatial resolution and richer contextual information. Finally, the detection head processes the aggregated features to predict object locations and classifications. Based on their architectural designs, object detectors are broadly categorized into two-stage and one-stage models, as well as anchor-based and anchor-free approaches.

2.1.1 Two-Stage Object Detectors

Two-stage object detectors follow a region proposal-based approach, where object detection is performed in two sequential steps: (1) generating candidate regions containing potential objects, and (2) refining these proposals through classification and bounding box regression. A pioneering work in this category is the R-CNN (Region-based CNN) model [5], which utilizes selective search to propose candidate regions, extracts features using a CNN, and subsequently classifies the detected objects. However, R-CNN suffers from inefficiencies due to redundant feature extraction for overlapping proposals.

To address these limitations, Fast R-CNN [39] was introduced, wherein a shared convolutional feature map is used to extract region-specific features, significantly reducing computational overhead as shown in Figure . Faster R-CNN [33] further enhanced this framework by replacing selective search with a Region Proposal Network (RPN), enabling end-to-end region proposal generation. Another notable variant, R-FCN (Region-based Fully Convolutional Network) [40], eliminates connected layers and processes all region proposals in a fully convolutional manner, improving efficiency while maintaining high detection accuracy. Additionally, the Spatial Pyramid Pooling Network (SPP-Net) [41] introduced scale-invariant feature pooling to enhance detection speed without compromising accuracy.

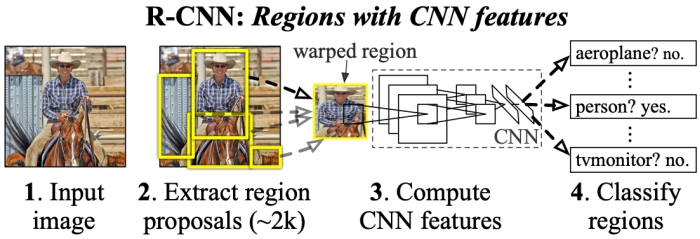

Figure 2.2 Overview of a two-stage object detector. © 2014 IEEE. Reprinted, with permission, from Girshick et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *CVPR*, 2014 [5].

Another enhancement in two-stage detection is the D2Det framework [42], which introduces dense local regression in place of conventional bounding box regression to improve detection accuracy. By employing discriminative RoI pooling, D2Det assigns adaptive weights to object regions, reducing background interference. However, like other two-stage models, D2Det exhibits limitations in detecting small objects due to its reliance on high-resolution feature maps.

Beyond standard RPN-based approaches, ClusterNet [43] integrates spatio-temporal CNNs to enhance video object detection. Unlike conventional two-stage models, ClusterNet processes multiple frames to generate region proposals, combining motion and appearance cues for improved object localization. The FoveaNet module within ClusterNet further refines object centers using heatmap estimation, demonstrating superior accuracy for temporally coherent object detection.

2.1.2 One-Stage Object Detectors

While two-stage detectors achieve high accuracy, their computational complexity limits real-time applications. One-stage detectors address this issue by eliminating the region proposal step and directly predicting object classes and bounding boxes over dense feature grids. YOLO (You Only Look Once) [6] is a prominent one-stage detector as shown in Figure , that partitions an image into a grid and predicts object probabilities and bounding boxes within each grid cell. This approach enables real-time detection with high efficiency but at the cost of reduced localization accuracy.

Subsequent iterations of YOLO, such as YOLOv2 [34] and YOLOv3 [44], introduced improvements in multi-scale detection and feature extraction. YOLOv4 [45] further refined these models by enhancing feature pyramids and introducing novel data augmentation tech-

niques.

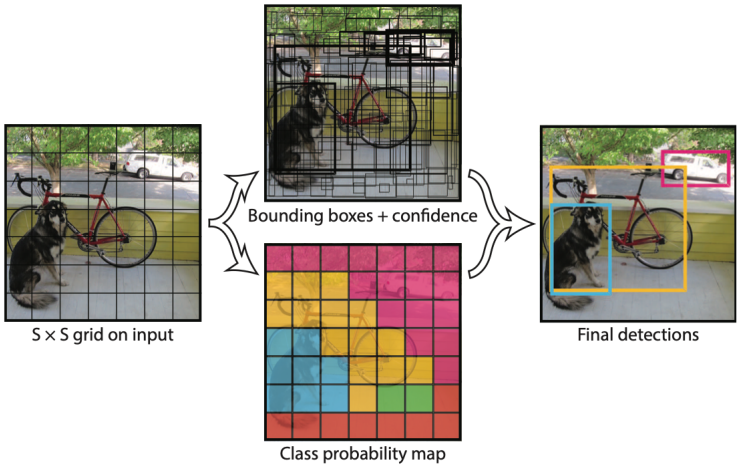

Figure 2.3 YOLO object detection architecture. © 2016 IEEE. Reprinted, with permission, from Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection," *CVPR*, 2016 [6].

Another widely used one-stage detector is the Single Shot MultiBox Detector (SSD) [46], which employs a VGG-based backbone and integrates multi-scale feature maps to enhance small object detection. Unlike YOLO, SSD applies predictions at multiple feature pyramid levels, improving detection across different object sizes. Enhanced versions of SSD, such as DSSD [47], incorporate deconvolution layers to improve feature resolution, benefiting small object detection.

Feature fusion techniques have also been explored to enhance detection robustness. Methods, such as Feature-Fused SSD [48], integrate multi-scale feature aggregation to improve contextual representation. Additionally, lightweight networks with attention mechanisms [49–51] combine feature pyramids with selective attention to boost detection accuracy while maintaining computational efficiency.

2.1.3 Differences Between Anchor-Based and Anchor-Free Detectors

Anchor-based and anchor-free detectors represent two primary paradigms in object detection. Anchor-based methods utilize predefined anchor boxes to handle variations in scale and aspect ratio, discretizing potential object locations into a fixed set of candidate boxes. These methods have been widely adopted in both one-stage and two-stage detectors due to their ability to improve detection accuracy. However, a significant drawback of anchor-based detectors is the need for a large number of predefined anchor configurations, which increases

computational complexity and memory usage. This issue becomes particularly critical for detecting small objects, where inappropriate anchor box selection can lead to poor localization and classification performance [33, 44].

In contrast, anchor-free detectors have gained popularity due to their simplicity and flexibility. Unlike anchor-based models, anchor-free approaches directly predict object locations and bounding box coordinates without relying on predefined anchor templates. Figure 2.4 illustrates the key distinction between these two approaches. Anchor-based models depend on predefined aspect ratios and sizes, whereas anchor-free methods estimate object locations directly from feature maps.

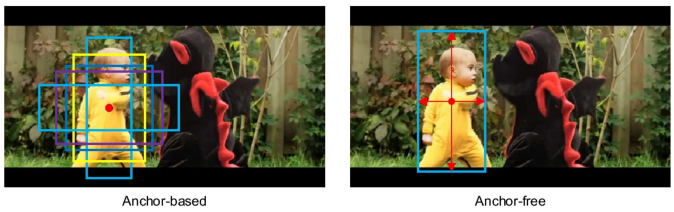

Figure 2.4 Differences between anchor-based and anchor-free detection. © 2021 Nature Publishing Group. Adapted, with permission, from Zhang et al., "A Siamese Query Network for Efficient Video Object Detection," *Nature Communications*, 2021 [7]. Used with permission for non-commercial academic purposes.

A common approach for improving anchor-based detection is the integration of feature pyramid networks (FPNs), where different size feature maps are used to associate large anchor boxes with higher-level features and small anchor boxes with lower-level features. This hierarchical design improves scale invariance and detection accuracy.However, anchor-based methods still have two primary limitations: (1) overlap-based anchor sampling, which can lead to redundant predictions, and (2) heuristic-based feature selection, which restricts adaptability across diverse datasets.

Recent research [52] has proposed feature-selective anchor-free modules that allow networks to dynamically select the optimal feature level for each instance. Unlike anchor-based approaches, which depend on predefined heuristics, anchor-free methods leverage instance-based learning to refine feature selection. This approach enhances adaptability in detecting objects of varying sizes and aspect ratios while reducing computational overhead.

Despite their advantages, anchor-free detectors also have limitations. They often struggle with precise localization, particularly in high-density object scenes where bounding box regression becomes more challenging. Additionally, the absence of predefined anchors may hinder their performance when detecting objects with extreme aspect ratios or varying orien-

tations. The choice between anchor-based and anchor-free detectors depends on the specific application, requiring a trade-off between accuracy, computational efficiency, and adaptability to different object scales and environments.

2.2 Survey on Object Detection Approaches

This section presents a comprehensive survey of object detection approaches, categorized based on the type of input data utilized (such as multi-frame information), the challenges addressed (including data imbalance and scale variations), and the architectural design choices employed (such as anchor-free methods).

2.2.1 Multi-Frame Object Detection

Multi-frame object detection enhances accuracy by incorporating temporal context, addressing issues such as motion blur, occlusion, and sudden appearance changes. Early methods like FGFA [53] and Deep Feature Flow [54] aggregate features or detections from adjacent frames, a strategy often categorized as box-level temporal reasoning. However, these approaches typically rely on precomputed motion or post-processing steps rather than modeling temporal dependencies in an end-to-end fashion.

FGFA attempts to address this by introducing flow-guided feature aggregation along motion trajectories, enabling temporally aligned feature fusion that improves per-frame representations. Similarly, RN-VID [55] leveraged temporal redundancy by merging multi-frame feature maps using channel reordering and 1×1 convolutions. This fusion architecture facilitates effective reuse of spatio-temporal information, yielding more robust detection compared to isolated frame-wise processing.

Similarly, FFAVOD (Feature Fusion Architecture for Video Object Detection) [56] utilizes the temporal redundancy in video frames to enhance object detection performance. The model aggregates feature representations from adjacent frames to refine object detection in the target frame. By incorporating a feature-sharing mechanism across consecutive frames and a dedicated feature fusion module, FFAVOD demonstrates superior detection accuracy compared to conventional single-frame detectors. This method has shown notable improvements in detecting moving objects in road user scenarios.

An alternative approach, CenterTrack [57], employs a point-based tracking framework that integrates object detection and tracking in a unified pipeline. Instead of bounding boxes, CenterTrack represents objects using single points located at their center. This method predicts object centers across frames, allowing the network to associate detections over time

efficiently. The framework generates a heatmap of object centers along with bounding box size maps and offset maps, ensuring robustness against occlusions and abrupt motion variations.

Another notable method, the Spatiotemporal Sampling Network (STSN) [58], introduces deformable convolutions to sample spatial features from adjacent frames dynamically. This approach enhances robustness against occlusions and motion blur by directly optimizing the sampling locations.

Furthermore, the Recurrent Multi-Frame Single Shot Detector (MF-SSD) [59] extends the Single Shot Detector (SSD) architecture by incorporating recurrent convolutional modules. This modification enables the integration of temporal information across multiple frames, significantly improving object detection in videos. MF-SSD extracts feature maps from consecutive frames and uses recurrent units to encode temporal dependencies. The recurrently generated feature maps are then processed by convolutional layers to produce refined bounding box predictions and class probabilities. This design allows the network to leverage prior frame information effectively, thereby enhancing detection accuracy for moving objects.

2.2.2 Detection of Small Objects

Various approaches have been proposed to enhance the resolution of low-resolution images to improve object detection performance. One widely used method is the Feature Pyramid Network (FPN) [60], which fuses features across multiple scales to construct feature pyramids, where higher-level feature maps detect larger objects and lower-level feature maps focus on smaller objects. However, small object detection remains a significant challenge due to the limited spatial information available in low-resolution images. While coupling features from different scales has shown improvements, it does not fully resolve the issue of small object detection.

To address these limitations, the Extended Feature Pyramid Network (EFPN) [61] introduces an additional high-resolution pyramid level explicitly designed for detecting small objects. EFPN incorporates a novel module called "Feature Texture Transfer" (FTT). This simultaneously enhances feature resolution and refines regional texture details. In the FTT framework, a feature extractor captures key semantic information, followed by sub-pixel convolution to increase the feature resolution. The texture extractor selectively enhances regional textures for small object detection, while residual connections fuse these textures with the super-resolved features, ultimately creating P3 for the extended feature pyramid. This framework aims to improve small object detection by providing more reliable and detailed feature representations.

Most CNN-based approaches treat channel-wise features uniformly, which limits their adaptability in enhancing small object representations. Image super-resolution methods aim to reconstruct high-frequency details that are often lost in low-resolution images. However, traditional super-resolution techniques primarily focus on recovering low-frequency details, neglecting the contextual relationships between objects of different scales. A more advanced strategy for multi-scale object representation is the Perceptual Generative Adversarial Network (Perceptual GAN) [8], which constructs super-resolved representations for small objects to improve detection accuracy. As shown in Figure 2.5, this network comprises a generator and a discriminator. The generator is a deep residual-based feature enhancement model that refines multi-scale object features by integrating fine-grained details from lower-level layers. This process effectively performs "super-resolution" on intermediate representations. The discriminator network supervises the learning process by distinguishing between super-resolved representations of small objects and real large objects, ensuring that the generated features are both discriminative and contextually meaningful. By employing adversarial loss and perceptual loss, this approach enhances detection performance by generating high-quality feature representations.

Several other methods utilize image super-resolution to refine object proposals selectively. Instead of performing super-resolution on entire images, some approaches focus only on enhancing small object regions within larger object proposals, reducing computational costs while improving detection accuracy. Traditional super-resolution methods often ignore contextual information, limiting their effectiveness in real-world object detection scenarios. Feature-level super-resolution addresses this limitation by extracting proposal features with broad receptive fields through successive convolutional operations, leveraging contextual information to refine object boundaries and maintain spatial relationships. By preserving the spatial structure and contextual information of objects, feature-level super-resolution improves object detection accuracy, particularly for small and occluded objects.

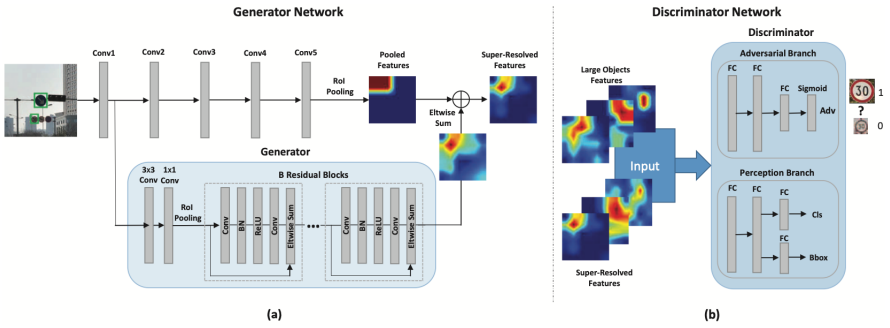

Figure 2.5 Architecture of Perceptual GAN for Super-Resolution in Object Detection. © 2017 IEEE. Reprinted, with permission, from Li et al., "Perceptual Generative Adversarial Networks for Small Object Detection," *CVPR*, 2017 [8].

2.2.3 Approaches for Addressing the Data Imbalance Problem

Many object detectors rely on a two-stage detection framework, such as R-CNN [33], [39]. These two-stage models mitigate class imbalance through cascaded detection and sampling heuristics. The region proposal stage significantly reduces the number of candidate object locations to a small set (typically 1-2k), filtering out most background samples. In the subsequent classification stage, sampling heuristics, such as maintaining a fixed foreground-to-background ratio (e.g., 1:3), help preserve a balanced distribution between object and background instances.

In contrast, one-stage detectors [6, 44, 46] suffer from severe class imbalance due to their dense sampling of anchor boxes across an image, leading to an overwhelming number of easy negative samples. To address this issue, Lin et al. [9] introduced focal loss, which reshapes the standard cross-entropy loss to down-weight well-classified examples, focusing training on a sparse set of hard examples. By reducing the influence of easily classified samples, focal loss prevents the model from being overwhelmed by the disproportionate number of simple background examples, thereby improving detection performance.

The combination of focal loss with Feature Pyramid Networks (FPN) [60] has enabled one-stage object detectors to achieve performance levels comparable to two-stage detectors, such as Faster R-CNN. Figure 2.6 illustrates the architecture of this method, which employs RetinaNet as the base network for computing convolutional feature maps over input images. The model integrates two task-specific sub-networks: one for classification and another for regression. Each level of the feature pyramid detects objects at different scales, leveraging FPN multi-scale feature representation to enhance detection across varying object sizes [62].

This approach has been extensively explored in the literature [45, 63–65], not only to address the class imbalance problem but also to improve scale-variant instance detection.

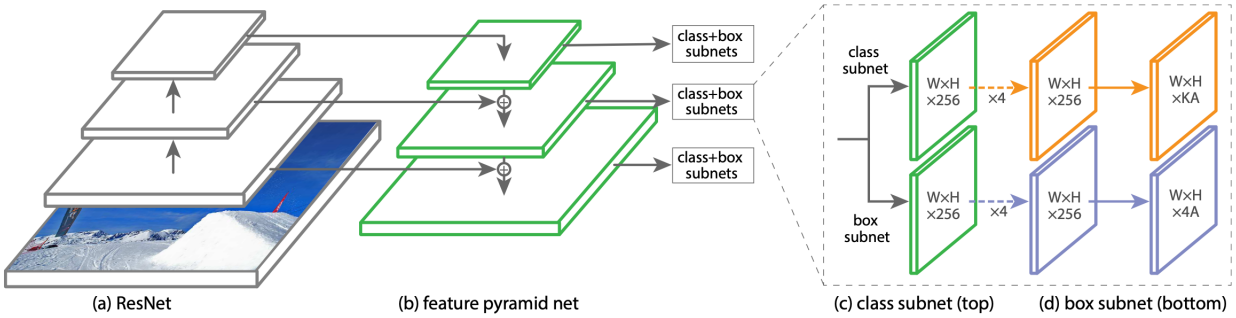

Figure 2.6 Architecture of the one-stage RetinaNet object detection network. © 2017 IEEE. Reprinted, with permission, from Lin et al., "Focal Loss for Dense Object Detection," *ICCV*, 2017 [9].

Another effective technique for mitigating class imbalance in object detection is data augmentation. Several studies have demonstrated that augmentation techniques improve data efficiency, particularly in imbalanced datasets where certain object classes are underrepresented. One simple yet effective approach, Copy-Paste augmentation [10], generates new training samples by copying objects from one image and pasting them onto another. This method involves randomly selecting two images, applying random scale jittering and horizontal flipping, and then copying objects from one image onto the other. Objects that are entirely occluded are removed, while partially occluded masks and bounding boxes are updated accordingly. Figure 2.7 illustrates the Copy-Paste process. This augmentation strategy enhances model generalization and robustness, making it an effective solution for addressing data imbalance in object detection tasks.

Figure 2.7 Copy-Paste data augmentation process. © 2021 IEEE. Reprinted, with permission, from Ghiasi et al., "Simple Copy-Paste is a Strong Data Augmentation Method for Instance Segmentation," *CVPR*, 2021 [10].

2.2.4 Handling Scale Variation

Detecting objects at varying scales remains a fundamental challenge due to differences in camera perspective and object size. Traditional CNNs often fail to consistently detect small or distant objects, especially in cluttered scenes. A common strategy to address this involves constructing multi-scale image pyramids [62], where the same image is resized to different scales and processed independently. While effective, this increases inference time and limits real-time applicability.

To overcome these constraints, modern detectors such as FPN [60] and its extensions [61,66] adopt feature pyramid architectures. These methods leverage the inherent hierarchical nature of CNNs by combining low-level, high-resolution features (better for small objects) with high-level, semantically rich features (better for large objects). This top-down pathway with lateral connections enables detectors to make predictions across multiple feature scales within a single forward pass, improving both accuracy and efficiency.

Another approach, explored in [11], leverages deformable convolution networks to increase the receptive field of the detector. Unlike conventional CNNs, deformable convolutions introduce spatial offsets to sampling locations, which are learned from the target task without additional supervision. This method can replace standard CNN layers and be trained end-to-end using backpropagation. Figure 2.8 illustrates the process, where RoI pooling generates pooled feature maps, and a fully connected (fc) layer computes normalized offsets, which are then applied to the transformation.

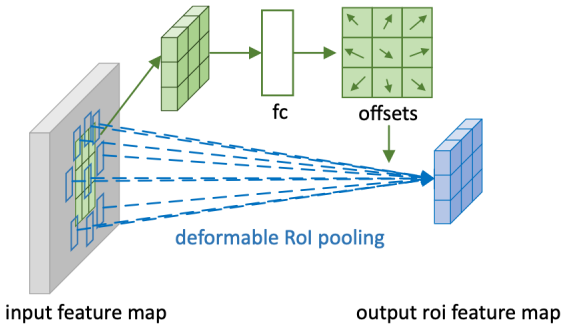

Figure 2.8 Deformable 3x3 RoI pooling. © 2017 IEEE. Reprinted, with permission, from Dai et al., "Deformable Convolutional Networks," *ICCV*, 2017 [11].

Singh et al. [12] investigated the difficulty CNNs face in detecting objects across a wide range of scales. They proposed Scale Normalization for Image Pyramids (SNIP), a technique that selectively trains and tests detectors on object instances only within appropriate scale ranges at each level of an image pyramid. This avoids training on extremely small or large objects at inappropriate resolutions, thereby improving both efficiency and detection accuracy. SNIP selectively backpropagates object gradients based on image scale, ensuring robust scale-invariant learning. The framework employs a multi-scale image pyramid that normalizes input representations for objects across different scales. Their approach utilizes ResNet-101 as the backbone and introduces CNN-B and CNN-S modules to process low- and high-resolution images efficiently, as depicted in Figure 2.9. CNN-S is trained on low-resolution images, while CNN-B is first trained on high-resolution images and then fine-tuned on up-sampled low-resolution images. This method optimizes computation by leveraging lower-resolution images during training and avoiding unnecessary backpropagation for large objects in high-resolution images, thereby improving efficiency.

The influence of receptive field size, network depth, and downsampling rate on object detection performance has been studied in [67]. The authors proposed TridentNet, a multi-branch architecture designed to handle scale variations efficiently. TridentNet employs scale-aware training and a trident block structure,where each branch shares parameters while using different dilation rates to produce scale-specific feature maps. Non-maximum Suppression (NMS) is used to aggregate final detections from multiple branches, ensuring robust multi-scale object detection. Additionally, the model scale-aware training strategy equips each branch with specialized abilities for detecting objects at different scales, leading to significant improve-

ments in detection performance.

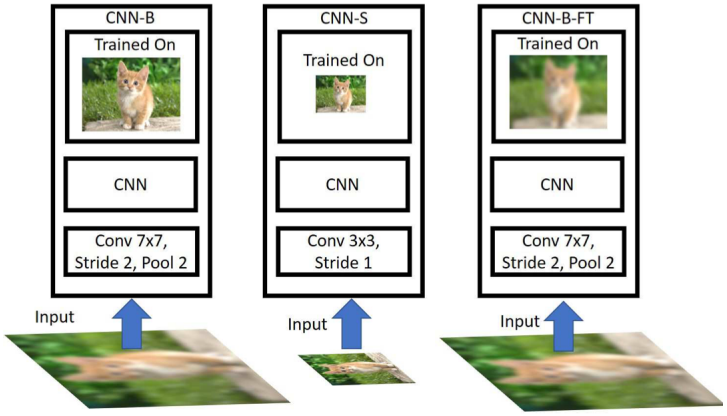

Figure 2.9 CNN architecture for Scale Normalization for Image Pyramids (SNIP). © 2018 IEEE. Reprinted, with permission, from Singh and Davis, "An Analysis of Scale Invariance in Object Detection," *CVPR*, 2018 [12].

Some object detection architectures combine multi-level features using fully connected or heuristic approaches to address multi-scale detection challenges. While several studies have introduced top-down pathways, such as Feature Pyramid Networks (FPN) [60], accumulating deeper features can weaken the network ability to capture fine details in shallow feature maps. Other methods [61, 68] enhance multi-scale detection by integrating deeper layers with shallower layers. However, excessive feature aggregation without balancing different resolutions can lead to increased computational complexity, high noise levels, and fusion difficulties. Low-resolution, high-level feature maps may limit object detection performance, making it challenging to detect fine details or recognize small objects accurately. Feature pyramid-based approaches aim to alleviate these issues by leveraging multi-scale information to enhance object detection.

2.2.5 Anchor-Free Approaches

The evolution of pixel-level image segmentation using dense prediction, as introduced by FCN [69] and FCOS (Fully Convolutional One-Stage Object Detection) [35], served as the foundation for modern anchor-free detectors. These approaches demonstrated that object detection can be effectively performed without predefined anchors, simplifying the detection pipeline while maintaining high accuracy.

One of the earliest keypoint-based anchor-free detectors is CornerNet [13], which introduced

a novel approach to detect objects by identifying their top-left and bottom-right corners as keypoints. This method employs a single convolutional neural network (CNN) to predict a heatmap for each corner and generates embedding vectors to associate them, thereby eliminating the need for predefined anchor boxes. Figure 2.10 illustrates the CornerNet architecture.

The CenterNet variant by Duan et al. [14], which extends CornerNet, introduces an additional center keypoint to validate object hypotheses formed by corner pairings. CenterNet is a one-stage, anchor-free detector that represents each object by its center point and predicts its size and offset directly, enabling fast and accurate per-frame object detection without region proposals. This version should not be confused with the anchor-free CenterNet by Zhou et al. [70], which detects objects by directly regressing center points from feature maps. During training, predicted bounding boxes with high Intersection over Union (IoU) with ground-truth boxes are assigned to the corresponding center keypoints. To improve detection accuracy, CenterNet employs center pooling and cascade corner pooling to refine center keypoint predictions. Figure 2.11 illustrates the network architecture, which generates heatmaps for corners and centers, proposes candidate bounding boxes, and refines them using center cues.

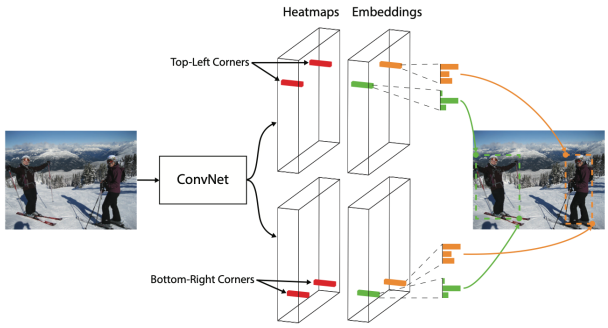

Figure 2.10 Architecture of an anchor-free detector. © 2020 Springer Nature. Reprinted, with permission, from Law and Deng, "CornerNet: Detecting Objects as Paired Keypoints," *International Journal of Computer Vision*, 2020 [13].

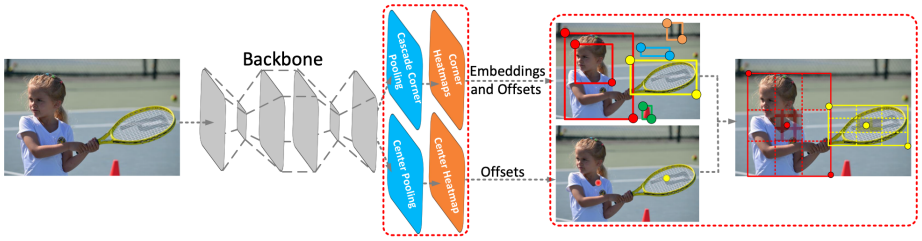

Figure 2.11 Architecture of CenterNet. © 2019 IEEE. Reprinted, with permission, from Duan et al., "CenterNet: Keypoint Triplets for Object Detection," *ICCV*, 2019 [14].

While the more widely known CenterNet by Zhou et al. [70] detects object centers directly using a heatmap-based approach, other keypoint-based methods rely on corner pairings. However, these approaches can suffer from incorrect keypoint matching. CentripetalNet [71] addresses this limitation by introducing a centripetal shift mechanism that guides paired corner keypoints toward the object center, improving association accuracy. It also incorporates corner pooling to enhance spatial context and uses a cross-star deformable convolution network to dynamically refine corner features.

Recent advancements in anchor-free object detection include YOLOX [15], a novel YOLO variant that eliminates anchor boxes while incorporating a decoupled detection head. The YOLOX architecture builds upon YOLOv3 [44] by separating the classification and localization tasks into distinct branches, thereby improving convergence and overall detection performance. As depicted in Figure 2.12, each Feature Pyramid Network (FPN) level applies a 1×1 convolution layer to constrain the feature channels to 256, followed by two parallel branches for regression and IoU estimation.

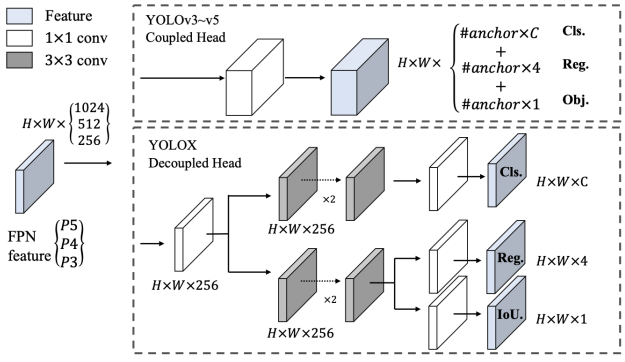

Figure 2.12 Comparison of YOLOv3 and YOLOX. Reprinted from Ge et al., "YOLOX: Exceeding YOLO Series in 2021," arXiv preprint arXiv:2107.08430 [15].

Anchor-free object detection methods, such as CenterNet [70] and YOLOX [15], have become popular for their streamlined architecture and competitive accuracy. These models eliminate the need for predefined anchor boxes by directly predicting keypoints or object centers. However, they still rely on heuristic rules, such as defining a fixed radius or center region around ground-truth points, to determine positive training samples. Such fixed-threshold strategies may not generalize well across varying object scales or aspect ratios, making it challenging to distinguish foreground from background under diverse conditions. Recent efforts explore adaptive label assignment to address this limitation and improve robustness.

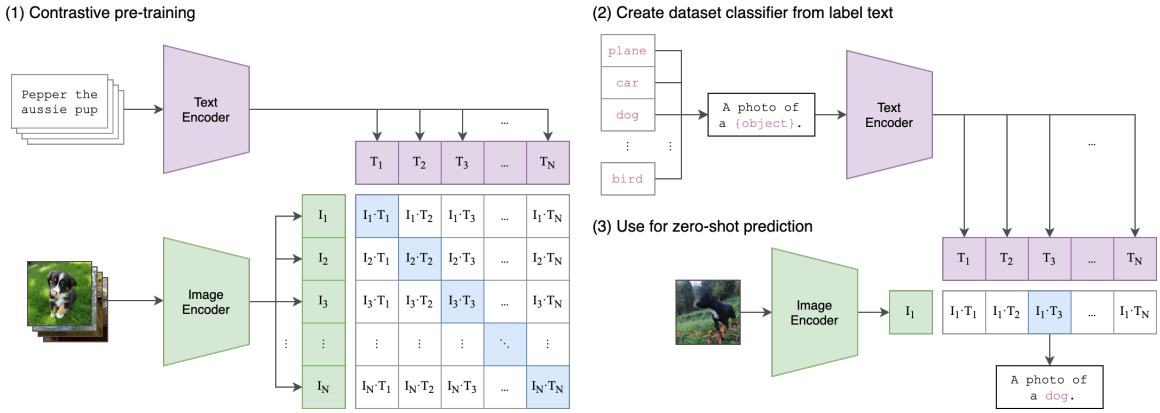

Figure 2.13 Vision-language alignment with CLIP for object detection. © 2021 by the authors. Reprinted, with permission, from Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR 139, 2021 [16].

2.2.6 End-to-End Detection with DETR and Deformable Variant

The DEtection TRansformer (DETR) [17] marked a significant departure from conventional object detectors by introducing an end-to-end detection framework built entirely on transformers. DETR formulates object detection as a direct set prediction problem using a bipartite matching loss to align predicted objects with ground-truth annotations. This eliminates the need for hand-crafted components such as anchor boxes, region proposal networks, and non-maximum suppression (NMS), streamlining the detection pipeline.

DETR employs a CNN backbone to extract image features, followed by a transformer encoder-decoder architecture that models global relationships between spatial features. A fixed number of learnable object queries are passed through the decoder, where each query attends to the encoded feature map to generate class labels and bounding box coordinates. Despite its conceptual elegance, DETR suffers from slow convergence and difficulty detecting

small objects due to its dense attention and global query design. These limitations restrict its applicability in high-resolution or real-time video settings.

To overcome these issues, Deformable DETR [28] introduces deformable attention modules that replace dense attention with a sparse sampling strategy. Instead of attending to all spatial locations, deformable attention dynamically selects a small set of relevant positions across multi-scale feature maps, guided by learnable offsets. This reduces the computational overhead and improves the model ability to localize small or crowded objects.

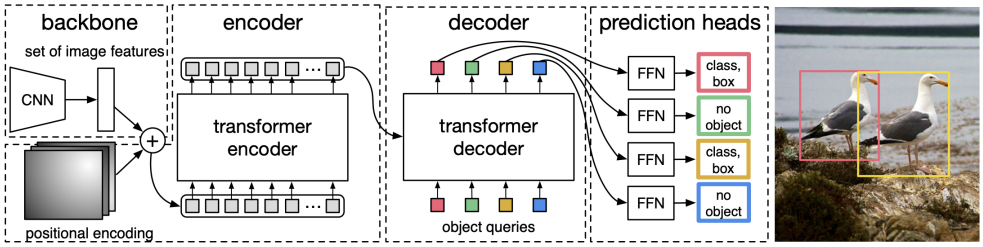

Figure 2.14 Illustration of DETR transformer-based encoder-decoder framework. © 2020 Springer Nature Switzerland AG. Reprinted, with permission, from Carion et al., "End-to-End Object Detection with Transformers," *ECCV*, 2020 [17].

2.2.7 Long-Term Temporal Modeling

Modeling long-term temporal dependencies is essential in video object detection, particularly when objects are occluded or exit and re-enter the scene. Traditional short-term feature aggregation methods often fall short in such cases due to their limited temporal receptive field.

Memory-Enhanced Global Attention (MEGA) [18] addresses this by introducing a memory-augmented transformer architecture that combines self-attention with an external memory module as shown in Figure . The memory stores representative key-value pairs from past frames and is dynamically updated as the video progresses. This allows the model to retrieve and integrate relevant historical information into current predictions, improving temporal consistency and robustness to occlusions. MEGA operates by maintaining both short-term tokens (from nearby frames) and long-term tokens (from memory), enabling dual-stream attention. The global attention mechanism computes interactions between all tokens, while the memory module ensures continuity in object-level reasoning across extended time spans. As a result, MEGA improves object detection accuracy, especially in challenging video sequences with motion blur, appearance changes, or intermittent visibility.

TubeDETR [72] builds upon the DETR framework by reformulating object detection in

videos as a tube generation task, where a tube refers to a sequence of bounding boxes tracking the same object across multiple frames. Unlike traditional per-frame detectors followed by post-hoc association, TubeDETR unifies detection and tracking within a single end-to-end transformer-based architecture. The model leverages spatio-temporal transformers to jointly encode visual features across frames and outputs space-time object tubes directly. Each tube is generated as a sequence of bounding boxes linked by shared object identity, improving temporal coherence and reducing identity switches. TubeDETR also incorporates a bipartite matching mechanism similar to DETR to match predicted tubes with ground-truth annotations during training. By modeling object trajectories holistically rather than as isolated detections, TubeDETR significantly improves both detection accuracy and identity consistency in video benchmarks. This tube-level formulation inspires the motivation behind modular memory and attention fusion strategies explored in this thesis.

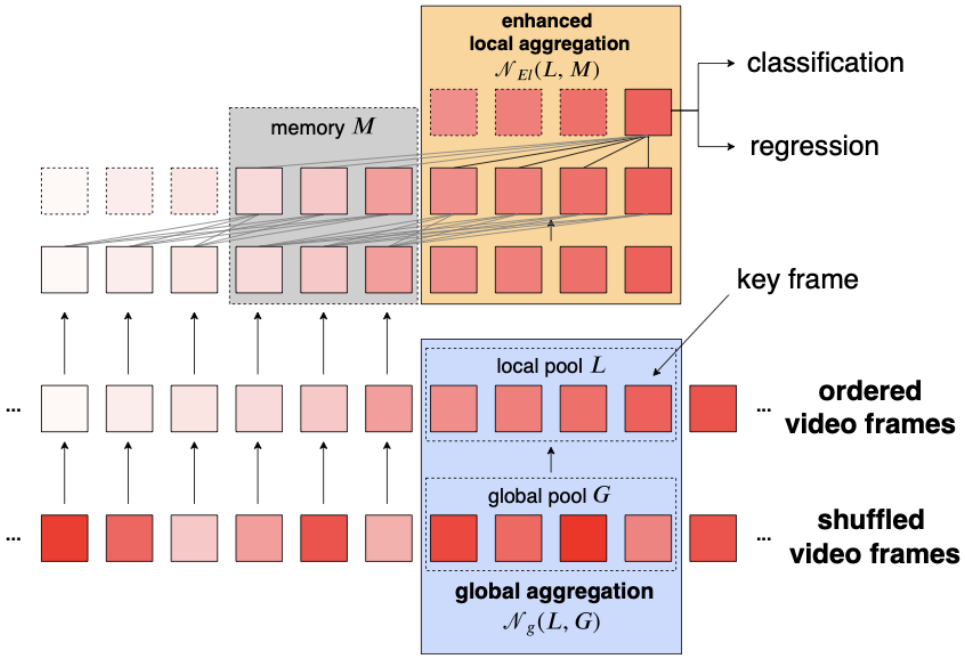

Figure 2.15 MEGA architecture integrating temporal memory for consistent object recognition. © 2020 IEEE. Reprinted, with permission, from Chen et al., "Memory Enhanced Global-Local Aggregation for Video Object Detection," *CVPR*, 2020 [18].

2.2.8 Scalability in Transformer-Based Detection

While transformer models offer global context and strong representational capacity, their quadratic complexity limits scalability in high-resolution tasks. To mitigate this, the Swin Transformer [19] was introduced as a hierarchical vision transformer that uses shifted windows

for local attention, significantly reducing computational cost. Swin architecture has been successfully integrated into object detection frameworks such as Cascade Mask R-CNN and RetinaNet through the MMDetection toolbox [73], where it serves as an efficient and scalable backbone. Its hierarchical design enables multi-scale feature extraction, making it well-suited for dense detection tasks.

As shown in Figure 2.16, the shifted window mechanism allows the model to maintain locality while building long-range dependencies across layers. This balance of efficiency and representational power makes Swin Transformer a practical backbone for detection in both resource-constrained and high-resolution settings.

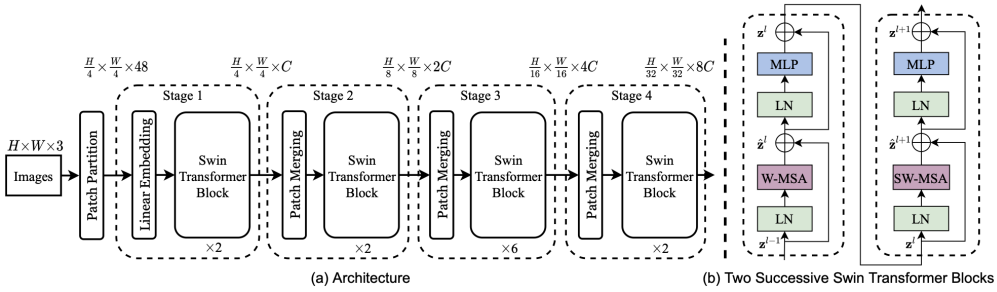

Figure 2.16 Hierarchical design of Swin Transformer using shifted windows for efficient computation. © 2021 IEEE. Reprinted, with permission, from Liu et al., "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," *ICCV*, 2021 [19].

2.3 Vision-Language Models, LLMs, and Open-Vocabulary Detection

The integration of vision-language models and large language models (LLMs) has transformed object detection by enabling open-vocabulary recognition and contextual grounding. Models like ContextDET [37] enhance detection by conditioning object queries on contextual sentences, while DetGPT [36] leverages large language models to frame detection as a generative task, producing text-aligned predictions. VOLTRON [38] grounds object proposals using phrase-level supervision, bridging region-text alignment. Grounding DINO [20] introduces contrastive learning within a DETR framework to link text phrases with image regions, enabling phrase grounding and zero-shot detection. GLIP [21] pre-trains on region-text pairs to learn strong grounding representations that generalize across tasks. OWL-ViT [22] supports zero-shot detection by matching region features with textual queries via cosine similarity, while Detic [74] separates localization from classification and uses image-level labels with vision-language embeddings for scalable category generalization. ViLD [75] employs CLIP-based embeddings for region classification, achieving strong results on both seen and novel

categories. Collectively, these approaches leverage semantic priors and compositional reasoning to address vocabulary limitations and enable detection under minimal supervision or in previously unseen categories.

Grounding DINO [20] introduces a robust mechanism for aligning visual and textual modalities through contrastive learning within a transformer-based architecture. By leveraging a pretrained language encoder alongside a vision transformer, Grounding DINO effectively maps textual queries to relevant image regions. This method refines object detection by explicitly grounding language descriptions to visual features, enabling precise object localization driven by semantic guidance. Its effective cross-modal alignment facilitates strong generalization capabilities, particularly in open-vocabulary scenarios, making it suitable for diverse and dynamic real-world environments. Figure 2.17 illustrates the general architectural framework of Grounding DINO.

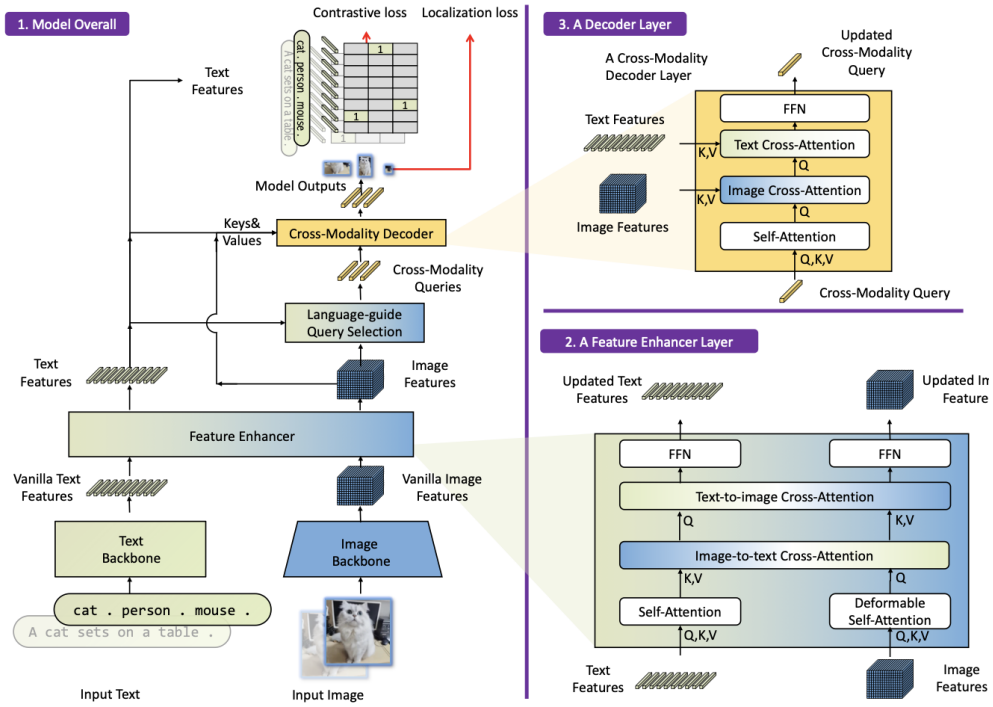

Figure 2.17 Overview of Grounding DINO architecture, illustrating contrastive alignment between textual and visual embeddings to facilitate semantic-guided object detection. © 2023 IEEE. Reprinted, with permission, from Liu et al., "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection," *CVPR*, 2023 [20].

GLIP [21] introduces a unified framework that blends language grounding with object detection by pre-training on a large set of region-text pairs. Unlike Grounding DINO [20], which integrates vision-language alignment into the DETR architecture using contrastive training

for phrase grounding, GLIP adopts a region classification approach, training a two-stream model where visual and textual embeddings are aligned in a shared space. This enables GLIP to perform both grounding and detection without requiring dense annotations for every category. Its design allows the model to generalize to novel categories and compositional phrases with minimal supervision. Figure 2.18 illustrates GLIP language-image pretraining strategy, highlighting how joint embedding improves zero-shot and open-vocabulary detection performance.

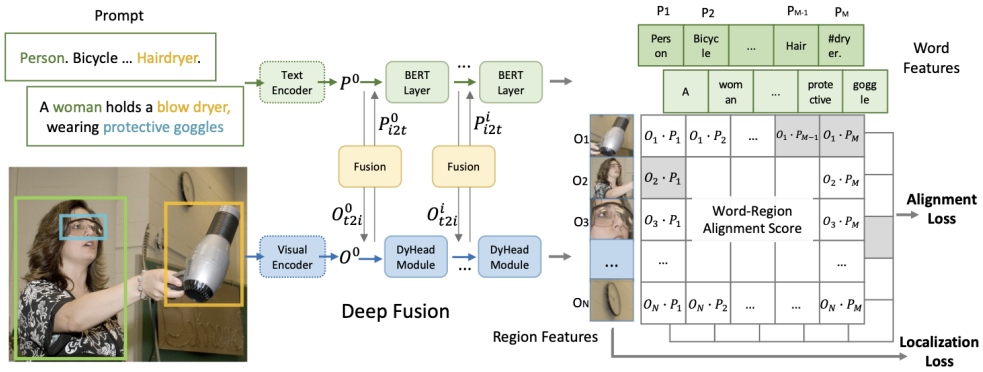

Figure 2.18 Illustration of GLIP unified language-image pre-training approach, highlighting the joint embedding of textual and visual features to enhance generalization in object detection tasks. © 2022 IEEE. Reprinted, with permission, from Li et al., "Grounded Language-Image Pre-training," *CVPR*, 2022 [21].

OWL-ViT [22] is a transformer-based framework developed for open-world object detection. It extends Vision Transformers (ViTs) by pairing them with a text encoder to enable open-vocabulary generalization. During training, OWL-ViT learns to align visual features from image regions with language embeddings derived from class names or descriptions. At inference, given a set of textual prompts (e.g., "a bicycle," "a person"), the model computes similarity scores between region features and these language queries to localize and classify objects, without requiring explicit bounding box annotations for all categories. This text-conditioned detection mechanism enables OWL-ViT to dynamically detect novel objects and scale to large label spaces. The model inference pipeline and architecture are illustrated in Figure 2.19.

Beyond task-specific open-vocabulary detectors, recent work has begun to explore *object-centric foundation models* that jointly handle images and videos. A representative example is GLEE (General Object Foundation Model for Images and Videos at Scale) [76]. GLEE is trained on millions of images from diverse benchmarks with multi-granularity supervision and adopts a unified architecture that combines an image encoder, text encoder, visual prompter,

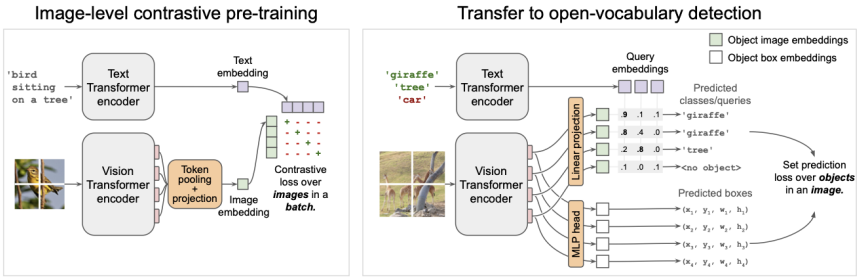

Figure 2.19 OWL-ViT architecture demonstrating integration of language embeddings with visual transformer outputs, enabling open-world detection of unseen categories. © 2022 IEEE. Reprinted, with permission, from Minderer et al., "Simple Open-Vocabulary Object Detection with Vision Transformers," *CVPR*, 2022 [22].

and object decoder. Within a single framework, it can perform detection, segmentation, tracking, grounding, and identification of arbitrary objects in open-world scenarios, and exhibits strong zero-shot transfer to new datasets and tasks. [76]

These video-oriented foundation models implicitly encode both temporal and semantic coherence in their learned object representations and can serve as powerful generic backbones or teachers for downstream perception systems. However, their scale and broad training objectives make direct deployment challenging in domain-specific, latency and resource-constrained scenarios such as road-user video detection. In this thesis we adopt a complementary perspective: instead of training a monolithic foundation model, we design mid-scale, modular components (STF, LAQEM, DAMM) that can be integrated into standard detectors, leverage pretrained vision–language features where appropriate, and explicitly expose temporal fusion and memory mechanisms tailored to the transportation setting.

2.4 Conclusion

Despite significant advancements in object detection, current methodologies often fall short when addressing real-world video challenges such as partial or complete occlusions, abrupt appearance changes, and low temporal coherence. Recent approaches incorporating transformer-based architectures and vision-language models have enhanced object detection capabilities but still frequently lack mechanisms for effective temporal memory and semantic generalization in dynamic scenarios. Consequently, there remains a clear gap in robust, temporally coherent object detection that effectively leverages multimodal semantic embeddings and memory mechanisms. Addressing these gaps forms the core motivation and objective of this thesis.

CHAPTER 3 OVERVIEW OF THE APPROACHES

Based upon the problems of occlusion, temporal instability, distractors, and generalisation identified in Chapter 1, this chapter introduces three complementary frameworks, each addressing specific limitations incrementally. Each of the three following chapters corresponds to a peer-reviewed published or submitted publication and can be read independently. However, the overall coherence of the research lies in the complementarity of the proposed methods and the incremental development of ideas aimed at overcoming the limitations of existing video object detection approaches.

3.1 First Contribution

This first article introduces the **Spatio-Temporal Fusion (STF)** framework, a novel architectural design aimed at improving object detection robustness in video streams. The development of STF was directly motivated by empirical failures observed in traffic surveillance videos from urban intersections. During our qualitative assessments, we consistently identified three key issues that traditional image-based detectors struggled with: (1) vehicles disappearing during sharp turns due to motion blur, (2) pedestrians becoming fragmented when partially occluded by poles or street furniture, and (3) persistent localization jitter during camera panning sequences.

These recurring detection failures revealed a fundamental limitation of single-frame models: the absence of temporal coherence. We initially hypothesized that adding a form of visual saliency could help stabilize detections across frames. To this end, we experimented with augmenting the YOLOX detector using Grad-CAM, leveraging the highlighted activation regions to reweight or reinforce detections in temporally adjacent frames. Specifically, we aimed to use Grad-CAM heatmaps to guide attention toward consistent object regions across time, hoping to improve robustness under occlusion or motion blur.

As illustrated in Figure 3.1, Grad-CAM successfully highlighted semantically meaningful regions such as vehicles or pedestrians. However, the activation maps were coarse and lacked the resolution and spatial precision necessary to accurately recover object centers or delineate tight bounding boxes. Consequently, this approach proved insufficient for precise temporal alignment, limiting its usefulness for downstream fusion tasks where consistent spatial anchoring is critical.

The limitations encountered with Grad-CAM led us to rethink our approach entirely. Given

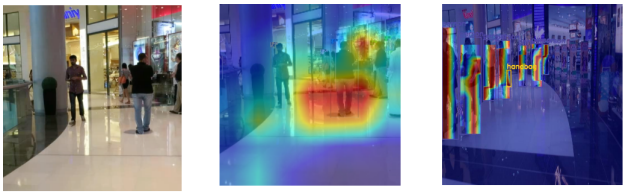

Figure 3.1 Limitation of Grad-CAM on YOLOX: highlights regions but lacks precise center probabilities.

Grad-CAM's coarse localization, we transitioned to a CenterNet [14]-inspired heatmap approach, explicitly modelling object centers to improve spatial-temporal alignment. This approach allowed us to treat detections as probability distributions over spatial locations, enabling soft, differentiable supervision and spatially-aware fusion.

The STF framework is structured around two complementary branches that extract different types of information from the video sequence. One branch focuses on consecutive frames to capture fine-grained motion cues and short-term temporal continuity. The second branch samples temporally distant yet semantically rich frames to preserve appearance consistency and mitigate the effects of occlusion or abrupt changes. The outputs from both branches are then combined using a learned weighting strategy that adapts to the alignment quality and scene dynamics, ensuring robust and temporally stable object detection. Further details are provided in Chapter 4.

We evaluated STF on standard video detection benchmarks including KITTI, Cityscapes, and UAVDT. The model consistently outperformed frame-level baselines in both AP and detection consistency metrics. Its robustness was particularly notable in scenarios involving motion blur, camera panning, and short-term occlusion, validating the effectiveness of its heatmap-driven temporal reasoning.

Nevertheless, STF has its limitations. Because its temporal context is confined to short frame offsets (e.g., t and $t-1$ or $t-3$), it struggles to recover objects that disappear for extended periods, such as those temporarily leaving the camera field of view or being completely occluded for more than a few frames. This issue was especially obvious in tunnel driving scenes and drone-based occlusion challenges, where targets failed to reappear without a long-term memory mechanism.

Definition of robustness and choice of datasets. In this first contribution, *robust* video object detection refers to a detector that maintains reliable performance when the

visual conditions are degraded or depart from the training distribution. In practice, this includes (i) short-term occlusions and partial visibility, (ii) motion blur and camera egomotion (e.g., panning or shaking), (iii) cluttered scenes with nearby distractors, and (iv) variations in scale and viewpoint for road users. The three benchmarks used for STF, namely KITTI MOT, Cityscapes and UAVDT, were selected because they specifically exhibit these factors in complementary ways. KITTI MOT and Cityscapes contain urban driving scenes with frequent occlusions, sudden motion and strong perspective changes, while UAVDT emphasizes small, distant vehicles, dense traffic and aerial viewpoints. The gains obtained by STF over frame-level baselines on these datasets therefore directly reflect improved robustness in this sense, and not only higher accuracy on clean or static images.

Added value of STF beyond the HRNet backbone. The HRNet backbone used in Chapter 4 is a strong per-frame feature extractor, but by construction it processes each frame independently and does not exploit temporal coherence. STF is precisely the component that turns this per-frame backbone into a temporally aware detector. The Multi-Frame Attention (MFA) module learns adaptive temporal weights over past and current frames so that the network can lean more heavily on frames that are less affected by blur or partial occlusion. The Single-Frame Attention (SFA) module then refines spatial and channel information within the current frame, suppressing background clutter and reinforcing object regions. Finally, the dual-frame fusion module combines these feature maps across scales in a learnable way instead of using fixed concatenation or averaging. The ablation study in Table 5.5 quantifies this added value: the baseline HRNet+CenterNet head reaches 92.10% mAP on Cityscapes; adding SFA increases this to 93.50%, adding MFA alone to 94.91%, and combining MFA+SFA with the dual-frame fusion module yields 95.73%. This shows that the improvement is not due to HRNet alone, but to the STF modules that explicitly model temporal and spatial attention.

Relation to the CenterNet detection head. The detection head used with STF is intentionally close to CenterNet to keep the comparison fair and to isolate the effect of temporal fusion. However, it is adapted to the video setting. In CenterNet, both the heatmap and the bounding box regressors operate on the same single-frame feature map. In STF, the fused object heatmap is predicted from the temporally aggregated feature map, while the box size and offsets are regressed from the current-frame features only. This decouples object presence decisions (which benefit from temporal context) from precise geometry (which must remain anchored to the current image). This modification is specific to the robustness objective of STF and goes beyond a straightforward reuse of the original CenterNet head.

Model capacity and fairness of comparisons. To keep the comparisons in Tables 5.1–5.3 as fair as possible, all detectors are trained under the same protocol (optimizer, learning rate schedule, number of epochs and data splits). STF reuses the same HRNet backbone and adds relatively lightweight attention and fusion layers; most of the parameters remain in the backbone, and the additional overhead is modest compared to widely used detectors such as YOLOv5, YOLOX or two-frame methods such as PPNet. The observed gains are therefore better explained by the way STF exploits temporal and spatial information than by a disproportionate increase in model size.

3.2 Second Contribution

While STF substantially improves local temporal robustness, it still cannot fully recover objects that remain invisible for longer intervals or that belong to rare or unseen categories. In sequences with long occlusions, heavy clutter, or unusual road users (e.g., construction vehicles or cargo bikes), we observed that purely appearance-based temporal fusion frequently failed to re-establish a stable track once the object reappeared. These failure cases motivated a second contribution that injects higher-level semantic priors and an explicit long-term memory into the detection process.

To address these limitations, we proposed **Language-Augmented Queries with Evolving Memory (LAQEM)**, which encodes persistent, semantically enriched object representations using CLIP-based vision–language embeddings and a dynamic query memory module. LAQEM augments the standard detector queries with three ingredients: (1) text-conditioned semantic anchors derived from CLIP, (2) a memory bank that stores high-quality per-frame embeddings, and (3) an evolving update policy that selects, filters, and refreshes memory entries over time. At each frame, current visual features are compared against this memory to retrieve consistent object-level representations, which are then injected back into the transformer decoder as enriched queries.

This design allows LAQEM to re-identify objects after extended occlusion, to maintain category-consistent predictions under appearance shifts, and to improve recall on rare or unseen categories. Empirically, LAQEM achieves strong generalization and long-term consistency on challenging benchmarks such as VisDrone, Cityscapes, UAVDT, and UA-DETRAC, with particularly clear gains on rare classes and sequences with frequent disappearances and re-entries. Further architectural and experimental details are provided in Chapter 5.

Robustness to long-term occlusions and rare categories. In the context of the second contribution, robustness primarily refers to the ability to maintain consistent identities

and category labels over longer temporal horizons and under semantic shifts. Concretely, LAQEM is designed to handle (i) objects that disappear for dozens of frames due to strong occlusions or leaving the field of view, (ii) road users with unusual appearance or rare categories (e.g., service vehicles, trailers, cargo bikes), and (iii) drift in the visual signature of an object across time. The VisDrone, Cityscapes, UAVDT and UA-DETRAC benchmarks were chosen because they jointly exhibit these phenomena: long and cluttered sequences, highly imbalanced class distributions, and frequent disappearances and re-entries. The gains brought by LAQEM over its baselines on these datasets therefore directly measure robustness in this long-term, semantic sense rather than only per-frame accuracy.

Embedding filter and memory design. A central component of LAQEM is the *embedding filter*, which decides which frame-level embeddings should be written to memory. The goal is to keep a compact memory that stores only informative, stable object representations. At each frame, candidate embeddings are first ranked according to their cosine similarity with CLIP text anchors and their detector confidence. Only the top-k candidates that exceed a similarity threshold are considered for insertion. Among these, the filter discards near-duplicates that are already well represented in memory (based on cosine distance) and keeps embeddings that either improve the coverage of rare classes or reduce uncertainty for existing tracks. In Chapter 5, we detail the choice of k, similarity thresholds and memory size, and show through ablations that the final configuration balances three factors: performance gains, memory stability over long sequences, and computational cost.

On the DETR baseline and absolute AP values. In the LAQEM experiments, transformer-based baselines such as DETR obtain lower AP than what is typically reported on large benchmarks like COCO. This behaviour is expected in our setting for two reasons. First, the traffic datasets we use are smaller, more imbalanced, and visually quite different from COCO, which makes convergence harder without additional pretraining. Second, for fairness, all baselines and LAQEM variants are trained under the same schedule and data regime, without any DETR-specific tuning or extra data. Under these constraints, DETR serves as a representative transformer detector rather than an upper bound on achievable AP. The consistent improvements of LAQEM over this baseline are therefore more informative than the absolute AP values themselves, and they confirm that the proposed semantic memory is beneficial even in this challenging regime.

3.3 Third Contribution

Building explicitly on STF's short-range temporal stability and LAQEM's semantic memory, the third contribution targets a different failure mode: scenes that are extremely cluttered, with many look-alike objects and ambiguous boundaries. In dense intersections or drone footage over highways, we observed that even with semantic memory, detectors could still confuse nearby instances, produce duplicate boxes, or drift toward background structures with similar texture.

To mitigate these issues, we introduce **DAMM (Dual-Stream Attention with Multi-Modal Queries)**, which refines detector queries using both semantic and spatial cues before they enter the main transformer decoder. DAMM capitalizes on three kinds of queries: appearance-based queries from a vision–language model, polygon-based positional queries derived from instance masks (e.g., SAM polygons), and generic learnable queries for background and context coverage. These queries are processed through two coupled attention streams. A semantic stream focuses on aligning each query with language-conditioned appearance features, while a spatial stream emphasizes geometric layout, using polygonal embeddings to preserve object extent and separation.

By jointly ranking and filtering queries across these two streams, DAMM selects a compact set of context-consistent, non-redundant queries that better respect object boundaries and reduce interference between nearby instances. This leads to improved localization in crowded scenes, fewer false positives on distractors, and better transfer to previously unseen object types in transportation scenarios. Chapter 6 provides the full architectural description and experimental analysis.

It is important to note that each article progressively builds on the insights and limitations of the previous one:

- The STF module establishes a temporal fusion mechanism that improves local robustness across frames.
- LAQEM extends this approach by incorporating semantic memory for long-term occlusion recovery and better generalisation to rare or unseen categories.
- DAMM complements the architecture by refining and filtering multimodal queries to ensure relevance and adaptability in cluttered and open-world scenarios.

Together, these three contributions form a coherent framework for robust video object detection. They respond incrementally and effectively to the major challenges outlined in the

problem statement, namely, occlusion, distractors, motion, generalization, and real-time feasibility.

Robustness in cluttered and open-world scenes. For the third contribution, robustness refers to reliable detection and localization in scenes that are both densely populated and open-world. Typical failure modes include (i) confusion between neighbouring instances with similar appearance (e.g., adjacent vehicles of the same type), (ii) drift of queries toward background structures such as poles, barriers or building edges, and (iii) loss of previously unseen or rare object types when the scene becomes crowded. DAMM addresses these aspects by jointly ranking queries according to semantic consistency (vision–language alignment) and spatial plausibility (polygon-based geometry), so that only a compact set of non-redundant and well-separated queries is fed to the main detector. The resulting improvements in crowded intersections and highway scenes are therefore a direct measure of robustness to clutter and distractors.

Interpreting low absolute scores in open-world experiments. In Chapter 6, some of the metrics reported in Table 6.1 are relatively low in absolute value for all compared methods. This reflects the difficulty of the evaluation setting rather than an implementation issue. The experiments combine several challenging factors at once: small and partially occluded road users, long sequences with frequent re-entries, and in some cases open-vocabulary or cross-dataset testing where the distribution at test time differs from the training labels. Under such conditions, even strong baselines suffer a drop in AP. DAMM is thus evaluated primarily in terms of *relative* gains over these baselines. The consistent improvements it brings, especially on crowded and open-world scenarios, indicate that the dual-stream query refinement is effective despite the low absolute numbers.

Positioning with respect to video foundation models. Recent video foundation models, such as large-scale architectures that jointly encode space, time and semantics, implicitly integrate temporal coherence and semantic reasoning within a single model. The contributions in this thesis, including DAMM, take a complementary and modular perspective: they are designed as plug-in components that can sit on top of existing detectors or vision–language backbones. In particular, DAMM can be interpreted as a lightweight query filtering and refinement layer that could, in principle, be combined with future video foundation models to improve their behaviour in dense traffic scenes. This clarifies the scope of the work: the goal is not to replace foundation models, but to provide practical modules that offer explicit control over temporal and multimodal fusion in safety-critical video detection settings.

CHAPTER 4 ARTICLE 1: SPATIO-TEMPORAL FUSION FOR VIDEO OBJECT DETECTION

Full Reference:
Noreen Anwar, Guillaume-Alexandre Bilodeau, and Wassim Bouachir. "STF: Spatio-Temporal Fusion for Robust Video Object Detection." In Proceedings of the 21st Conference on Robots and Vision (CRV), Guelph, Canada, May 2024.Publication date : February 16,2024. Published.

Statement of Contribution:
This article was co-authored with Professors Guillaume-Alexandre Bilodeau and Wassim Bouachir.

Noreen Anwar was primarily responsible for conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing – original draft, and visualization.

Professor Guillaume-Alexandre Bilodeau contributed to supervision, writing, review and editing, methodology, and project administration. He provided continuous research supervision, critical feedback on the methodology, and contributed to the structuring and proofreading of the paper.

Professor Wassim Bouachir contributed to writing, review and editing, supervision, and validation. He supported the scientific validation of results and advised on model design decisions.

All authors contributed to the revision and improvement of the final manuscript.

This article was accepted and presented at the 21st Conference on Computer and Robot Vision (CRV 2024), held in Guelph, Ontario, Canada.

4.1 Abstract

Consecutive frames in a video contain redundancy, but they may also contain relevant complementary information for the detection task. The objective of our work is to leverage this complementary information to improve detection. Therefore, we propose a spatio-temporal fusion framework (STF). We first introduce multi-frame and single-frame attention modules that allow a neural network to share feature maps between nearby frames to obtain

more robust object representations. Second, we introduce a dual-frame fusion module that merges feature maps in a learnable manner to improve them. Our evaluation is conducted on three different benchmarks, including video sequences of moving road users. The performed experiments demonstrate that the proposed spatio-temporal fusion module leads to improved detection performance compared to baseline object detectors. Code is available at `https://github.com/noreenanwar/STF-module`

4.2 Introduction

Recent advances in computer vision have significantly improved the capability of object detection systems, especially in single-frame localization and classification tasks [35, 60]. Despite this progress, relying solely on single-frame approaches is often inadequate in practical scenarios involving occlusions, motion blur, or small-scale objects, all of which compromise object visibility and feature distinctiveness [53, 56–59, 77]. Single-frame object detectors are subject to errors in the case of poor or improper visibility of objects that can be caused by occlusions, motion blur, or small object sizes. When objects are occluded or in the case of motion blur, their appearance features can be severely altered. The object detector should be robust to the fact that objects can exist on a spectrum of scales and sizes. Furthermore, small objects have less distinctive features making them harder to detect.

To overcome these limitations, researchers have increasingly focused on leveraging multiple frames from video sequences to enrich object representation and improve detection performance [53, 56–59, 77–79]. Historically, methods employing multiple frames have utilized heuristic post-processing techniques, such as bounding box integration through optical flow or motion estimation, applied after single-frame detection [78, 79]. Although such methods have improved detection consistency over time, these methods fail to fully exploit the inherent temporal continuity present in sequential frames, resulting in suboptimal handling of degraded object features.

Recent advances have shifted toward more sophisticated methods that fuse features across multiple frames in an end-to-end learnable manner [53, 56, 57, 59, 77]. However, this approach is inherently challenging due to misalignment issues caused by object movements or visibility changes across consecutive frames. As a result, effective feature fusion demands careful design and a robust mechanism to adaptively select and combine the most relevant features from each frame.

In this chapter, we introduce a *Spatio-Temporal Fusion* (STF) framework for video object detection that addresses these challenges. The key idea is to improve detection by combining

features from the current frame with those from a nearby past frame in a learnable way. Our STF framework consists of two attention-based modules and a fusion mechanism. First, a **Multi-Frame Attention (MFA)** module learns to highlight relevant features across two consecutive frames, allowing the model to capture temporal context and handle occlusion or motion blur. Second, a **Single-Frame Attention (SFA)** module focuses on the current frame alone, enhancing important spatial regions and feature channels to reduce false positives from background clutter. Finally, a **Dual-Frame Fusion** module merges the outputs of MFA and SFA, producing a unified feature representation that benefits from both temporal cues and single-frame refinement. By leveraging spatio-temporal cues in this manner, STF aims to produce more robust object representations and improved detection accuracy compared to single-frame detectors.

To evaluate our approach rigorously, we conducted experiments using three widely recognized traffic-related datasets: KITTI MOT [80], Cityscapes [81], and UAVDT [82]. Our results demonstrate significant performance improvements over existing single-frame and multi-frame object detection methods, affirming the effectiveness and practical relevance of the STF framework.

Leveraging multiple frames for object detection has been explored extensively due to its potential to enhance detection precision and robustness by exploiting spatial-temporal correlations. While single-frame object detection methods have been the primary focus of much computer vision research, multi-frame detection methods remain relatively less investigated despite their significant practical importance in areas such as video surveillance, autonomous navigation, and intelligent driving systems. Sequential video frames inherently contain complementary information about object instances, offering valuable insights that can improve detection performance, especially in challenging scenarios involving occlusions or rapid object movement.

Early methods utilizing multiple frames predominantly adopted heuristic, box-level post-processing techniques. For instance, Kang et al. [78] and Lee et al. [79] initially applied single-frame detectors independently and subsequently integrated the resulting bounding boxes across frames using standard off-the-shelf motion estimation methods. Although these approaches improved temporal consistency, they relied heavily on manually crafted rules and heuristic post-processing steps without incorporating learnable components, thereby limiting their adaptability and accuracy in complex real-world environments.

In contrast, recent approaches have aimed to integrate temporal context more systematically and end-to-end through deep learning architectures. Zhu et al. [53] introduced Flow-Guided Feature Aggregation (FGFA), which utilizes optical flow to warp and align feature maps

from adjacent frames, enabling improved feature representation for detection tasks. Similarly, Bertasius et al. [58] proposed an approach to calculate spatial offsets between temporally adjacent frames, effectively sharing relevant features across frames to enhance detection accuracy.

A further advancement was achieved by methods that employed fully learnable feature fusion architectures. For example, Perreault et al. [56] proposed the Feature Fusion Architecture for Video Object Detection (FFAVOD), a method designed to learn the optimal fusion of feature maps across adjacent frames, thereby directly enhancing detection and classification performance. Likewise, RN-VID [55] proposed merging feature maps from neighboring frames through channel-wise convolution and rearrangement to reinforce object representation.

Zhou et al. [57] developed CenterTrack, which introduced a point-based tracking framework capable of simultaneously detecting and tracking objects across video frames. CenterTrack concatenates information from two consecutive frames along with a prior heatmap, achieving temporal association concurrently with detection. This method significantly improves robustness in dynamic video scenes.

Explorations into integrating temporal memory through recurrent architectures have also proven beneficial. Xiao et al. [83] introduced a Spatio-Temporal Memory Module (STMM), employing a recurrent neural network to aggregate spatial-temporal information across multiple frames, thus enriching detection features. Liu et al. [77] leveraged Long Short-Term Memory (LSTM) networks to interpolate features temporally, significantly enhancing inference speed. Similarly, Broad et al. [59] proposed the Multi-frame Single Shot Detector (MF-SSD), integrating a recurrent convolutional module to fuse spatial and temporal features across multiple frames effectively.

Despite significant advances, current temporal fusion strategies predominantly utilize straightforward aggregation techniques such as concatenation or summation of feature maps, lacking a sophisticated, fully learnable integration mechanism. To address this limitation, our work introduces a novel approach focusing on a fully learnable fusion-based module. Our method adaptively integrates temporal, spatial, and channel-based information from current and previous frames in an end-to-end manner, thereby overcoming the inherent limitations of existing feature fusion strategies and substantially enhancing object detection performance.

4.3 Methodology

4.3.1 Overview

The overview of our attention-based framework, STF, is shown in Figure 5.2. Given a pair of frames, a pre-trained HRNet [84], where we froze the first and third layers, is used to extract features. After that, the features go through two attention modules: 1) a multi-frame attention (MFA) module that uses the two extracted feature maps to perform temporal and spatial attention, assigning adaptive temporal weights to them, and 2) a single-frame attention (SFA) module that uses spatial and channel dimension attention for improving current frame feature maps. To use the temporally prior frame, the idea here is to combine in a learnable manner the extracted features of the past and current frames for object detection. To combine features from two frames after applying attention, our proposed network fuses temporal, channel, and spatial information by aggregating them at the same time with our dual-frame fusion module. In the following, we introduce these modules in detail.

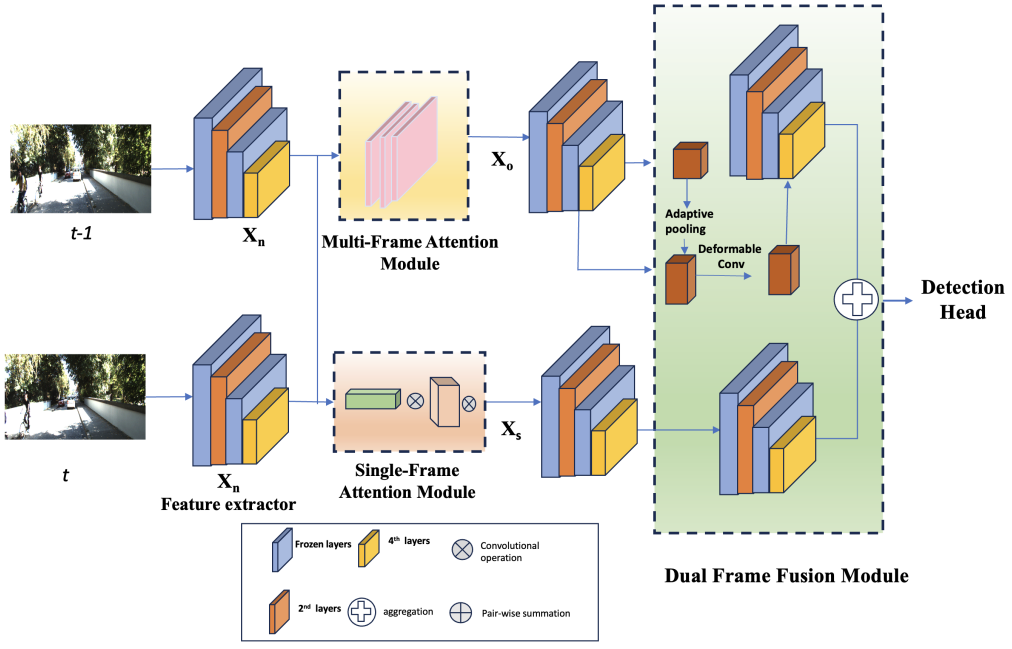

Figure 4.1 Overview of our spatio-temporal based fusion framework (STF), illustrating the key components: MFA, SFA, and dual-fusion module

4.3.2 Multi-Frame Attention (MFA) module

Given an input video, the multi-scale feature maps of two frames (the current frame and a past frame) are extracted with the HRNet backbone. Then, our goal is to merge the features

of these two frames. The Tada Convolution, introduced in the work by Huang et al. [85], efficiently addresses temporal modeling by introducing flexibility to the temporal invariance of 2D convolutions. This is achieved through the incorporation of adaptive temporal weights, which are superimposed onto the convolutional process. Similarly, Cao et al. proposed TCTrack [86], which exemplifies the application of Tada Convolution for improving object tracking. This approach employed Tada Convolutions to incorporate adaptive temporal weights, contributing to improved temporal modeling. Inspired by this previous research, to get adaptive temporal weights for each frame, we designed a Multi-Frame Attention (MFA) module (see Figure 4.2). The key idea is to adjust the model behavior in real-time as it processes each sequence of frames. This deals with size variations, movement, overlapping, or interaction of objects in frames.

Global information in object detection refers to semantic details that are consistent across frames, helping in identifying objects based on shared characteristics, while local temporal information involves using nearby frames to gather information, such as motion, helping to localize objects, especially in cases of uncertainty about their existence in a specific frame. This module improves the representation ability with multi-frame features by: 1) assigning adaptable weights to each frame to enhance the ability to detect and analyze changes over time, 2) combining both global and local information from multi-frames, and 3) better capturing both detail and broader spatial and temporal information using a multi-scale integrator.

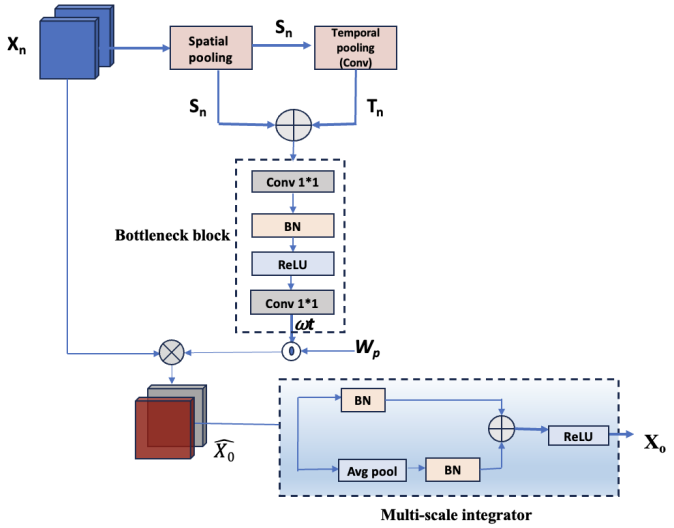

Figure 4.2 Multi-Frame Attention module with multi-scale integrator.

Our MFA module works as follows. Let us assume that we have an input sequence of frames I_n and we get a sequence $X_n \in \mathbb{R}^{B\times C\times T\times H\times W}$ of features outputted by the HRNet backbone,

where B is the batch size, C is the number of channels, T is the temporal dimension, and H and W are the spatial dimensions. For capturing the global spatial context, we start with global average pooling (GAP) across the spatial dimension of the past and current frame features. We then obtain frame descriptor $S_n = \mathrm{GAP}_S(X_n)$, that encompasses global spatial context. To integrate local temporal context effectively, global average pooling across both spatial and temporal dimensions is applied to obtain spatio-temporal descriptor

$$T_n = \mathrm{GAP}_{st}(X_n). \tag{4.1}$$

Global spatial context and local temporal information are then aggregated, and this combined information is passed through a bottleneck block (BNB). The output of the bottleneck block results in obtaining local weights ω_t, as illustrated in Figure 4.2. These weights combine the spatial and temporal descriptors after the bottleneck block with

$$\omega_t = \mathrm{BNB}(S_n + T_n). \tag{4.2}$$

Then, the total weights that we used in our model are the element-wise product of these weights ω_t and weights W_p that refer to the initial set of weights in the convolution kernel that is shared across all frames. Note that the local weights ω_t are set to 0 during initialization, which has the advantage of reducing the training time. An adaptive convolution is then applied to the current frame with

$$\hat{X}_o = (\omega_t \odot W_p) * \ X_n \tag{4.3}$$

where $\odot$ denotes element-wise multiplication.

To effectively integrate spatio-temporal information and address the limitations in spatial features for a given frame, we finally apply a multi-scale integrator as shown in Figure 4.2. It is expressed as

$$X_o = \lambda(\hat{X}_o) + \gamma(\mathrm{AvgPool}(\hat{X}_o)), \tag{4.4}$$

where $\hat{X}_o$ is the output from the adaptive convolution . The operators λ and γ represent distinct normalization functions. The goal behind using an average pooling (AP) layer is to enlarge the receptive field to capture a wider range of spatial contexts.

4.3.3 Single-Frame Attention Module

Besides temporal attention, attention in the spatial and channel dimensions also provides a potential enhancement for feature maps derived from single-frame images. In the context of Convolutional Neural Networks (CNNs), the attention mechanism assigns an additional weight to individual pixels in a specific dimension, indicating the significance of particular information. These learned weights strengthen valuable features and weaken less useful ones, facilitating feature screening and enhancement. Furthermore, in videos with generally stable backgrounds, spatial and channel attention, as explained by the methodology proposed in Hou et al. [87], can efficiently suppress false positive detection in the background area.

Inspired by this work, we propose a Single-Frame Attention module (SFA) that uses channel and spatial attention mechanisms, as illustrated in Figure 4.3. The SFA module aims to refine feature representation within a single frame. In the SFA module, each frame denoted as I_n, is processed to enhance the channel and spatial information of its feature maps X_n. First, channel attention with average pooling (AP) and max pooling (MP) are applied to condense the spatial information. To help our model learn complex feature representation, we integrate 1×1 convolutional layer as shown in Figure 4.3. This results in channel attention A_c formulated as

$$A_c = Conv_{1\times1}(AP(X_n)) + (MP(X_n)). \tag{4.5}$$

For spatial attention, a comparable approach is applied, but it operates within the spatial domain. Here, the features influenced by channel attention are subjected to average and max pooling operations (MP), focusing on spatial features. The resulting features are then concatenated and processed through a 5×5 convolutional layer to enhance the spatial aspects of the frame. This gives spatial attention A_s formulated as:

$$\begin{aligned} A_s = \mathrm{Conv}_{5\times5}\Big(&\mathrm{Conv}_{1\times1}(\mathrm{AP}(A_c * X_n)) \\ &+ (\mathrm{MP}(A_c * X_n))\Big) \end{aligned} \tag{4.6}$$

where $Conv_{5\times5}$ is the convolution operation using a 5×5 filter and $*$ symbolizes convolution.

Finally, the two attention tensors are concatenated with X_n to obtain the new features X_s. This fusion process allows the model to focus on relevant information captured by the attention mechanisms, enhancing the representation of the feature maps. We observed that by using convolutional layers, the module can more effectively capture and enhance the intricate patterns in the features. This ensures that the model is capturing well the spatial features in each frame.

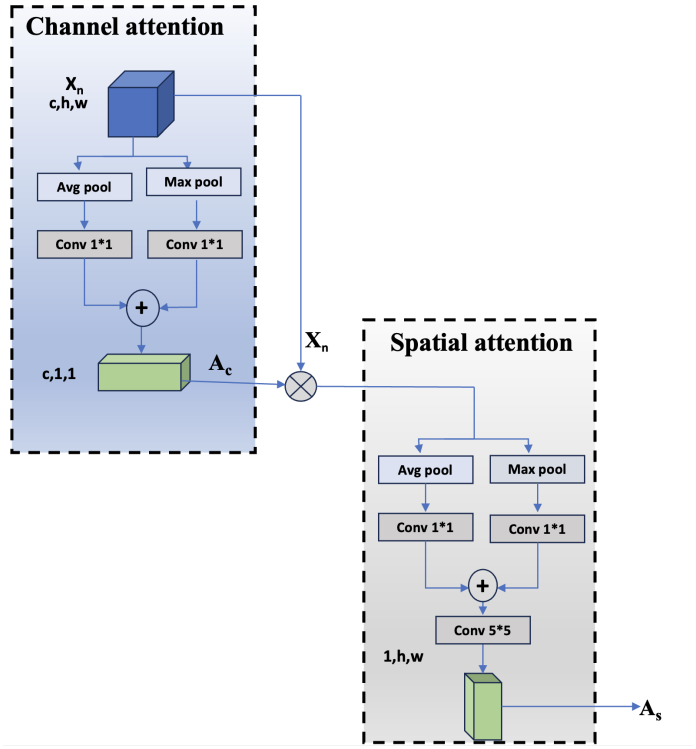

Figure 4.3 The channel and spatial attention modules of our proposed single-frame attention module

4.3.4 Dual-Frame Fusion Module

Figure 5.2 illustrates the feature maps X_o and X_s obtained after the SFA and MFA modules, which serve as inputs to our dual-frame fusion module. The proposed dual-frame fusion module combines semantic information of the high-level feature maps and spatial information of low-level feature maps. Instead of traditional up-sampling, inspired by [88], we use Adaptive Feature Pooling for a more flexible approach. This offers an expanded receptive field, facilitating improved integration of both core and contextual semantics.

The high-level feature map is adaptively pooled to match the size of the low-level feature maps. These feature maps are then combined via pixel-wise summation and further processed through deformable convolutions. This offers better adaption to different object sizes, shapes, and other geometric deformations. The input has a total of four layers, and the aforementioned convolution and up-sampling process is iterated 2 times to have the final output. With the help of the above process, we obtained channel and spatial attention feature maps on single frames and temporal attention feature maps for multiple frames. These feature maps are aggregated to generate a fused feature map.

4.3.5 Detection Head

Our detection head is similar to CenterNet [14]. We performed the computation of the fused object probability heatmap on the merged feature maps. However, the size and offset of the bounding boxes are generated from single-frame features. The loss function comprises three components: a fusion heatmap loss calculated with Focal Loss, and two regression losses (offset and size) computed with L1 Loss. The formulation of each loss is as follows. L_Z is the unique fusion heatmap loss,

$$L_Z = -\frac{1}{M}\sum_{ij}\begin{cases}(1-\hat{Q}_{ij})^{\epsilon}\log(\hat{Q}_{ij}) & \text{if } Q_{ij}=1\\ (1-Q_{ij})^{\zeta}(\hat{Q}_{ij})^{\epsilon}\log(1-\hat{Q}_{ij}) & \text{otherwise}\end{cases}, \tag{4.7}$$

where $\hat{Q}_{ij}$ indicates the predicted heatmap value for each pixel, $Q_{ij}=1$ signifies the pixel is the center of an object, and ϵ and ζ are the modified focal loss hyper-parameters. L_Y represents the loss for heatmap offset,

$$L_Y = \frac{1}{M}\sum_{q}|\hat{P}_q - T - \tilde{q}|, \tag{4.8}$$

where $\hat{P}_q$ is the predicted offset, T is the position after downsampling, and $\tilde{q}$ is the actual center point. L_X calculates the loss for the size of the bounding box,

$$L_X = \frac{1}{J}\sum_{j=1}^{J}|\hat{R}_j - R_j|, \tag{4.9}$$

where $\hat{R}_j$ is the predicted size and R_j is the ground truth size. The overall training objective is

$$L_{total} = L_Z + \lambda_{dim}L_X + \lambda_{pos}L_Y, \tag{4.10}$$

where λ_{dim} and λ_{pos} are the adjusted hyper-parameters for the size and offset loss components, respectively.

4.4 Experiments

In this section, we assess the performance of our proposed method compared to SOTA methods and perform an ablation study.

4.4.1 Datasets and Evaluation Metrics

Datasets: As our method relies on more than a frame, the evaluation requires the use of video datasets. Our selected evaluation domain focuses on traffic surveillance given its significant relevance to our research. We used datasets with videos, but some are not standard datasets for object detection. Nevertheless, they were used in previous work on video object detection. We chose: KITTI MOT (Multi-Object Tracking) [80] and Cityscapes [81], both not used for object detection usually but provide videos, and UAVDT [82] used for object detection in videos. Each of these datasets provides unique challenges and contains sequences at different viewpoints with different sizes of objects. As we are using non-standard datasets (KITTI, Cityscapes) for object detection, we needed to compute some results ourselves for competing SOTA methods for a fair comparison. However, this is not true for the UAVDT dataset, where we use the standard data training and test split.

Evaluation Metrics: We use Average Precision (AP) for multiple scales of objects and Mean Average Precision (mAP) across varying IoU thresholds and mAP50 and mAP75, respectively at 0.5 and 0.75 IoU thresholds, to evaluate detection accuracy. Intersection over Union (IoU) is used to evaluate bounding box precision on all datasets.

4.4.2 Implementation Details

For features extraction, we used HRNet [84], and pre-trained it on the COCO dataset [89], following the methodology described in [14]. Our global architecture follows CenterNet [14]. However, our training process is done in two steps. First, our backbone is fine-tuned on each dataset starting from the pre-trained weights on COCO. Then, the first and third layers of the backbone are frozen and the MFA, SFA, and dual-fusion modules as well as the network heads are trained. Training is conducted over 250 epochs utilizing the Adam optimizer, starting with a learning rate of 1×10^{-4}, which undergoes a decimation by a factor of 10 after the 130^{th} and 140^{th} epochs. To ensure training stability, we use gradient clipping. The same training protocol was used for the overall architecture as well as for all the base detectors to demonstrate the contribution of our approach.

4.4.3 Results and Discussion

Comparisons with SOTA methods on the Cityscapes dataset are reported in Table 5.1.

They show that our attention-based fusion detector consistently outperforms the other SOTA detectors. There is a significant improvement in the detection results when using our STF model as compared to SOTA detectors. The improvement in detection results is due to our

two attention modules and our dual-fusion module, all contributing positively to detecting objects better (especially small or occluded ones). In Table 5.1, we also compare our model with the vanilla HRNet as we use a feature extractor based on the HRNet architecture. This allows us to examine our results in comparison to vanilla HRNet to observe the impact of our STF module on a similar backbone. This comparison demonstrates a gain in accuracy for all sizes of objects. Furthermore, we changed the backbone of Centernet [14] to observe how HRNet affects its performance as it uses a detection similar as ours. It can be concluded from the results that using HRNet alone does not yield significant improvement. This is another demonstration that our method using a classification head similar to CenterNet performs better due to our SFA and MFA modules. By comparing our results with YOLOv5 and the recent YOLOX, our model shows improvement in terms of precision and accuracy, as well. Finally, we also perform better than PPNet which uses multiple frames.

Table 4.1 Comparison of our method with SOTA methods on the Cityscapes validation dataset. **Boldface** indicates best results.

Method Type	**Method**	**Backbone**	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SOTA (Single Frame Detectors)	RetinaNet [90]	RetinaNet-50	92.8	94.1	93.2	43.1	58.2	94.0
	Vanilla HRNet*	HRNet	92.2	94.9	91.1	45.3	59.9	93.1
	CenterNet [90]	Hourglass-104	92.1	93.9	93.0	43.2	56.7	93.5
	CenterNet*	Resnet-18	92.7	92.5	92.7	44.8	57.6	93.2
	CenterNet*	HRNet	92.8	93.3	93.1	45.9	58.7	93.4
	YOLOv5 [90]	CSPDarknet53	93.6	93.4	91.8	43.7	59.5	95.1
	YOLOX [90]	CSPDarknet53	93.9	94.9	92.7	44.8	61.5	96.7
SOTA (Two Frame Detectors)	PPNet [90]	Resnet-50	94.8	96.2	92.5	43.9	57.4	95.8
	STF (Ours)	HRNet	**95.7**	**97.2**	**95.3**	**49.3**	**65.3**	**97.3**

*Trained by ourselves.

Table 5.2 presents the results of the KITTI validation dataset. The conclusions are the same as for Cityscapes with similar improvements compared to baseline methods. By comparing it with other SOTA detectors, our proposed method outperforms them with improvements for all object size categories (small, medium, and large). Our method demonstrates an improvement in detection results when compared to SOTA single-frame and two-frame detectors.

Results on the UAVDT test dataset are reported in Table 5.3. Our Spatio-Temporal Fusion (STF) module consistently outperforms the base detectors. Also, when compared to SOTA multi-frame detectors, such as FFAVOD and RN-VID that fuse features without attention, we can notice that although this helps compared to single-frame detectors, a more sophisticated fusion approach, like the one we propose, is required to obtain even better results.

Table 4.2 Comparison of our method with SOTA methods on the KITTI MOT validation dataset. **Boldface** indicates the best result.

Method Type	**Method**	**Backbone**	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SOTA methods (Single Frame Detectors)	RetinaNet [91]	RetinaNet-50	56.6	-	-	29.9	62.8	73.1
	Vanilla HRNet*	HRNet	79.1	81.6	70.0	48.3	65.2	74.3
	CenterNet [92]	Hourglass-104	-	85.3	-	-	-	-
	CenterNet*	Resnet-18	80.5	83.4	74.5	50.2	66.8	78.7
	CenterNet*	HRNet	81.7	83.3	74.1	50.0	66.8	77.4
	YOLOv5*	CSPDarknet53	84.3	86.8	76.3	52.9	70.4	83.5
	YOLOX*	CSPDarknet53	85.9	87.7	79.8	53.8	71.7	84.9
SOTA methods (Two Frame Detectors)	Mf-SSD [59]	SqueezeNet	83.0	-	-	-	-	-
	MFCN [93]	ResNet101	84.6	-	-	-	-	-
	PPNet [90]	ResNet50	86.2	-	-	-	-	-
	STF (Ours)	HRNet	**88.7**	**90.0**	**82.9**	**57.1**	**74.6**	**88.1**

*Trained by ourselves.

Table 4.3 Comparison of UAVDT test dataset with different methods. **Boldface** indicates the best result.

Method Type	**Method**	**Backbone**	mAP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
SOTA methods (Single Frame Detectors)	RetinaNet [94]	RetinaNet-50	16.2	34.0	13.7	8.8	30.1	23.8
	CenterNet [95]	Hourglass-104	16.4	29.7	16.6	12.2	25.1	11.3
	YOLOv5 [95]	CSPDarknet53	18.0	33.5	17.2	11.0	29.6	37.5
	YOLOX [96]	CSPDarknet53	26.0	43.3	21.4	-	-	-
	SpotNet [97]	U-Net	53.4	-	-	-	-	-
SOTA methods (Two Frame Detectors)	STDnet-ST [98]	RCN	13.3	36.4	-	-	-	-
	AdNet-MS [94]	Darknet53	13.3	43.5	18.3	12.1	37.9	27.9
	RN-VID [55]	-	39.4	-	-	-	-	-
	FFAVOD-SpotNet [56]	-	53.8	-	-	-	-	-
	FFAVOD-CenterNet [56]	-	52.1	-	-	-	-	-
	STF (Ours)	HRNet	**58.4**	**79.5**	**46.3**	**35.8**	**59.4**	**61.9**

4.4.4 Ablation Study

An ablation study was performed to evaluate the contribution of the different parts of the proposed method: the multi-frame attention module, single-frame attention module as well as the single-frame and multi-frame attention with the dual-frame fusion module, and show the effect of each component in Table 5.5. We find that the method with the MFA module or the SFA module detects better than the baseline method (HRNet + CenterNet head). We also observe that the proposed method (STF) with both two modules and dual-frame fusion performs the best. According to the proposed STF module, our MFA module plays a crucial role in combining features from two frames. Similarly, the SFA module aims to improve the accuracy of detection within a single frame. This is achieved through the combination of single-frame channel and spatial attention, which effectively suppresses false positive detec-

Table 4.4 Ablation study on the MFA, SFA, and Dual-fusion modules

Configuration	**mAP (%)**
Baseline (HRNet+CenterNet head)	92.10
Baseline + SFA	93.50
Baseline + MFA	94.91
Baseline (MFA+SFA)	**95.73**

tion in background regions. In our observations, we noted that each module independently contributes to performance enhancement. Moreover, a synergistic effect is observed when both modules are combined, leading to a more significant improvement in results. Therefore, for better efficiency and accuracy, our proposed model demonstrates superior results as compared to other configurations.

To illustrate the specific contributions of our proposed dual-frame fusion method, we also conducted an ablation study on it. We aimed to understand the individual impact of different fusion strategies on the overall performance of our model. For that, we use different strategies of combining two frames, i.e. concatenation, median, mean, and max fusion. In all cases, that decreased the performance by a large margin as shown in Table **??**. We attribute this to the misalignment of features across frames, necessitating a more intricate operation for aggregating these features. Admittedly, our model requires additional parameters to effectively learn the optimal combination of feature maps. However, as indicated in Table 4.5, our findings strongly support the benefit of our dual-frame fusion method in integrating feature maps.

Table 4.5 Ablation study of the different fusion strategies on Cityscapes dataset.

Fusion Methods	**mAP**
Concatenation	88.60
Median	91.50
Mean	91.70
Max	91.89
Dual-frame fusion (Ours)	**95.73**

4.5 Conclusion

In this work, we designed a spatio-temporal fusion module as a new approach for multi-frame object detection. Specifically, we identified the ineffectiveness and inadequacy issues present in single-frame object detectors. Then, we proposed to solve these problems using multi-frame and single-frame attention modules, as well as a dual-frame fusion module to

improve object representation. Our results show that by exploiting sequential frames, we can improve the efficiency and accuracy of detection under challenging conditions. This dual-frame approach, with its unique combination of temporal, channel, and spatial attention, represents a significant advancement in feature map processing for sequential analysis, to improve object detection.

CHAPTER 5 ARTICLE 2: LAQEM: TRANSFORMER WITH LANGUAGE-AUGMENTED QUERIES AND AN EVOLVING MEMORY FOR OBJECT DETECTION

Full Reference:
Noreen Anwar, Guillaume-Alexandre Bilodeau, and Wassim Bouachir. "LAQEM: Transformer with Language-Augmented Queries and an Evolving Memory for Object Detection." Submitted to Pattern Recognition Journal, Submitted on May 30 2024.

Statement of Contribution:
This article was co-authored with Professors Guillaume-Alexandre Bilodeau and Wassim Bouachir.

Noreen Anwar was primarily responsible for conceptualization, methodology, software development, formal analysis, investigation, data curation, visualization, and writing – original draft.

Professor Guillaume-Alexandre Bilodeau contributed to supervision, academic guidance, project administration, writing review, and editing. He also supported the acquisition of research funding.

Professor Wassim Bouachir contributed to conceptualization, validation, writing, review, and editing. He provided feedback on model design and interpretation of results.

All authors participated in revising and improving the final version of the manuscript.

This article is currently under review for publication in the Pattern Recognition Journal.

5.1 Abstract

Object Detection faces significant challenges in identifying novel objects and incremental learning of new classes over time. To address these challenges, we propose a transformer with **L**anguage-**A**ugmented **Q**ueries and an **E**volving **M**emory (LAQEM). Our novel model extends the capabilities of a transformer-based object detector by introducing a language-augmented memory, which integrates the Vision Language Model (VLM) driven object queries with an evolving memory mechanism. By reformulating the learning objective as a binary matching task between input features and corresponding object embeddings, our method enables effective generalization to seen and unseen classes. Our model is further

strengthened by conditioning the transformer decoder on VLM-derived embeddings, while the evolving memory retains features seen in the past to capitalize on historical knowledge. Experiments show that LAQEM achieves improvement over state-of-the-art methods for detecting known object classes and in the open-vocabulary setting for unseen object detection.

5.2 Introduction

Object detection, a core task in computer vision, involves identifying objects within images by predicting their bounding boxes and class labels. Traditional approaches, predominantly based on convolutional neural networks (CNNs), decompose this task into separate classification and regression problems over a multitude of proposals such as regions [99, 100], anchors [9], or keypoints [70, 101].

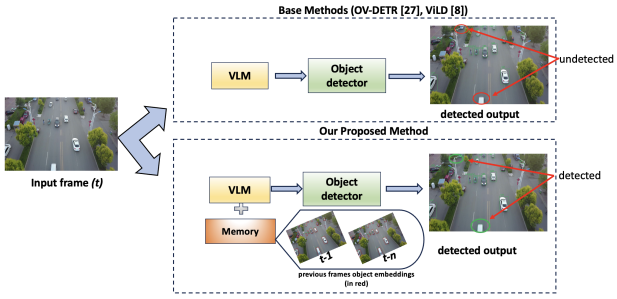

Figure 5.1 Comparison of the base methods like OV-DETR [23], ViLD [24](top) with Our Proposed Method (bottom) for object detection. The base methods perform frame-by-frame detection, while Our Proposed Method integrates a memory to retain information from previous frames (e.g., $t-1$ and $t-n$), enhancing detection accuracy and consistency across frames.

Recent advancements in object detection have seen the adoption of transformer architectures, reframing object detection as a direct set prediction problem using encoder-decoder models [17, 102]. While detectors like YOLOv5 [103], EfficientDet [65], and Cascade R-CNN [104] have made significant contributions, they may have limitations as they often rely on handcrafted components, such as non-maximum suppression. The introduction of DETR [17] marked a paradigm shift by offering an end-to-end transformer-based solution that simplifies object detection into a set prediction task, achieving notable improvements on benchmarks like COCO.

However, on the other hand, transformers such as Vision transformers [19, 63, 105, 106] have recently emerged as robust backbones for conventional detection frameworks such as Faster R-CNN [100], Mask R-CNN [66], and RetinaNet [9]. Despite their high precision, these

transformer-based detectors are computationally demanding (e.g., at least 200 GFLOPs), making them impractical for resource-constrained real-world applications. There is an end-to-end encoder-decoder transformer type detector named DEtection TRansformer (DETR) [17] that addresses this issue by simplifying object detection into a prediction task. DETR [17]innovative decoder allows object queries to focus on various regions of interest within a single-level representation, significantly enhancing inference efficiency (ranging from 86 to 253 GFLOPs). However, this improvement comes with a trade-off, leading to training convergence times that are approximately 10 to 20 times slower. Consequently, the challenge remains to develop transformer-based detectors that can achieve high precision while maintaining efficiency in both the training and inference stages.

Despite significant advancements, many existing models focus on static images, often neglecting the inherent temporal information in video data. This limitation reduces their effectiveness in handling dynamic scenes where object appearances, positions, and contexts continually evolve [53, 107]. Fully leveraging the temporal dimension in video object detection introduces substantial challenges, including managing motion blur, and occlusions, and ensuring the continuity of object tracking across frames [28, 55]. Nevertheless, the temporal aspect presents considerable opportunities to enhance detection accuracy, improve robustness in dynamic environments, and achieve a more comprehensive understanding of scene context and object interactions [2, 80].

Traditional frame-by-frame detection methods fail to capture the temporal continuity and contextual cues inherent in videos. By utilizing temporal information from preceding frames, detection accuracy, and robustness can be significantly improved, particularly in addressing challenges such as motion blur, occlusions, and dynamic scene transitions [53, 55, 107, 108]. Although methods like Flow-Guided Feature Aggregation (FGFA) [53] and Spatiotemporal LSTM (ST-LSTM) [107] demonstrate the benefits of integrating temporal context, they often come at the cost of increased computational complexity.

As illustrated in Figure 5.1, many video object detection methods struggle with retaining contextual information across frames, often resulting in missed detections or inconsistent recognition of objects as they move through a scene. In the figure, the base methods (OV-DETR [23] and ViLD [24]) employ a zero-shot Visual Language Model (VLM) alongside an object detector to perform single-frame detection, generating outputs based solely on the current frame. While this approach is reasonable for individual frame analysis, it lacks temporal coherence, potentially leading to missed detections in complex or occluded scenarios. In contrast, our proposed method incorporates a memory bank that retains visual information from previous frames (e.g., times $t-1$ and $t-n$), effectively fusing temporal cues to enhance

detection robustness. By leveraging the memory bank, the model integrates historical data with the current frame features, enabling improved identification of recurrent objects and achieving greater detection accuracy over time. This temporal fusion allows the model to make more informed decisions, particularly in sequences where objects experience occlusions or partial visibility.

Vision-language models are yielding promising results across various downstream vision tasks, including 2D and 3D perception and image generation [16,109–111]. The emergence of Vision-Language Pre-trained Models (VLMs) has revolutionized the field by mapping a broader array of visual concepts to natural language, enhancing model generality. Models such as CLIP [16] and FLAVA [111] have demonstrated significant improvements in visual recognition tasks, like object detection, segmentation, and image retrieval by leveraging extensive image-text datasets. These models effectively align features across modalities, resulting in impressive zero-shot classification abilities. Various adaptation strategies have been developed to improve performance on downstream tasks, often utilizing frozen VLM encoders in zero-shot and few-shot scenarios [16].

Building on these advancements, we propose a novel object detection model that leverages VLMs and the DETR model [28] by incorporating adaptive object queries specific to the classes detected in the current frame. Our approach integrates temporal context by leveraging visual embeddings from previous frames via CLIP [16], allowing object queries to prioritize relevant objects based on the scene temporal evolution. For example, if a car and a pedestrian are detected in the frame *t-1*, the queries in the frame t focus on these classes, improving detection accuracy and efficiency for frequently occurring objects. This integration with the Deformable DETRs attention mechanism enhances detection robustness and performance on video datasets, as we show in the results.

By leveraging the temporal continuity and contextual information inherent in sequential video frames, our approach aims to mitigate the performance limitations of existing video object detection methods. Additionally, we draw on vision-language pre-training [109,110] by adopting CLIP to align features across different modalities. This integration enables our enhanced DET-LIP model to harness the extensive knowledge embedded in vision-language models, thereby improving performance across various video object detection scenarios. his stream retains and refines features from previous frames, leveraging both spatial-temporal context and CLIP rich semantic representations. By embedding multimodal semantic information, the memory stream enables continuous adaptation across video sequences, significantly enhancing detection performance in few-shot scenarios and ensuring robust object detection in dynamic environments.

Furthermore, we introduce a dynamic memory network that preserves historical features and adaptively weights cached information, enabling improved identification of recurrent objects. These queries are dynamically adapted using contextual information from the CLIP-augmented memory stream, allowing the model to concentrate on relevant objects with increased precision. This approach improves detection stability and accuracy, particularly in scenarios involving object motion, appearance changes, and occlusions, thereby achieving superior performance in video object detection tasks. By incorporating insights beyond the provided training samples, our model demonstrates superior performance, particularly in few-shot scenarios. Extensive evaluations of benchmark video datasets show significant improvements in both accuracy and robustness, underscoring the effectiveness of our approach in detecting both known and novel objects.

The key contributions of our paper are as follows:

- We introduce a CLIP-augmented memory stream that integrates CLIP-driven multimodal alignment with a dynamic memory mechanism.

  This stream retains and refines features from previous frames, leveraging both spatial-temporal context and CLIP-rich semantic representations.

  By embedding multimodal semantic information, the memory stream enables continuous adaptation across video sequences, significantly enhancing detection performance in few-shot scenarios and ensuring robust object detection in dynamic environments.

- We enhance the Deformable DETR model by introducing object queries tailored to the current frame, specifically targeting known classes.

  These queries are dynamically adapted using contextual information from the CLIP-augmented memory stream, allowing the model to concentrate on relevant objects with increased precision. This approach improves detection stability and accuracy, particularly in scenarios involving object motion, appearance changes, and occlusions, thereby achieving superior performance in video object detection tasks.

- We advance the state-of-the-art in video object detection by demonstrating significant gains in detection accuracy and robustness, particularly in dynamic and complex video environments.

5.3 Related Work

5.3.1 Video Object detection

Video object detection has gained significant attention due to its applications in autonomous driving, surveillance, and robotics. Traditional approaches have focused on frame-by-frame detection, but recent methods leverage temporal information to enhance detection accuracy and robustness. Earlier methods relied on optical flow-based warping for feature aggregation. Some works, like Flow-Guided Feature Aggregation (FGFA) [53], aggregate features from multiple frames using optical flow to align features, improving robustness against motion blur and occlusion. However, these methods have notable drawbacks: 1) Training a model for optical flow extraction demands extensive optical flow data, which can be difficult and costly to obtain. 2) Integrating an optical flow network with a detection network into a single model poses challenges due to multi-task learning and model complexities. Integrating temporal information from previous frames into current frame predictions has been a widely adopted strategy for enhancing video object detection. Several approaches utilize information from previous or nearby frames at varying intervals to improve detection. For example, Anwar et al. [108] proposed a Spatio-Temporal Fusion (STF) module that selectively fuses information from both the current and previous frames to improve accuracy in complex scenarios. Similarly, Perreault et al. [55] and [56] introduced methods that combine recurrent networks and feature aggregation to enhance detection by capturing temporal continuity across multiple frames. These methods focus on incorporating short-term temporal context, but they face challenges when capturing long-term dependencies, which may limit overall robustness. This issue is reminiscent of the approach introduced by Zhu et al. [53], where features from adjacent frames are aggregated to improve current frame representations. While effective for short-term temporal modeling, these methods struggle with long-term context, thus requiring more advanced strategies for sustained accuracy in longer sequences.

5.3.2 Object Detection with Transformers

End-to-end object detection represents a significant shift from traditional detection pipelines, offering more streamlined and efficient architectures. DETR [17] pioneered this approach by introducing a transformer-based architecture and utilizing a Hungarian loss to achieve one-to-one matching predictions, effectively eliminating the need for handcrafted components and post-processing steps. Following DETR, several variants (Deformable DETR [28], DINO [112], RT-DETR [113]) have been developed to further enhance their performance and efficiency.

Deformable DETR [28] employs a multi-scale deformable attention module to accelerate convergence. DINO [112] integrates techniques such as contrastive denoising, mixed query selection, and the “look forward twice" scheme to improve DETR overall effectiveness. RT-DETR [113] introduces an efficient hybrid encoder and an uncertainty-minimal query selection strategy to optimize both accuracy and latency. Additionally, SDPDET [114] presents a Spatial-Dynamic Processing framework that adaptively scales feature maps based on object spatial distributions, thereby improving detection performance in varied spatial contexts. QueryDET [115] advances the query-based detection paradigm by implementing a Dynamic Query Generation mechanism, which adjusts query parameters in real time to better handle objects of diverse scales and appearances. These advancements collectively contribute to the evolution of transformer-based object detectors, each addressing the specific limitations of their predecessors. Our proposed approach distinguishes itself by integrating a CLIP-augmented memory stream with tailored object queries, thereby leveraging both temporal context and rich semantic representations to achieve superior performance in video object detection tasks.

5.3.3 Open-vocabulary Object Detectors

Foundation models such as [16,116,117] have recently garnered significant attention for their application in downstream tasks. Numerous vision-language pre-trained models [16,111,118] have emerged, trained on extensive image-text datasets in an unsupervised manner. These models typically include both image and text encoders, which produce features that can be aligned within a cross-modality representational space for effective image-text matching. Leveraging these alignment spaces enables zero-shot transfer to various downstream visual recognition tasks, such as object detection [119–121], with a particular focus on zero-shot and few-shot scenarios.

The Open-Vocabulary Detection Transformer (DETR) integrates the DETR architecture to address open-vocabulary detection tasks. OWL-ViT [22] employs a simplified Vision Transformer (ViT) architecture, leveraging a pre-trained CLIP model that is fine-tuned using the DETR objective. Similarly, OV-DETR [23] incorporates features from a pre-trained CLIP model into DETR object queries. This approach maintains the original DETR queries and augments each of them with a CLIP feature corresponding to each class, leading to a quadratic increase in the number of queries as it scales with both the original DETR queries and the number of classes considered.

Despite these advancements, current open-vocabulary DETR models face challenges in terms of scalability and computational complexity, particularly when applied to large-scale video

datasets. Moreover, the lack of temporal context integration limits their effectiveness in dynamic environments where object appearance and context evolve. Our work addresses these gaps by extending the capabilities of open-vocabulary models, specifically focusing on the seamless incorporation of temporal information to enhance performance in video-based object detection. This distinguishes our approach from previous methods and sets a new benchmark for video object detection using transformer-based vision-language models.

Despite these advancements, challenges remain in effectively applying vision-language models to video object detection, where temporal information and dynamic scenes present unique difficulties. Current approaches often struggle with the integration of temporal context, leading to limitations in accuracy and robustness.

In this work, we propose a method that enhances the capabilities of the Deformable DETR model by incorporating augmented queries from CLIP, specifically tailored to seen (known) classes. Our approach leverages temporal context in conjunction with the attention mechanisms of Deformable DETR to enhance detection accuracy and robustness in video datasets. With a memory, our method exploits the continuity and contextual information inherent in sequential video frames, addressing the limitations of existing approaches.

5.4 Methodology

Our proposed method is based on DETR. The Detection Transformer (DETR) [113] redefined the object detection process by structuring it as a unified, feed-forward network. In DETR and object detectors derived from it, object queries, essentially arbitrary feature vectors, are transformed into labeled bounding boxes through a series of cross-attention layers within the decoder. In the original DETR-based model [113], object queries are learned as free-form parameters, without any predefined structure. However, more recent DETR models [112, 122, 123] have adopted a two-stage approach, similar to the methodology used in R-CNNs [100]. In any case, the query mechanism plays a crucial role in guiding the detector internal processes, dictating which regions of the image the detector focuses on and which object classes are given priority. Our approach capitalizes on a VLM to define the object queries and it empowers DET-LIP to identify a broad range of both seen and unseen object classes, without being constrained by a predefined set of classes.

As well, the object queries are enhanced by leveraging temporal context from preceding frames within video sequences using a memory. By integrating this temporal information, our proposed DET-LIP model adaptively learns and recognizes objects over time, effectively utilizing multi-frame visual cues. Our method is detailed in the following.

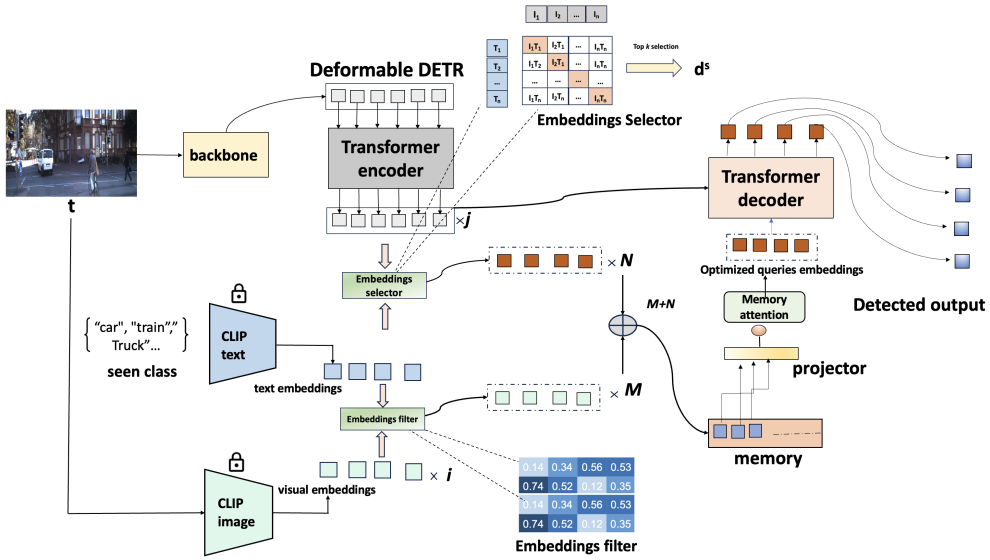

Figure 5.2 DET-LIP extends standard DETR by addressing the limitations of fixed class sets. Unlike traditional DETR, which operates within a predefined set of classes, DET-LIP leverages CLIP-based text-conditioned image embeddings to achieve flexible object recognition. This enables DET-LIP to detect a broader range of object categories based on dynamic historical inputs. The model introduces a dynamic filtering mechanism for embedding selection and utilizes memory attention to refine query embeddings, enhancing detection performance without being constrained by fixed class definitions.

5.4.1 Enhancing DETR Detection Mechanism

For an input image z, the standard DETR model generates K object predictions $\hat{v}$, where K is determined by a fixed number of object queries w, which are learnable positional encodings. The DETR pipeline consists of two primary steps during training: (i) object set prediction and (ii) optimal bipartite matching. These steps require ground-truth annotations to match predicted objects with actual objects in the image.

Given an input image z, a global context representation d is first extracted by a CNN backbone g_ξ and then processed by a transformer encoder j_χ, summarized as

$$d = j_\chi(g_\xi(z)), \tag{5.1}$$

where d represents a sequence of feature embeddings of z. Using the context feature d and object queries w as inputs, the transformer decoder l_θ (with prediction heads) produces the set prediction $\hat{v} = \{\hat{v}_k\}_{k=1}^{K}$ with

$$\hat{v} = l_\theta(d, w), \tag{5.2}$$

where $\hat{v}$ includes both bounding box predictions $\hat{b}$ and class predictions $\hat{p}$ for a predefined set of training classes.

Optimal bipartite matching is used during the training. Its goal is to find the best match between the K predictions $\hat{v}$ and the set of ground-truth objects $y = \{y_k\}_{k=1}^{M}$ (including no object $\emptyset$). This involves searching for a permutation of K elements $\sigma \in S_K$ that minimizes the matching cost:

$$\hat{\sigma} = \arg\min_{\sigma \in S_K} \sum_{k=1}^{K} L_{\text{cost}}(y_k, \hat{v}_{\sigma(k)}), \tag{5.3}$$

where $L_{\text{cost}}(y_k, \hat{v}_{\sigma(k)})$ is a pairwise matching cost between ground truth y_k and the prediction $\hat{v}_{\sigma(k)}$ with index $\sigma(k)$. The matching cost L_{cost} comprises losses for both class prediction $L_{\text{cls}}(\hat{p}, p)$ and bounding box localization $L_{\text{box}}(\hat{b}, b)$. This bipartite matching process results in one-to-one label assignments, where each prediction $\hat{v}_k$ is matched to a ground-truth annotation y_j or $\emptyset$ (no object). The optimal assignment can be efficiently computed using the Hungarian algorithm [124].

We enhance the DETR framework by incorporating temporal context from preceding frames and leveraging the cross-modal capabilities of CLIP, as depicted in Fig. 5.2. Traditional DETR models are constrained by single-frame detection and a fixed set of object classes. By integrating the CLIP visual encoder with DETR, specifically with respect to defined classes, our approach leverages temporal consistency and enriched contextual features, significantly improving detection accuracy. This integration enables the model to adaptively learn and recognize seen and unseen objects over time, resulting in superior performance in dynamic and complex environments. Our new predictions $\hat{v}_t$ are formulated as

$$\hat{v}_t = l_\theta(d_t, w^M), \tag{5.4}$$

where $\hat{v}_t$ is the prediction at time t, utilizing context from both the current frame embeddings d_t. The queries in w^M are composed of queries conditioned by a VLM and queries from previous frames.

5.4.2 Queries Conditioned by a VLM through Embedding Filtering

We enhance the DETR model by introducing a two-step mechanism consisting of a Query Selector and an Embedding Filter, which condition object queries using a Vision-Language Model (VLM). This mechanism leverages text embeddings derived from the CLIP text encoder to represent the seen classes.

First, visual embeddings are extracted from the Deformable DETR encoder for each current frame. These embeddings are passed through the Query Selector, which identifies the top k embeddings based on their relevance, focusing the model's attention on potential object locations, as shown in Figure 5.2.

Next, the Embedding Filter refines these selected embeddings by leveraging CLIP-generated text and visual embeddings. This filtering process involves computing the cosine similarity between each selected visual embedding and the corresponding CLIP-derived text embedding. Visual embeddings with similarity scores surpassing a predefined threshold are retained, aligning them more closely with seen classes.

To achieve open-vocabulary detection, we utilize CLIP alignment capabilities by generating text embeddings d_j^{text} for the seen classes using the CLIP model [16]:

$$d_j^{\text{text}} = \text{CLIP}_{\text{text}}(a_c), \tag{5.5}$$

where a_c denotes the text labels of the seen classes. Corresponding visual embeddings d_i^{img} are extracted by conditioning the image input z on these text embeddings:

$$d_i^{\text{img}} = \text{CLIP}_{\text{img}}(z \mid d_j^{\text{text}}), \tag{5.6}$$

where d_i^{img} represents the resulting visual embeddings aligned with d_j^{text}.

We combine these aligned embeddings in a shared latent space using contrastive learning, following the methodology in CLIP [16]. The selected visual embeddings from the Query Selector are further combined with these aligned embeddings, forming a comprehensive set of optimized query embeddings. These embeddings are stored in a memory bank, denoted as w^M:

$$w^M = [\, q^{\text{selected}};\, d_i^{\text{img}};\, d_j^{\text{text}} \,]. \tag{5.7}$$

This combined storage allows the model to retrieve and reuse relevant embeddings dynami-

cally, enhancing detection performance across successive frames. Furthermore, by leveraging the embedding filtering strategy, the memory bank is updated with only those embeddings that exceed a predefined similarity threshold, enabling the model to maintain high-quality historical embeddings while discarding irrelevant or noisy information. This adaptive filtering approach enhances model robustness and stability in challenging scenarios such as appearance changes, occlusions, and dynamic object movements. For more details, see the supplementary material.

5.4.3 Adaptive Memory Conditioning in DETR

We also introduce an adaptive memory, which plays a crucial role in accumulating historical visual information over time. The memory bank serves as a dynamic storage mechanism, continuously updating with new information to improve model performance in both current and future frames. It effectively allows the model to retain past information and leverage it for enhanced decision-making and accuracy. The memory bank is particularly useful for managing visual features associated with different object classes. When the model encounters an object that it has not seen before, it retrieves relevant features from the memory bank, which stores embeddings of previously detected objects. This retrieval process allows the model to leverage prior knowledge and recognize new objects based on their similarity to known visual patterns. These retrieved features refine the current queries, enabling the model to generalize to new objects based on past information. This enriched representation improves object detection performance, even for classes not explicitly encountered in earlier frames. Given the updated memory A_d and the visual features from CLIP d_i^{img}, the memory is updated based on the following equation:

$$A_d \leftarrow \begin{cases} A_d \cup \{d_i^{\text{img}}\}, & \text{if } \left\| d_i^{\text{img}} - A_d \right\| > \text{threshold}, \\ \left(A_d \setminus \{d_k^{\text{img}}\}\right) \cup \{d_i^{\text{img}}\}, & \text{if } \left\| d_k^{\text{img}} - A_d \right\| \leq \text{threshold and memory is full}, \\ A_d, & \text{otherwise.} \end{cases} \tag{5.8}$$

5.4.4 Visual Matching

With the enriched data stored in the memory bank, our next objective is to measure the similarity between the embeddings from the memory bank and the detection outputs from DETR.

To perform visual matching, we begin by projecting w^M into the same dimensional space as

the object queries q using a fully connected layer F_{proj}.

The input to the DETR decoder, denoted as q', is then computed as:

$$q' = q \oplus F_{\text{proj}}(w^M) \tag{5.9}$$

where the $\oplus$ operation combines class-agnostic object queries q with visual-specific embeddings $F_{\text{proj}}(w^M)$, enriching the queries with image-specific information. Further details of the memory bank integration process, including the algorithm and visual representation of the query replication strategy, can be found in the supplementary material.

Given the conditioned query features q', our binary matching loss for label assignment is defined as:

$$L_{\text{cost}}(y, \hat{y}_\sigma) = L_{\text{match}}(p, \hat{p}_\sigma) + L_{\text{box}}(b, \hat{b}_\sigma), \tag{5.10}$$

where $L_{\text{match}}(p, \hat{p}_\sigma)$ replaces the classification loss $L_{\text{cls}}(p, \hat{p}_\sigma)$. In this framework, p represents a 1-dimensional sigmoid probability vector that encodes the similarity of objects, where values close to 1 indicate that an object is "matched" (i.e., correctly identified as the target object), and values near 0 represent "not matched" instances. To optimize the matching process, we employ Focal Loss [9] as the loss function, calculated between the predicted similarity scores $\hat{p}_\sigma$ and the ground truth labels p. This approach helps the model focus on harder-to-classify examples, improving robustness in identifying correct matches, especially in challenging detection scenarios.

The matching loss is designed to allow the model to associate all instances of a queried object within an image while treating instances of other classes as "not matched." This process enables the model to selectively focus on relevant objects in the detection task, ensuring it learns robust and adaptive representations. By dynamically handling mismatches and correct matches, the model becomes well-suited for open-vocabulary detection, where it must generalize across unseen object classes. This loss mechanism is particularly beneficial in multi-frame contexts, where consistent detection across frames is crucial for accurate object tracking and recognition.

Upon optimizing Eq. (5.10), we derive the optimized label assignments ω for various object queries. This yields a set of detected objects with associated box coordinates $\hat{b}$ and a 2-dimensional matching probability $\hat{p}$. These outputs are utilized to calculate our final loss function during model training. For our final model training loss, we integrate L_{embed} with the bounding box losses $L_{\text{match}}(p, \hat{p})$ and $L_{\text{box}}(b, \hat{b})$ as follows:

$$L_{\text{loss}}(y, \hat{y}) = \lambda_{\text{Focal}} L_{\text{Focal}} + \lambda_{L1} L_{L1} + \lambda_{\text{GIoU}} L_{\text{GIoU}} + \lambda_{\text{embed}} L_{\text{embed}}, \tag{5.11}$$

where L_{box} includes the L1 loss and the generalized IoU (GIoU) loss for bounding boxes. The parameters λ_{Focal}, λ_{L1}, λ_{GIoU}, and λ_{embed} serve as the respective weighting factors.

5.4.5 Inference

During testing, for each image, we send the text embeddings, d_j^{text}, of all the base and novel classes to the model and merge the results by selecting the top-k predictions with the highest prediction scores. We follow prior work [16] to use $k = 100$ for the COCO dataset. To obtain the context representation d in Eq. (1), we forward the input image through the CNN backbone g_ξ and Transformer encoder j_χ. Note that d is computed only once and shared for all conditional inputs for efficiency. Then, the conditioned object queries from different classes are sent to the Transformer decoder in parallel.

5.5 Experiments

In this section, we evaluate the performance of our proposed method by comparing it with state-of-the-art (SOTA) techniques and conducting an ablation study.

5.5.1 Datasets and Evaluation Metrics

Datasets: To overcome the limitations of relying solely on single-frame datasets, we propose utilizing video datasets to exploit the temporal information inherent in sequential frames. Specifically, we employ the Cityscapes [2] dataset, which, although commonly used for urban scene analysis, provides valuable video sequences suitable for our object detection tasks. Additionally, we incorporate the UAVDT [82], UA-DETRAC [125], and VisDrone [126] datasets, all explicitly designed for object detection and tracking in video sequences captured from diverse perspectives, including aerial viewpoints. These datasets present unique challenges by offering sequences with varying viewpoints, object scales, and motion patterns. For a comprehensive evaluation, we partition the datasets into 'seen' and 'unseen' classes, training our model exclusively on the 'seen' classes. Bounding boxes not labeled with a 'seen' class are excluded from the training data, and images without any remaining annotations are discarded. This setup enables a robust comparison with state-of-the-art methods, which typically train on single frames and consider all classes simultaneously.

Evaluation Metrics: Detection accuracy is evaluated using a comprehensive set of metrics,

with particular attention to the distinction between seen and unseen classes during training to test the zero-shot capabilities of our proposed method. Average Precision (AP) is computed separately for seen and unseen classes across various object scales. Mean Average Precision (mAP) is reported over a range of Intersections over Union (IoU) thresholds, with specific emphasis on AP_{50} and AP_{75} values, corresponding to IoU thresholds of 0.5 and 0.75, respectively. IoU serves as the standard criterion for assessing bounding box precision across all datasets, ensuring a rigorous evaluation of detection performance for both seen and unseen classes.

5.5.2 Implementation Details

Our approach builds upon the vanilla Deformable DETR framework [113]. We leverage the open-source CLIP model [16] utilizing ViTB/32 to extract image embeddings from text embeddings. We trained our model with AdamW optimizer, where we adopted the cosine annealing learning schedule with the initial learning rate of 1e-4. Code is available at `https://github.com/DET-LIP/DET-LIP`

5.6 Results and Discussion

We first report our results by training on all the classes of the datasets. Meaning that all the classes are seen. This allows us to assess the ability of our method to leverage multiple frame information and generic CLIP features for better detection.

5.6.1 Results

We report the performance of our DET-LIP model across several benchmark datasets in Tables 5.1–5.4, demonstrating significant improvements over state-of-the-art (SOTA) methods. As shown in Table 5.1, on the Cityscapes validation dataset, our model achieves substantial gains across various COCO metrics, including overall AP, AP_{50}, and AP_{75}. Particularly, DET-LIP demonstrates robust performance on small and medium objects (AP_S and AP_M), addressing typical challenges in urban scene detection tasks. These improvements can be attributed to the integration of CLIP features with our memory bank, which leverages multi-frame information to enhance detection accuracy and maintain temporal consistency, particularly for larger objects (AP_L).

Moving to the VisDrone dataset, shown in Table 5.2, DET-LIP outperforms baseline methods, highlighting its adaptability to the aerial perspective and small object detection (AP_S), which is critical in this context. Our method benefits from memory embeddings that capture

historical frame information, enabling it to retain context and improve detection in sequences with occlusions or varying object scales. This results in a notable increase in both AP_{50} and AP_{75} metrics, confirming the model's capability to deliver high precision across diverse aerial scenes.

Table 5.3 presents our results on the UAVDT test dataset, where DET-LIP shows consistent performance gains over existing detectors, especially in overall AP and AP_{75} scores. UAVDT, with its challenges in scale variance and motion blur, is well-suited for demonstrating the strengths of DET-LIP. By incorporating temporal embeddings, the model effectively detects small objects and improves recognition of medium and large objects in UAV imagery, thereby enhancing overall detection accuracy. These results underscore the effectiveness of our model in handling the complexities inherent in UAV-based detection tasks.

Finally, in the UA-DETRAC test dataset results (Table 5.4), DET-LIP achieves higher AP scores across all size categories. The integration of our memory bank enables the model to track objects with temporal coherence, enhancing its performance in detecting medium and large vehicles (AP_M and AP_L) common in traffic scenarios. Additionally, DET-LIP improved AP_S demonstrates its ability to handle smaller vehicles and occluded objects, which are frequently encountered in surveillance applications. This consistent performance across multiple datasets validates DET-LIP's robustness and adaptability to varied detection contexts, achieving SOTA results in various challenging scenarios.

Table 5.1 Comparison of our method with SOTA methods on the Cityscapes validation dataset. **Boldface** indicates best results.

Method	**Backbone**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR [113]	ResNet-50	14.7	26.5	11.3	4.6	11.2	28.6
Deformable DETR [28]	ResNet-50	22.5	41.7	22.5	7.9	20.7	42.1
DELA-DETR [127]	ResNet-50	25.2	46.8	22.8	6.5	23.8	44.3
Moco-v2 [128]	ResNet-50	27.2	55.2	26.7	10.9	28.4	46.7
FP-DETR [129]	Resnet-50	29.6	53.6	28.4	11.2	30.9	47.4
DenseCL [130]	Resnet-50	30.1	53.5	35.7	11.8	32.6	55.2
ViLD [24]	Resnet-50	29.7	54.3	52.5	-	-	-
OV-DETR [23]	Resnet-50	31.5	54.3	36.2	11.1	34.5	56.1
DET-LIP (our model)	Resnet-50	**34.9**	**59.3**	**40.1**	**15.0**	**39.5**	**60.8**

5.6.2 Ablation Study

We conducted an ablation study on DET-LIP to evaluate our model architecture. We first assess the importance of a memory bank in our model. Our findings indicate that incorporating a memory bank significantly enhances performance. Table 5.5 demonstrates that

Table 5.2 Comparison of our method with SOTA methods on the VisDrone validation dataset. **Boldface** indicates best results.

Method	**Backbone**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR [113]	ResNet50	29.0	41.2	27.5	22.1	27.2	42.9
Deformable DETR [28]	ResNet50	27.2	44.1	28.5	13.5	25.2	37.8
QueryDet [115]	ResNet50	28.3	48.1	28.8	19.8	35.9	40.3
SDPDet [114]	ResNet50	33.7	56.6	34.3	26.7	42.9	45.7
OV-DETR [23]	Resnet-50	33.9	57.3	35.4	27.5	43.6	46.3
DET-LIP (our model)	ResNet50	**35.5**	**59.2**	**37.2**	**29.6**	**44.7**	**47.5**

Table 5.3 Comparison of UAVDT test dataset with different methods. **Boldface** indicates best results.

Method	**Backbone**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR [113]	ResNet50	11.1	21.0	9.1	5.3	10.1	13.6
Deformable DETR [28]	ResNet50	13.6	23.2	9.8	5.7	14.3	17.6
QueryDet [115]	Resnet-50	14.3	27.2	16.6	11.1	24.2	14.7
SDPDet [114]	ResNet50	20.0	32.0	23.1	13.3	33.0	21.1
OV-DETR [23]	Resnet-50	28.3	35.1	22.4	12.5	27.5	24.1
DET-LIP (our model)	ResNet50	**30.4**	**40.2**	**25.7**	**16.1**	**23.2**	**26.7**

integrating information from previous frames improves detection results, underscoring the value of historical data. Optimal results are achieved by leveraging memory networks and validating their complementary strengths. However, this approach may pose challenges for applications with limited storage resources.

To enhance the detection of unseen classes, we incorporated class-agnostic object proposals into the training process. These proposals, inspired by ViLD [24], identify regions likely to contain objects, including those from unseen classes. However, since class labels for unseen classes are unavailable during training, label assignments for these regions can be incorrect, resulting in reduced AP for unseen classes.

Performance Across Multiple Datasets. Table 5.6 presents a comprehensive comparison of our method against the baseline OV-DETR across various datasets. Our approach, DET-LIP, consistently outperforms OV-DETR in both AP^s (AP for seen classes) and AP^u (AP for unseen classes) across the Cityscapes, VisDrone, UAVDT, and UA-DETRAC datasets. The significant improvements, highlighted in bold, demonstrate the effectiveness of integrating memory networks and class-agnostic object proposals in enhancing detection performance, particularly for unseen classes.

Effect of Object Proposals and CLIP Embeddings. To further investigate the contri-

Table 5.4 Comparison of UA-DETRAC test dataset with different methods. **Boldface** indicates best results.

Method	**Backbone**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
DETR [113]	ResNet50	16.3	27.5	10.1	6.5	16.8	23.1
Deformable DETR [28]	ResNet50	18.5	35.6	10.4	7.0	17.1	25.7
OV-DETR [23]	Resnet-50	28.8	37.3	23.9	13.2	32.5	28.6
DET-LIP (our model)	ResNet50	**30.7**	**43.2**	**29.5**	**19.6**	**33.5**	**35.9**

Table 5.5 Ablation study on memory network

Configuration	**AP (%)**
DETR	14.7
Deformable DETR	22.5
DET-LIP (without memory)	31.6
DET-LIP (with memory)	**34.9**

butions of different components in our model, we conducted an ablation study focusing on the use of object proposals (P) and CLIP embeddings (C). Table 5.7 illustrates the performance under various configurations. The inclusion of object proposals alone (P) results in a moderate improvement in both AP_s and AP_u, highlighting the importance of incorporating object proposals for enhancing detection accuracy. However, when both object proposals and CLIP embeddings (P+C) are utilized, there is a substantial increase in performance, achieving the highest AP_s and AP_u. This indicates that the combination of these components synergistically enhances the model's ability to detect both seen and unseen classes effectively.

The ablation studies validate the critical components of our DET-LIP architecture. The memory bank significantly boosts detection performance by leveraging historical frame information. Additionally, the integration of class-agnostic object proposals and CLIP embeddings plays a pivotal role in enhancing the detection of both seen and unseen classes. These findings collectively demonstrate the effectiveness of our proposed enhancements in creating a robust and adaptable object detection framework.

5.7 Conclusion

In this work, we introduced DET-LIP, a groundbreaking open-vocabulary object detector that synergies the strengths of DETR and CLIP. By addressing the challenge of classifying novel objects without relying on labeled data, our approach employs a binary matching objective between input queries and object embeddings derived from a pre-trained vision-

Table 5.6 Comparison of methods across different datasets. **Boldface** indicates best results.

Method	**Cityscapes**		**VisDrone**		**UAVDT**		**UA-DETRAC**	
	$\mathbf{AP^s}$	$\mathbf{AP^u}$	$\mathbf{AP^s}$	$\mathbf{AP^u}$	$\mathbf{AP^s}$	$\mathbf{AP^u}$	$\mathbf{AP^s}$	$\mathbf{AP^u}$
OV-DETR [23]	25.4	16.7	21.6	12.6	19.8	9.4	18.7	8.5
DET-LIP (our method)	**29.1**	**19.3**	**24.1**	**15.5**	**22.1**	**12.3**	**21.2**	**10.8**

Table 5.7 Ablation study on using object proposals (P) and CLIP embeddings (C).

P	**C**	AP_s	AP_u
		27.6	11.5
✓		22.2	8.7
✓	✓	35.5	23.4

language model. Furthermore, we enhance the models adaptability in few-shot scenarios with a dynamic memory bank. Our extensive experiments across diverse datasets underscore the effectiveness of DET-LIP, demonstrating substantial improvements over state-of-the-art methods. This study marks a significant advancement in open-world object detection and sets a new benchmark for future research in vision-language integration.

CHAPTER 6 ARTICLE 3: DUAL-STREAM ATTENTION WITH MULTI-MODAL QUERIES FOR OBJECT DETECTION IN TRANSPORTATION APPLICATIONS

Full Reference:
Noreen Anwar, Guillaume-Alexandre Bilodeau, and Wassim Bouachir. "Dual-Stream Attention with Multi-Modal Queries for Object Detection in Transportation Applications." Accepted to British Machine Vision Conference (BMVC), Submitted on July 25 2025.

Statement of Contribution:
This article was co-authored with Professors Guillaume-Alexandre Bilodeau and Wassim Bouachir.

Noreen Anwar was responsible for conceptualization, methodology, software implementation, experimentation, data curation, formal analysis, visualization, and writing – original draft.

Professor Guillaume-Alexandre Bilodeau contributed to supervision, academic guidance, project administration, writing, review, and editing. He also supported funding acquisition.

Professor Wassim Bouachir contributed to conceptualization, validation, writing, review, and editing. He provided insights into the attention mechanism and multimodal integration strategy.

All authors contributed to revising and refining the final manuscript.

This article is accepted at the British Machine Vision Conference (BMVC 2025).

6.1 Abstract

Transformer-based object detectors often struggle with occlusions, fine-grained localization, and computational inefficiency caused by fixed queries and dense attention. We propose DAMM, Dual-stream Attention with Multi-Modal queries, a novel framework introducing both query adaptation and structured cross-attention for improved accuracy and efficiency. DAMM capitalizes on three types of queries: appearance-based queries from vision-language models, positional queries using polygonal embeddings, and random learned queries for general scene coverage. Furthermore, a dual-stream cross-attention module separately refines

semantic and spatial features, boosting localization precision in cluttered scenes. We evaluated DAMM on four challenging benchmarks, and it achieved state-of-the-art performance in average precision (AP) and recall, demonstrating the effectiveness of multi-modal query adaptation and dual-stream attention. Source code is at: GitHub.

6.2 Introduction

Traditional approaches for object detection in transportation applications focused on classifying and localizing objects within predefined categories using rigid bounding boxes. While these methods have achieved remarkable success, they struggle to generalize to real-world scenarios characterized by arbitrary objects, occlusions, and complex spatial structures. Recent transformer-based detectors, such as DETR and its variants, have improved detection accuracy but remain constrained by static object queries, computationally expensive global attention mechanisms, and limited spatial granularity. These limitations are particularly pronounced in dynamic environments where object shapes, occlusions, and distributions deviate significantly from what was seen during training, as exemplified in aerial and urban scenarios. UAVDT [82] and VisDrone [131] highlight the challenges of detecting small and highly occluded objects in UAV-based imagery. Existing methods such as RT-DETR [132] and UAV-DETR [132] attempt to optimize transformers for UAV-based detection but remain limited by their reliance on predefined object distributions.

Recent breakthroughs in vision-language models (VLMs) [20, 22, 121] have opened new avenues for open-world recognition by narrowing the gap between visual and textual representations. However, existing detection frameworks fail to fully capitalize on these advancements due to three shortcomings: (1) reliance on static query embeddings, which lack adaptability to diverse object appearances and contexts; (2) computational inefficiencies stemming from dense self-attention mechanisms in the decoder; and (3) the use of rigid, axis-aligned bounding box positional queries, which provide suboptimal positioning for objects with irregular geometries. We address these shortcomings with four contributions:

- Multi-Modal Queries: We introduce a unified query set that integrates appearance-based queries derived from vision-language embeddings to capture semantics, position-based queries from segmentation-driven polygonal embeddings to capture spatial information, and random learned queries to ensure robust general scene coverage. This multi-modal approach enables the model to adapt to diverse object appearances and contexts dynamically, addressing the first shortcoming.
- Adaptive Query Fusion: We designed a learnable mechanism to dynamically refine

static and dynamic queries within the transformer decoder. This adaptive fusion accelerates convergence while improving generalization across diverse detection scenarios, complementing our first contribution.

- Dual-Stream Cross-Attention: We propose a structured attention mechanism that decouples semantic and spatial representations, optimizing feature alignment from unified query adaptation while significantly reducing computational overhead. This dual-stream design enhances the models ability to reason about both object semantic and spatial relationships efficiently, addressing the second shortcoming.

- Polygonal Positional Embeddings: We introduce an original scheme that encodes object boundaries as polygonal positional embeddings, enabling our model to precisely capture irregular shapes and occlusions, thereby achieving superior localization accuracy, and addressing the third shortcoming.

These contributions are integrated in DAMM (Dual-Stream Attention with Multi-Modal Queries), a novel transformer-based video detection framework fusing appearance queries from Grounding DINO [133], positional queries from SAM [116], and random learned queries into a unified representation. These queries are processed via a dual-stream decoder that separately refines semantic and spatial cues to enhance detection accuracy across diverse scenarios. We evaluated DAMM on four challenging datasets, Cityscapes [2], UAVDT [82], VisDrone [3], and UA-DETRAC [125], demonstrating state-of-the-art performance in both mean average precision (mAP) and recall. DAMM consistently outperforms existing methods, particularly in scenarios involving occlusions and fine-grained spatial structures.

6.3 Related Work

Object detection has evolved significantly with the advent of transformer and vision-language models. Despite these advancements, existing methods still face challenges in query representation, spatial precision, and cross-modal feature integration. Our work builds upon recent developments in structured detection models, multi-modal fusion, and query refinement strategies to introduce a more flexible and robust detection framework.

Transformer-Based Object Detection: The introduction of DETR [17] transformed object detection by eliminating traditional heuristics, such as non-maximum suppression (NMS) and anchor generation. However, its slow convergence and global dense attention motivated the development of more efficient variants. Deformable DETR [28] significantly improved computational efficiency by introducing sparse spatial sampling, while DINO [112] and DN-

DETR [134] refined query-based learning (that is, how object queries are generated, updated, and refined during training) to accelerate training and enhance feature refinement. Other methods, such as Conditional DETR [135], introduced conditional queries to better align detection predictions, while Sparse DETR [132] and QueryDet [115] focused on optimizing similar queries efficiency by propagating only the most informative ones. Despite these advances, existing DETR-based models remain limited by static object queries that lack adaptability in dynamic scenes. Our approach addresses this issue by incorporating multi-modal query adaptation, allowing queries to dynamically fuse semantic and spatial embeddings for better detection performance.

Spatial Representation Learning: Traditional bounding boxes often struggle with irregular shapes in aerial (VisDrone [**?**]) and urban (Cityscapes [**?**]) datasets. Deformable DETR [28] enhanced spatial sensitivity through offset learning, while SDPDet [114] improved with scale-aware localization by dynamically fusing multi-scale features and incorporating scale-specific modulation into its regression head, enabling more accurate detection of objects with varying sizes. Recent contour-based methods like Poly-YOLO [136] and PolarMask [137] demonstrated the benefits of polygonal representations, but require complex post-processing. We address this by integrating segmentation-derived polygonal embeddings directly into the transformer decoder, achieving more precise localization without additional refinement steps.

Vision-Language Integration in Detection: Recent vision-language models (VLMs) have extended detection capabilities beyond closed-set categories, allowing for more flexible and open-world recognition. CLIP [16] and ALIGN [138] demonstrated strong zero-shot classification capabilities, leading to their adaptation in detection frameworks such as OWL-ViT [22] and GLIP [21]. Grounding DINO [20] further incorporated text-driven region supervision to enhance detection performance. However, most existing methods rely on static text embeddings, limiting their adaptability to unseen categories. OV-DETR [23] addressed this by integrating open-vocabulary learning into a DETR-based framework. However, it still lacks dynamic query adaptation. Our model builds on these works by introducing adaptive query fusion, which integrates both vision-language embeddings and structured spatial cues, leading to improved generalization across diverse object categories.

6.4 Methodology

To improve generalization, DAMM introduces adaptive query mechanisms that dynamically refine representations across multiple detection stages. By combining multi-modal feature fusion, fine-grained embeddings, and structured attention mechanisms, DAMM achieves supe-

rior robustness across diverse benchmarks, as demonstrated in our results. DAMM overcomes the constraints of fixed object queries by dynamically integrating multi-modal cues and decoupling semantic and spatial processing through a dual-stream cross-attention mechanism. As illustrated in Figure 6.1, DAMM fuses image semantic features from Grounding DINO, positional cues from SAM-generated polygons, and randomly learned embeddings to generate dual-stream queries that are appearance-based and position-based. By fusing these representations within a unified query adaptation mechanism, we enable the DETR-like decoder to generate more accurate object predictions. These queries undergo independent refinement via a dual cross-attention module. This design enhances detection robustness by dynamically adapting queries and leveraging multi-modal information for improved localization and recognition.

6.4.1 Multi-Modal Queries

Conventional DETR-based detectors rely on fixed object queries, limiting their adaptability to variations in object appearance and spatial configuration. To address this, we propose unified query adaptation, which constructs a dynamic query representation by fusing cues from multiple modalities. Our formulation consists of three distinct query types:

1) **Appearance-Based Queries** ($\mathbf{Q}_{\text{app}}$): Leveraging Grounding DINO [133] strengths in open-vocabulary recognition, we compute the appearance-based queries for a given image $\mathbf{I}$ and text prompt $\mathcal{T}$ as:

$$\mathbf{Q}_{\text{app}} = \phi_{\text{proj}}(\text{GDINO}(\mathbf{I}, \mathcal{T})); \tag{6.1}$$

2) **Positional Queries** ($\mathbf{Q}_{\text{pos}}$): These are obtained from segmentation masks generated by SAM [116] based on prompts, $\mathcal{P}$, from Grounding DINO. A polygonal approximation $\mathbf{B}$ of each mask $\mathbf{M}$ is made and transformed into an embedding. The polygonal representation efficiently captures the boundary information of a potential object to detect, making those queries more apt to deal with occlusions. By converting $\mathbf{B}$ into a flattened form and processing it through an MLP, we encode rich geometric details that enhance the models ability to localize objects. For a polygonal approximation $\mathbf{B}$, we have:

$$\mathbf{Q}_{\text{pos}} = \text{MLP}\Big(\text{Flatten}(\mathbf{B})\Big), \tag{6.2}$$

where $\mathbf{M} = \text{SAM}(\mathbf{I}, \mathcal{P})$, $\mathbf{B} = \text{PA}(\mathbf{M})$, and $\text{PA}(\cdot)$ computes the polygonal approximation of the mask.

3) **Random Learnable Queries** ($\mathbf{Q}_{\text{ran}}$): These are learnable queries inherited from the DETR framework, initialized from a Gaussian distribution:

$$\mathbf{Q}_{\text{ran}} \sim \mathcal{N}(0, \mathbf{I}) \tag{6.3}$$

Although DETR treats these queries as learnable, their initialization from a Gaussian distribution ensures diverse starting points, which are refined during training.

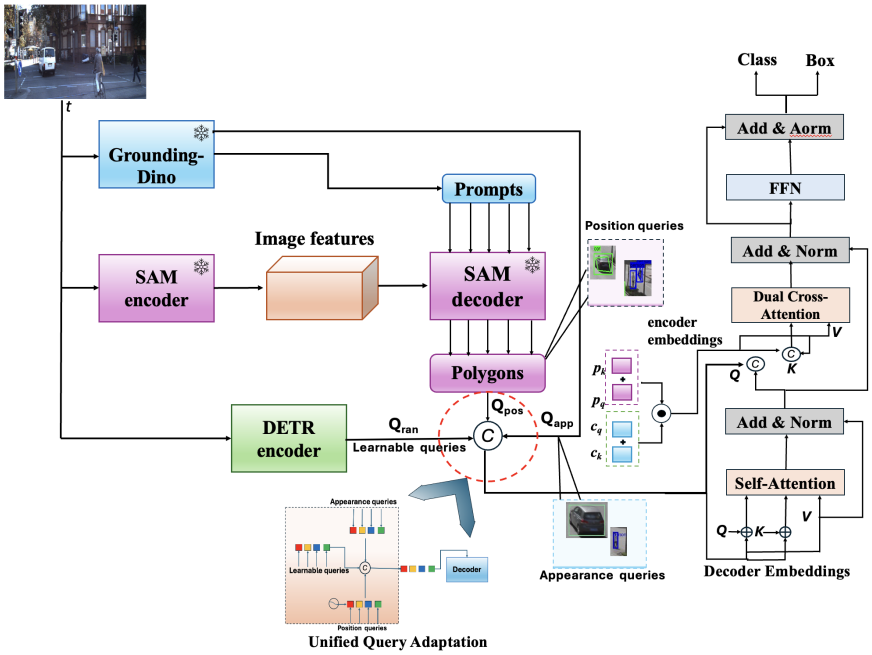

Figure 6.1 **DAMM Framework.** Our approach builds upon transformer-based detection by integrating multi-modal queries, unified query adaptation fusion, and dual-stream cross-attention. Object queries dynamically incorporate appearance-based and positional cues.

The overall query set is formed by concatenating the three components:

$$\mathbf{Q}_{\text{combined}} = [\mathbf{Q}_{\text{app}} \,\|\, \mathbf{Q}_{\text{pos}} \,\|\, \mathbf{Q}_{\text{ran}}]\,, \tag{6.4}$$

where $\|$ denotes channel-wise concatenation. This multi-scale, multi-modal representation captures both semantic and spatial information effectively.

6.4.2 Transformer-Based Detection Pipeline

DAMM adheres to a standard transformer-based detection architecture comprising an encoder, a decoder, and prediction heads for classification and localization. They can be summarized as follows. **Encoder:** Feature maps are extracted from a backbone network (e.g., ResNet-50 or Swin Transformer). The encoder utilizes multi-head self-attention to capture long-range dependencies and employs deformable attention [28] to efficiently sample salient spatial regions. **Decoder:** The decoder refines object queries through two mechanisms: 1) **Self-Attention:** Enables interaction among queries, promoting context-aware refinement and reducing redundancy and 2) **Dual-Stream Cross-Attention:** Decouples semantic and spatial processing to effectively integrate category-level and localization cues.

Dual-Stream Cross-Attention

Standard cross-attention in DETR aggregates all feature components uniformly, potentially obscuring the distinction between semantic and spatial cues. Our dual-stream cross-attention module addresses this issue by decomposing the attention mechanism into two independent streams: one that processes appearance-based queries from Grounding DINO, and another that leverages position-based queries derived from SAM polygonal embeddings. Each object query is represented as $\mathbf{Q}_i = \left[\mathbf{Q}^i_{\text{app}} \; ; \; \mathbf{Q}^i_{\text{pos}}\right]$, where $\mathbf{Q}^i_{\text{app}}$ and $\mathbf{Q}^i_{\text{pos}}$ denote the semantic (appearance) and spatial (positional) parts of the i-th query. These components are processed independently in their respective attention streams. The cross-attention operation for the i-th query is computed as:

$$\text{Attention}(\mathbf{Q}_i, \mathbf{K}, \mathbf{V}) = \mathbf{c}_q^\top \mathbf{c}_k + \mathbf{p}_q^\top \mathbf{p}_k, \tag{6.5}$$

where $\mathbf{c}_q = \mathbf{Q}^k_{\text{app}}$ and $\mathbf{c}_k = \mathbf{K}_{\text{app}}$ represent the semantic query and key vectors, respectively, and $\mathbf{p}_q = \mathbf{Q}^j_{\text{pos}}$ and $\mathbf{p}_k = \mathbf{K}_{\text{pos}}$ denote the spatial query and key vectors. This decoupled formulation allows the network to concentrate on content and spatial localization independently, yielding improved performance in occluded and cluttered scenes.

Unified Query Adaptation Fusion

Static query embeddings limit the adaptability of transformer-based detectors in dynamic scenes. We propose **adaptive query fusion**, a mechanism that iteratively refines object queries across decoder layers. At each decoder layer t, the query representation $\mathbf{Q}_t$ is updated using features derived from the current cross-attended output $\mathbf{F}_t$:

$$\mathbf{Q}_{t+1} = \text{FFN}\left(\text{LN}(\mathbf{Q}_t + \mathbf{F}_t)\right), \tag{6.6}$$

where $\mathbf{F}_t$ is the cross-attention output at step t, $\text{LN}(\cdot)$ denotes LayerNorm, and $\text{FFN}(\cdot)$ is a two-layer feedforward network with ReLU activation. This formulation enables the queries to incorporate up-to-date contextual cues from the encoder features at each stage. Importantly, the subscript t indexes decoder layers (not time steps). By continuously merging semantic and spatial context through learned fusion, DAMM dynamically adapts to scene structure and object interactions, improving robustness to occlusions and complex layouts.

Reference Points

Reference points serve as the spatial anchors for box prediction in transformer-based detection frameworks. Specifically, the decoder embedding $\mathbf{f}$ encodes the displacements of the four corners of a bounding box relative to a reference point $\mathbf{s}$. The final box is obtained by predicting these displacements in the unnormalized space and then normalizing the result to

the range $[0, 1]$. In the original DETR framework, the reference point is statically initialized as $\mathbf{s} = [0, 0]^\top$ for all decoder queries. In DAMM, we extend this approach by exploring two dynamic formulations for initializing reference points: 1) **Global Learnable**: Unnormalized 2D coordinates $\mathbf{s}^* \in \mathbb{R}^2$ are treated as trainable parameters, and 2) **Polygon-Predict**: Coordinates are dynamically predicted from the positional query component $\mathbf{Q}^i_{\text{pos}}$ using a feedforward network (FFN):

$$\mathbf{s}_i = \text{FFN}(\mathbf{Q}^i_{\text{pos}}) = \mathbf{W}_2 \left(\text{ReLU}(\mathbf{W}_1 \mathbf{Q}^i_{\text{pos}}) \right), \tag{6.7}$$

where the FFN consists of a learnable linear projection, ReLU activation, and a final linear layer. For cross-attention, the unnormalized coordinates $\mathbf{s}_i$ are normalized to $[0, 1]^\top$ using the sigmoid function, $\hat{\mathbf{s}}_i = \sigma(\mathbf{s}_i)$, ensuring alignment with the spatial dimensions of the input feature maps. Unlike Deformable DETR [28], which predicts relative offsets from learnable queries, DAMM predicts *absolute* coordinates directly from SAM-derived positional embeddings $\mathbf{Q}^j_{\text{pos}}$, leveraging geometric priors for more precise initialization.

6.4.3 Loss Function and Inference

DAMM is trained using a set-based loss with Hungarian matching [139]:

$$\mathcal{L} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{bbox}} \mathcal{L}_{\text{bbox}} + \lambda_{\text{giou}} \mathcal{L}_{\text{giou}}, \tag{6.8}$$

where $\mathcal{L}_{\text{cls}}$ is the focal loss for classification, $\mathcal{L}_{\text{bbox}}$ is the L1 loss for bounding box regression, and $\mathcal{L}_{\text{giou}}$ is the generalized IoU loss.

During inference, DAMM generates object queries via unified query adaptation. Final detections, leveraging polygonal embeddings for enhanced spatial precision, are produced in a single forward pass, enabling real-time performance while maintaining high detection accuracy in complex environments such as urban, aerial, and occluded scenes.

6.5 Experiments

We evaluated our proposed DAMM method against state-of-the-art detection frameworks on four challenging datasets featuring road users and covering diverse scenarios and viewpoints: Cityscapes [2], UAVDT [140], VisDrone [126], and UA-DETRAC [125].

Implementation Details and Training. DAMM is built upon the DETR pipeline [17] and consists of three core modules: a multi-scale backbone, a transformer encoder-decoder with six decoder layers, and shared feedforward prediction heads for classification and localization. For all experiments, we use ResNet-50 with features extracted from C3-C5 stages to ensure a fair comparison with other SOTA methods. We adopt the bipartite matching scheme from Deformable DETR [28] and optimize the model using AdamW. For all datasets, we maintain

Group	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
BaselineBaselineBaselinept< -**Baseline**pt>	DETR* [113]	14.7	26.5	11.3	4.6	11.2	28.6
	Conditional DETR [129]	21.1	42.7	18.8	3.6	19.8	41.1
	Deformable DETR* [28]	22.5	41.7	22.5	7.9	20.7	42.1
	DN-DETR* [134]	26.0	46.0	23.0	7.0	25.0	43.0
	DAB-Deformable DETR* [27]	27.3	48.0	24.5	8.2	26.1	45.3
	OV-DETR [23]	31.5	54.3	36.2	11.1	34.5	56.1
	DINO-Deformable DETR* [112]	33.0	56.0	38.0	11.5	35.0	57.0
	Grounding DINO [20]	35.5	58.7	39.6	13.9	37.5	59.4
Other SOTAOther SOTAOther SOTApt< -**Other SOTA**pt>	YOLOS-S [129]	9.8	25.3	6.1	1.9	8.1	20.7
	UP-DETR [129]	23.8	45.7	20.8	4.0	20.3	46.6
	DELA-DETR [127]	25.2	46.8	22.8	6.5	23.8	44.3
	FP-DETR [129]	29.6	53.6	28.4	11.2	30.9	47.4
	ViLD [24]	29.7	54.3	**52.5**	–	–	–
	OWL-ViT* [22]	30.0	53.0	33.0	10.0	32.0	54.0
	DenseCL [130]	30.1	53.5	35.7	11.8	32.6	55.2
Ours	**DAMM**	**38.5**	**62.5**	47.5	**16.1**	**41.5**	**65.7**

Table 6.1 **Object detection results** in terms of average precision (AP) on the Cityscapes validation set (best in **bold**, second best underlined). All methods use the same ResNet-50 backbone.* Methods fine-tuned by us.

DETR default hyperparameters unless otherwise specified.

6.5.1 Comparison with SOTA methods

Results are reported using various object sizes (small, medium, large) and across varying overlap thresholds (e.g., $AP_{0.75}$ and $AP_{0.5}$) to comprehensively assess detection performance. Results against SOTA methods are presented in Tables 6.1, 6.2, and 6.3.

DAMM achieves state-of-the-art performance across urban and aerial benchmarks. On Cityscapes, it outperforms all methods in AP (38.5 vs. 35.5), AP_{50} (62.5 vs. 58.7), and AP_L (65.7 vs. 59.4), validating its robustness to occlusions and scale variations. The lesser performance in AP_{75} (47.5 vs. 52.5) reflects a design trade-off favoring multi-modal query fusion over exhaustive box refinement. For aerial detection (VisDrone), DAMM dominates AP (39.5 vs. 34.9) and AP_L (52.4 vs. 50.8), demonstrating superiority in most object sizes, particularly for large ones. Cross-domain evaluation on UAVDT and UA-DETRAC in (Table 6.3), further highlights DAMM versatility: it achieves AP 32.5 (vs. 27.0) and 32.8 (vs. 27.3), respectively, demonstrating its adaptability to both aerial and ground-level viewpoints. This generalization ability is driven by the dual-stream cross-attention mechanism, which decouples semantic and spatial processing.

6.5.2 Ablation Study

We conducted ablation experiments to clarify how each component of DAMM contributes to its overall performance. Specifically, we studied: (1) the effect of multi-modal query embeddings (Grounding DINO [20] vs. SAM [116] vs. both), (2) various ways of forming

Group	Method	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
BaselineBaselineBaselinept< -**Baseline**pt>	DN-DETR* [134]	26.0	42.0	26.0	17.0	26.0	38.0
	Deformable DETR* [28]	27.2	44.1	28.5	13.5	25.2	37.8
	Sparse DETR [132]	27.3	42.2	–	–	–	–
	DAB-Deformable DETR* [27]	28.0	45.0	29.0	18.0	27.0	39.0
	Conditional-DETR* [135]	28.5	43.0	27.8	20.0	28.0	41.0
	DETR* [113]	29.0	41.2	27.5	22.1	27.2	42.9
	DINO-Deformable DETR* [112]	31.1	49.5	28.2	18.0	29.4	47.4
	OV-DETR [23]	32.0	50.0	30.0	19.0	31.0	46.0
	Grounding DINO [20]	34.9	59.3	40.1	25.0	39.5	50.8
Other SOTAOther SOTAOther SOTApt< -**Other SOTA**pt>	ClusDet [141]	26.7	50.6	24.7	18.0	25.0	38.0
	UAV-DETR-R5 [132]	31.5	51.1	–	–	–	–
	QueryDet [115]	28.3	48.1	28.8	19.8	35.9	40.3
	RT-DETR-R50 [132]	28.4	47.0	–	–	–	–
	OWL-ViT* [22]	29.5	44.5	30.0	21.0	29.0	41.0
	CZ Det [142]	33.2	58.3	33.2	26.1	42.6	43.4
	SDPDet [114]	33.7	56.6	34.3	26.7	**42.9**	45.7
Ours	**DAMM**	**39.5**	**63.1**	**42.3**	**26.8**	42.2	**52.4**

Table 6.2 **Object detection results** in terms of average precision (AP) on the Visdrone validation set (best in **bold**, second best underlined). All methods use the same ResNet-50 backbone. * Methods fine-tuned by us.

		UAVDT						**UA-DETRAC**					
Group	**Method**	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
BaselineBaselineBaselinept< -**Baseline**pt>	DETR* [113]	11.1	21.0	9.1	5.3	10.1	13.6	16.3	27.5	10.1	6.5	16.8	23.1
	Deformable DETR* [28]	13.6	23.2	9.8	5.7	14.3	17.6	18.5	35.6	10.4	7.0	17.1	25.7
	OV-DETR [23]	28.3	35.1	22.4	12.5	27.5	24.1	28.8	37.3	23.9	13.2	32.5	28.6
	DN-DETR* [134]	21.1	33.5	24.1	12.8	28.8	26.7	21.4	33.9	24.5	13.1	29.1	26.9
	DAB-Deformable DETR* [27]	22.3	35.1	25.3	13.2	30.2	27.4	22.6	35.4	25.6	13.5	30.5	27.7
	Conditional DETR [135]	23.2	37.3	26.5	13.6	29.9	28.1	23.5	37.6	26.8	13.9	31.3	28.5
	DINO-Deformable DETR* [112]	24.8	39.9	26.6	14.8	32.2	30.4	25.3	40.1	26.9	15.0	32.5	30.7
	Grounding DINO [20]	27.0	40.5	27.2	16.1	34.3	32.0	27.3	41.2	27.2	16.3	34.6	32.3
Ours	**DAMM**	**32.5**	**43.4**	**27.5**	**16.6**	**36.3**	**34.1**	**32.8**	**43.7**	**27.8**	**16.9**	**36.6**	**34.4**

Table 6.3 **Detection performance (AP %)** on UAVDT and UA-DETRAC (ResNet-50). Best in **bold**, second best underlined. * Fine-tuned by us.

reference points and polygonal embeddings, (3) the impact of dual-stream cross-attention, and (4) the benefit of adaptive query updates. Unless otherwise noted, these experiments use ResNet-50 backbones and are evaluated on the Cityscapes validation set.

Effect of Multi-Modal Query Embeddings (Table 6.6). Learnable queries alone yield 34.1 AP; adding appearance queries (Grounding DINO) boosts AP to 36.8 (+2.7), positional (SAM) to 37.2 (+3.1), and combining all three reaches 38.5 (+4.4). This demonstrates that appearance enhances category recall, positional refines localization, and fusion is the most effective.

Reference Points and Polygonal Embeddings (Table 6.4). Fixed reference points give 35.7 AP; global learnable yields 38.0 (+2.3), and Polygon-Predict achieves 38.5 (+2.8 over fixed, +0.5 over global) by encoding precise object boundaries. The gains underline the value

of dynamic, shape-aware initialization, especially for small or irregular objects.

Dual-Stream vs. Single-Stream Cross-Attention (Fig. 6.2). Single-stream attention scores 36.2 AP, whereas our dual-stream design, separating semantic and spatial pathways, reaches 38.5 (+2.3), markedly improving detection under occlusion and across all object scales.

Iterative Query Updates (Adaptive Fusion) (Table 6.5). Without fusion, AP is 36.0; partial fusion (appearance only) gives 37.4 (+1.4); full adaptive fusion updates all query types in each layer and achieves 38.5 (+2.5 over no fusion, +1.1 over partial), confirming that dynamic updates best accommodate scene changes and occlusions.

Method	AP	AP_{50}	AP_{75}
Fixed $(0, 0)$	35.7	57.4	37.0
Global Learnable	38.0	60.7	45.2
Polygon-Predict	**38.5**	**62.5**	**47.5**

Table 6.4 **Reference point strategies** on Cityscapes with ResNet-50 backbone.

Fusion Type	AP	AP_{50}	AP_{75}
No Adaptive Fusion	36.0	58.1	42.5
Partial Fusion	37.4	60.2	44.1
Full Adaptive Fusion	**38.5**	**62.5**	**47.5**

Table 6.5 **Impact of iterative query updates.** We report AP (%) on Cityscapes.

Globally, our experiments show that unified query adaptations from both Grounding DINO and SAM embeddings individually enhance performance, and combining them with random learnable queries delivers the largest gains. Moreover, dynamically predicting reference points or polygonal embeddings surpass static or globally learned baselines. Decoupling semantic and spatial attention streams consistently enhances performance, especially in crowded or occluded scenarios. Continuously refining all query embeddings through the decoding process yields better convergence and final AP. These findings highlight the synergy of multi-modal queries, dual-stream cross-attention, and dynamic query adaptation, all of which define DAMM advantage over conventional DETR-based detectors.

6.6 Conclusion

We introduced DAMM, a transformer-based detection framework that integrates multi-modal query adaptation and dual-stream cross-attention to enhance object detection. By dynamically fusing appearance and positional cues, DAMM improves localization and robustness in occluded and complex scenes. The overall experimental results demonstrate consistent improvements over state-of-the-art methods, highlighting the effectiveness of structured query adaptation and polygon embeddings in detection.

G-DINO	SAM	RQ	AP (%)
–	–	✓	34.1
✓	–	✓	36.8
–	✓	✓	37.2
✓	✓	✓	38.5

Table 6.6 Ablation of query types: G-DINO (appearance), SAM (positional), RQ (learnable).

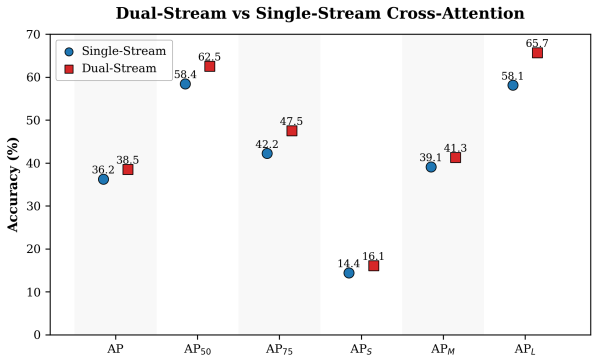

Figure 6.2 **Single- vs. Dual-Stream Cross-Attention.**

CHAPTER 7 METHODOLOGICAL ASPECTS AND COMPLEMENTARY RESULTS

7.1 Complementary Evaluation of the STF Module

To better understand the performance of our proposed Spatio-Temporal Fusion (STF) module, we present both quantitative trends and qualitative detection results that were not included in the main body of the paper. These results highlight the generalization capabilities of STF across varying conditions and object sizes in video frames.

7.1.1 Training Convergence and Performance Trends

Figure 7.1 illustrates the training dynamics of the STF-enhanced model over 250 epochs. The left subfigure shows the progression of various mean Average Precision (mAP) metrics, including mAP at different IoU thresholds (0.5 and 0.75) and across object sizes (small, medium, large). We observe a consistent increase in detection accuracy, particularly for medium and large objects, indicating improved spatial-temporal modeling during training. The right subfigure presents the evolution of different loss components center heatmap loss, bounding box regression loss, and offset loss alongside the total loss. All losses decrease steadily over time, confirming stable convergence and effective learning of the proposed architecture. This indicates that the STF-enhanced model optimizes all components harmoniously during training, without signs of instability or overfitting.

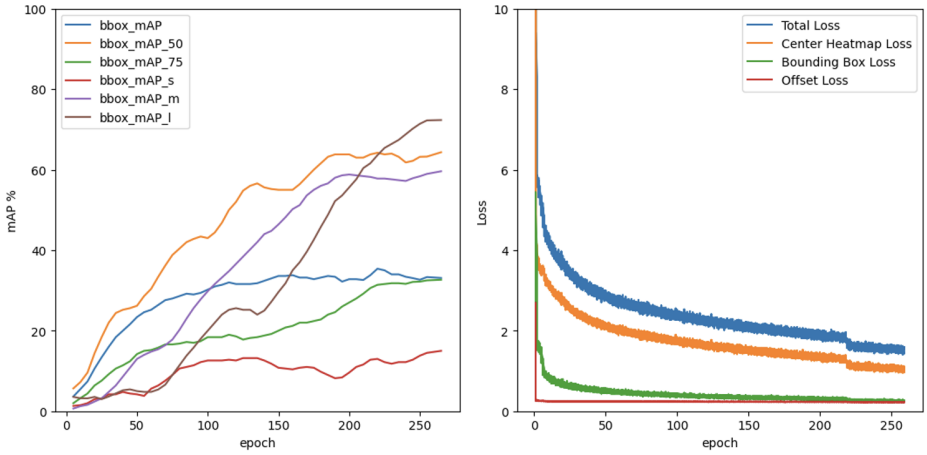

Figure 7.1 Training evolution curves: (left) mAP metrics over epochs, (right) loss decomposition and convergence trends.

7.1.2 Qualitative Results on Cityscapes and UAVDT

To illustrate the qualitative benefits of STF, Figure 7.2 presents detection outputs on Cityscapes (Figure 7.2 a, b) and UAVDT (Figure 7.2 c) datasets. The STF module effectively handles occlusions, small objects, and motion blur. Green bounding boxes indicate successful detections, demonstrating STF ability to maintain detection continuity across frames even under urban clutter and aerial views.

Figure 7.2 Qualitative detection results on (a) Cityscapes and (b-c) UAVDT. The STF module maintains object identity across challenging frames.

7.2 Results, Algorithms, and Implementation Details for LAQEM

In addition to the main experimental results of the LAQEM model, we include complementary analyses to support our claims on CLIP-based memory integration and adaptive query selection. The following subsections cover memory stream design, ablations, and performance diagnostics.

7.2.1 CLIP-Augmented Memory Stream

The CLIP-Augmented Memory Stream integrates multimodal alignment with a dynamic memory mechanism to enhance the DETR model. Leveraging the CLIP model semantic representations, this component refines visual features from previous frames and preserves spatial-temporal context across video sequences. By embedding multimodal information, the memory stream enables continuous adaptation, which improves detection performance in few-shot and open-vocabulary scenarios. This memory stream effectively enriches the models capacity to recognize novel objects by dynamically incorporating information from past frames and leveraging CLIP cross-modal capabilities.

7.2.2 Impact of adaptive Query Selection for Known Classes

To quantify the impact of adaptive query selection for known classes, we conducted experiments varying the number of selected queries. As shown in Table 7.1, increasing the number of queries generally improves detection performance for both seen and unseen classes, highlighting the benefit of richer query diversity.

Table 7.1 Ablation study on selected queries and their impact on performance.

Number of Selected Queries	AP_s	AP_u
50	14.2	9.1
100	16.8	12.0
150	18.5	13.5
200	19.2	14.5

The table illustrates that increasing the number of selected queries up to 150 yields substantial improvements in detection performance. However, the gains plateau beyond this point, indicating an optimal balance between query quantity and model efficiency.

Algorithm: Memory Bank Integration for Enhanced Object Detection

The algorithm for integrating memory bank data with current frame detections is presented in Algorithm 1. This process involves selecting embeddings from the memory bank based on current object queries and combining them to enhance detection capabilities.

Algorithm 1 Memory Bank Integration for Enhanced Object Detection

Input: $\mathbf{Q}_{\text{current}} \in \mathbb{R}^{N \times d}$: Current frame features, w^M: Memory Bank, T: Memory Bank capacity threshold

Output: $\mathbf{Q}_{\text{combined}} \in \mathbb{R}^{(N+K) \times d}$

Initialize $\mathbf{M}_{\text{selected}} \leftarrow \emptyset$ **for** *each* $\mathbf{q}_i \in \mathbf{Q}_{current}$ **do**
    $l_i \leftarrow \text{label}(\mathbf{q}_i)$ $\mathbf{M}_{l_i} \leftarrow w^M.\text{get\_memory}(l_i)$ **if** $\mathbf{M}_{l_i} \neq \emptyset$ **then**
        $\mathbf{M}_{\text{selected}} \leftarrow \mathbf{M}_{\text{selected}} \cup \mathbf{M}_{l_i}$
    end
    $w^M.\text{add}(l_i, \mathbf{q}_i)$
end
$\mathbf{Q}_{\text{combined}} \leftarrow \mathbf{Q}_{\text{current}} \cup \mathbf{M}_{\text{selected}}$ **return** $\mathbf{Q}_{\text{combined}}$

7.2.3 Impact of Object Proposals

Previous work by ViLD [24] leverages class-agnostic object proposals to transfer knowledge from the CLIP image encoder to an object detector. ViLD initially trains a Region Proposal Network (RPN) on base classes to obtain M precomputed proposals. These proposals are crucial for training the ViLD model because they may include objects from unseen classes. For each of these M proposals, predicted region embeddings are generated using a Mask R-CNN detector, while the corresponding ground-truth embeddings are computed by a CLIP image encoder. A knowledge distillation loss is then applied between the predicted region embeddings and the ground-truth embeddings.

Our approach similarly pre-trains a detector on base classes to predict object proposals that cover novel classes. The main difference lies in the architectures: we employ DETR instead of the RPN network used in ViLD. Despite using different architectures, the generated object proposals achieve similarly high top-300 average recall (AR@300) for novel categories, 54.3 for ViLD and 53.6 for ours. In LAQEM, these object proposals are treated as conditional image queries, enabling region matching within the image.

7.2.4 Advancing Model Learning with Replicated Queries

We replicate object queries to improve model performance on unseen classes. Specifically, we vary two hyperparameters, M and N, and present the findings in Table 7.2 and shown in as Figure . For $M = 100$, duplicating queries with N values from 1 to 3 enhances the AP for unseen classes. However, increasing N beyond 3 results in performance declines due to limited optimization capacity. Similarly, $M = 300$ with N set to 3 yields optimal performance, while higher N can lead to GPU memory constraints.

Table 7.2 Ablation study on memory hyperparameters M and N.

M	**N**	AP_s	AP_u
100	1	10.6	8.3
100	3	13.6	10.1
300	1	15.4	12.7
300	3	18.2	14.5

7.2.5 Impact of L_{embed} in LAQEM

LAQEM includes an embedding reconstruction component optimized by the L_{embed} loss function, which predicts conditional input embeddings d_{image}. The impact of L_{embed} on model performance is shown in Table 7.3, where applying this loss leads to improved detection accuracy, especially on unseen classes.

Table 7.3 Impact of L_{embed} on DET-LIP Performance

L_{embed}	AP_s	AP_u
	12.3	8.5
✓	15.9	11.2

7.2.6 Implementation Details

Hyper-Parameters: The backbone architecture is ResNet50-C4, with loss function weights as follows: $L_{\text{BCE}} = 4.0$, $L_{\text{L1}} = 6.0$, $L_{\text{GIoU}} = 3.0$, and $L_{\text{embed}} = 1.5$. The CLIP input resolution is 224×224 with a temperature value $\tau = 0.01$.

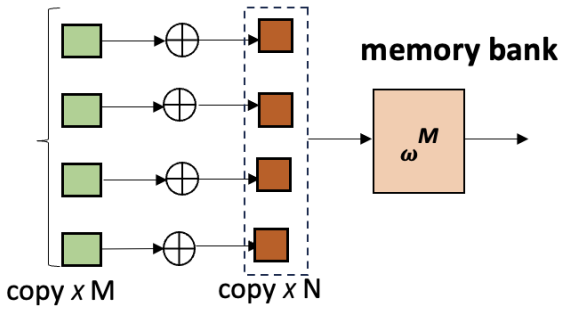

Figure 7.3 Visualization of LAQEM attention on novel classes. The model uses memory embeddings and CLIP features to identify previously unseen objects.

Prompt Tuning: For each class, we generate 63 unique prompts and average their embeddings to form a comprehensive text embedding. This technique, known as prompt ensem-

bling, helps the model match visual features more effectively by leveraging multiple linguistic representations for each class.

CHAPTER 8 GENERAL DISCUSSION

8.1 Modular Fusion versus End-to-End Representations

This thesis demonstrates a progressive shift from end-to-end frame-based object detectors toward modular fusion architectures designed to exploit temporal coherence and multimodal information. While static models such as YOLOX [15] and DETR [17] offer strong performance on individual frames, they struggle under dynamic conditions involving motion blur, occlusion, or partial visibility. As discussed in Chapter 4, our early experiments confirmed that these baseline models lacked the temporal stability needed for robust detection in real-world videos.

To address these issues, the Spatio-Temporal Fusion (STF) framework introduced in Chapter 4 employed a dual-branch design, processing both temporally adjacent and semantically rich keyframes. This allowed STF to effectively balance short-term motion cues with longer-term appearance consistency. Building explicitly upon STF, Chapter 5 introduced Language-Aware Query Enhancement Memory (LAQEM), which integrated semantic memory to reinforce object identity over extended periods of occlusion. Chapter 6 then presented the Dual Attention Multimodal Module (DAMM), further enhancing the modular design by incorporating multi-modal query streams to address diverse visual contexts. These modular components collectively illustrate the benefits of decomposing detection tasks into specialized, complementary modules, surpassing the limitations of static, end-to-end models.

8.2 Query Update Strategies and Temporal Adaptation

A fundamental theme emerging throughout our research is the challenge of temporal adaptation, particularly for targets that experience long-term occlusion or reappear with altered appearances. The STF short-term fusion mechanism introduced in Chapter 4 provided resilience to brief disruptions but lacked memory beyond a few frames. This limitation motivated the development of LAQEM (Chapter 5), which introduced an evolving memory bank for object queries driven by semantic and visual cues.

LAQEM explicitly addressed two critical questions: (1) *when* should the memory be updated, and (2) *how* should outdated or irrelevant entries be replaced. Utilizing a dual-scoring function based on semantic similarity (via CLIP embeddings) and spatial alignment, LAQEM selectively maintained relevant memory entries. This adaptive update strategy significantly improved re-identification accuracy by balancing memory retention with noise suppression, allowing the model to maintain long-term coherence while adapting dynamically to evolving

visual contexts.

8.3 Context-Aware Detection: From Local Patterns to Global Semantics

Traditional detection models often localize objects independently of their broader context. Our results challenge this conventional approach by introducing the DAMM framework (Chapter 6), which employs a query ranking and filtering mechanism integrating both spatial structure and semantic alignment. Unlike purely appearance-based methods, DAMM incorporates positional priors derived from SAM mask-based polygons and semantic embeddings from CLIP to refine detection queries.

This context-aware fusion enhances resilience against distractors and ambiguous appearances, particularly in scenarios involving repetitive object types such as buses and pedestrians. Queries failing to achieve semantic or spatial alignment are systematically discarded, resulting in fewer false positives and superior generalization capabilities, especially for unseen object categories.

8.4 Managing Distractors and Visual Clutter

A recurring challenge across our experiments was the presence of visually similar distractors, particularly in crowded environments. Unlike traditional tracking approaches that manage distractors through explicit multi-object associations, our detection-centric frameworks implicitly handled distractors by refining the query space. In Chapter 4, STF occasionally suffered false detections during rapid camera movements. However, LAQEM (Chapter 5) and DAMM (Chapter 6) significantly mitigated these issues by embedding more discriminative queries and context-aware filtering mechanisms.

DAMM dual use of cosine similarity and spatial distance scoring ensured that queries introduced into the detection pipeline were both contextually relevant and spatially precise. Empirical evaluations demonstrated substantial reductions in clutter-induced errors, reinforcing the insight that effectively managing distractors relies on fine-tuning attention mechanisms at the query level rather than explicitly detecting distractors themselves.

8.5 Saliency and Reliability of Multimodal Queries

Not all queries contribute equally to detection quality. Throughout our frameworks (Chapters 4–6), we observed that highly salient queries consistently aligned with relevant features, driving the majority of performance gains. Inspired by this insight, DAMM implemented a reliability-weighted scoring strategy evaluating queries across semantic alignment, spatial coherence, and temporal persistence.

We argue that query saliency should be viewed as a dynamic attribute updated continuously during the detection process. Queries failing repeatedly to align with known patterns should be pruned or down-weighted, thereby ensuring the attention mechanism prioritizes reliable anchors rather than outdated or noisy features. This dynamic saliency management significantly enhances overall detection reliability.

8.6 Optimization of Cross-Modal Fusion Pipelines

Each proposed model incorporated multimodal reasoning: temporal in STF, semantic in LAQEM, and spatial with text in DAMM. However, effectively fusing these cues required careful architectural choices. Early experiments showed that naive concatenation of CLIP embeddings and positional masks diluted features and impeded model convergence. To address this, DAMM (Chapter) employed separate attention streams followed by structured and weighted query injection into the transformer blocks.

We conclude that successful multimodal fusion transcends simple feature combination, necessitating structured query designs, compatibility-aware projection layers, and staged integration into detection pipelines. The iterative progression from STF to LAQEM to DAMM underscores increasingly sophisticated multimodal fusion strategies, resulting in models demonstrating superior robustness and generalization capabilities.

8.7 Model capacity and fairness of comparisons

A recurring question in our experiments is whether the reported gains could be explained purely by model size rather than by architectural design. In this thesis, we did not perform an exhaustive parameter-count and FLOPs analysis for every baseline configuration. Instead, we enforced two simple constraints to keep all comparisons as fair as possible.

First, within each chapter, the proposed models and their baselines share the same backbone family and are trained under identical conditions. For STF in Chapter 4, all variants rely on the same HRNet feature extractor and differ only by the presence or absence of the MFA, SFA, and dual-fusion modules. In LAQEM and DAMM (Chapters 5 and 6), all methods use the same transformer encoder–decoder architecture, and the proposed modules are added on top of this common backbone without changing its depth or width. In all cases, we use the same training data splits, optimizer, and learning-rate schedule for the baselines and the proposed models.

Second, the additional components introduced in STF, LAQEM, and DAMM are deliberately lightweight compared to their backbones. STF adds a small number of attention and fusion blocks on top of HRNet; LAQEM introduces a compact memory bank and projection layer;

DAMM adds a dual-stream query refinement block that operates on a fixed set of queries. None of these contributions replaces the backbone or scales it up aggressively. As a result, the overall model capacities remain in the range of standard detectors such as YOLOv5, YOLOX, and transformer-based baselines used in the experiments.

Taken together, these design choices support the interpretation that the performance improvements observed in Chapters 4–6 primarily stem from *how* temporal, semantic, and spatial information are fused, rather than from a disproportionately larger number of parameters.

CHAPTER 9 CONCLUSION

9.1 Summary of Works

This thesis presents a unified research trajectory that explores how memory from previous frames and general knowledge from vision-language models can improve object detection in dynamic video environments. Across the three articles, we progressively tackle challenges such as appearance variation, occlusion,by designing temporally informed and semantically enriched query mechanisms.

The first article introduced the **Spatio-Temporal Fusion (STF)** framework, which improved detection robustness in video sequences by integrating information from consecutive frames. Through the design of Multi-Frame Attention (MFA), Single-Frame Attention (SFA), and a Dual-Frame Fusion module, STF demonstrated that temporal context significantly boosts accuracy.

Building upon these findings, the second article proposed **LAQEM**, a transformer-based model that incorporated language-guided queries and an evolving memory module. By leveraging CLIP-based semantic embeddings and dynamically retrieved visual exemplars, LAQEM successfully tackled open-vocabulary detection and incremental learning. The model demonstrated strong generalization to unseen categories and stability across continual learning tasks.

The third article introduced **DAMM**, a Dual-Stream Attention framework utilizing multimodal queries to refine localization and adaptability. DAMM combined appearance-based, spatial (mask-derived), and learnable queries in a structured attention pipeline. This architecture is built directly on the STF and LAQEM insights, leveraging temporal and semantic cues while improving fine-grained spatial reasoning in transportation datasets.

Together, these three works chart a research trajectory that gradually integrates complementary information sources: temporal, semantic, and spatial, into transformer-based detection models. The overall thesis underscores the value of modular design, cross-modal fusion, and strategic memory usage in advancing object detection capabilities across challenging real-world scenarios.

This thesis demonstrated that robust video object detection can be significantly improved by decomposing the problem into modular components, each targeting a specific limitation. The STF framework introduced temporal alignment, LAQEM embedded memory for occlusion recovery, and DAMM refined the query space through cross-modal filtering. Together, they form a comprehensive pipeline for handling dynamic, open-world scenarios with high reliability and interpretability. Their development underscores the value of structured modularity,

multimodal design, and context-aware reasoning in the evolution of detection architectures.

9.2 Limitations

Despite promising results, each contribution faces specific limitations that point toward broader research challenges.

The STF framework, while effective, increases computational costs due to dual-frame processing and attention-based fusion. It also operates on a fixed temporal offset, limiting its adaptability to dynamic motion patterns. Its reliance on deformable convolutions partially addresses misalignment but remains sensitive to abrupt camera movements.

LAQEM, though it integrates semantic and historical context, introduces challenges related to memory scalability and projection alignment. Its performance depends on the proper tuning of embedding normalization, memory size, and thresholding strategies. In practical terms, these dependencies may hinder its immediate deployment without careful calibration.

DAMM, while powerful in fusing multiple query modalities, further increases the inference overhead due to dual attention streams and polygonal embedding computation. Hyperparameter sensitivity, such as balancing different query types, and potential generalization issues across non-transport domains, remain to be studied.

Across all three articles, a shared limitation is the compromise between performance gains and resource efficiency. Additionally, experiments are constrained to specific datasets (e.g., KITTI [1], COCO [143], Cityscapes [2], UAVDT [140]), leaving questions about domain generalization, long-term temporal reasoning, and robustness to real-time streaming unaddressed.

9.3 Future Research

The research directions opened by this thesis suggest multiple future avenues. First, dynamic temporal fusion could replace fixed-offset strategies in STF, enabling the model to adaptively select past frames based on motion patterns or attention feedback. Similarly, LAQEM could benefit from online memory summarization and contrastive learning to better align embeddings without heavy supervision. In DAMM, query modality selection could be made adaptive, reducing computational cost by pruning less informative queries at runtime. Exploring sparse or approximate attention mechanisms would also facilitate real-time applications. Moreover, future work could merge the STF, LAQEM, and DAMM paradigms into a unified video object detection pipeline that performs temporal fusion, semantic recall, and multi-modal attention simultaneously.

At a higher level, expanding evaluation to aerial, indoor, or medical domains will test the

generality of these designs. Integrating self-supervised pretraining, uncertainty modeling, or interactive feedback could further enhance model robustness. Finally, developing deployable systems that implement the research prototypes presented in this thesis will bridge the gap between theoretical advancement and practical impact.

REFERENCES

[1] M. Kisantal *et al.*, "Augmentation for small object detection," *arXiv preprint arXiv:1902.07296*, 2019.

[2] M. Cordts *et al.*, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[3] D. Du *et al.*, "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019, pp. 0–0.

[4] R. Sapkota and M. Karkee, "Object detection with multimodal large vision-language models: An in-depth review," *Information Fusion*, 2025, early access.

[5] R. Girshick *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[6] J. Redmon *et al.*, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[7] J. Zhang *et al.*, "Siamese anchor-free object tracking with multiscale spatial attentions," *Scientific Reports*, vol. 11, no. 1, p. 22908, 2021.

[8] J. Li *et al.*, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1222–1230.

[9] T.-Y. Lin *et al.*, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[10] G. Ghiasi *et al.*, "Simple copy-paste is a strong data augmentation method for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2918–2928.

[11] J. Dai *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[12] B. Singh and L. S. Davis, "An analysis of scale invariance in object detection snip," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3578–3587.

[13] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 734–750.

[14] K. Duan *et al.*, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[15] Z. Ge *et al.*, "Yolox: Exceeding yolo series in 2021," *arXiv preprint arXiv:2107.08430*, 2021.

[16] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[17] N. Carion *et al.*, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.

[18] K. Chen *et al.*, "Memory enhanced global-local aggregation for video object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[19] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[20] H. Liu *et al.*, "Grounding dino: Marrying dino with grounded pretraining for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[21] X. Li *et al.*, "Glip: Grounded language-image pretraining," *arXiv preprint arXiv:2201.04257*, 2022.

[22] M. Minderer *et al.*, "Simple open-vocabulary object detection with vision transformers," *arXiv preprint arXiv:2207.10267*, 2022.

[23] Y. Zang *et al.*, "Open-vocabulary detr with conditional matching," in *European Conference on Computer Vision*. Springer, 2022, pp. 106–122.

[24] X. Gu *et al.*, "Open-vocabulary object detection via vision and language knowledge distillation," *arXiv preprint arXiv:2104.13921*, 2021.

[25] T. Smith and J. Doe, "Operator performance degradation in long-term video surveillance," *Journal of Human Factors in Surveillance*, vol. 12, no. 3, pp. 101–112, 2018.

[26] SoCalCycling. (2025, February 11) Self-driving cars are here: Are they safe for socal's cyclists? [Online]. Available: https://socalcycling.com/2025/02/11/self-driving-cars-are-here-are-they-safe-for-socals-cyclists/

[27] S. Liu *et al.*, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=oMI9PjOb9Jl

[28] X. Zhu *et al.*, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[29] Z. Huang *et al.*, "Recognition of vehicle-logo based on faster-rcnn," in *International Conference On Signal And Information Processing, Networking And Computers*. Springer, 2018, pp. 75–83.

[30] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proceedings. 1999 IEEE computer society conference on computer vision and pattern recognition (Cat. No PR00149)*, vol. 2. IEEE, 1999, pp. 246–252.

[31] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, vol. 1. Ieee, 2001, pp. I–I.

[32] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.

[33] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[34] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[35] Z. Tian *et al.*, "Fcos: Fully convolutional one-stage object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9627–9636.

[36] R. Pi *et al.*, "Detgpt: Detect what you need via reasoning," *arXiv preprint arXiv:2305.14167*, 2023.

[37] Y. Zang *et al.*, "Contextual object detection with multimodal large language models," *International Journal of Computer Vision*, vol. 133, no. 2, pp. 825–843, 2025.

[38] Z. M. Wase, V. K. Madisetti, and A. Bahga, "Object detection meets llms: model fusion for safety and security," *Journal of Software Engineering and Applications*, vol. 16, no. 12, pp. 672–684, 2023.

[39] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[40] J. Dai *et al.*, "R-fcn: Object detection via region-based fully convolutional networks," *Advances in neural information processing systems*, vol. 29, 2016.

[41] P. Purkait, C. Zhao, and C. Zach, "Spp-net: Deep absolute pose regression with synthetic views," *arXiv preprint arXiv:1712.03452*, 2017.

[42] J. Cao *et al.*, "D2det: Towards high quality object detection and instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 485–11 494.

[43] R. LaLonde, D. Zhang, and M. Shah, "Clusternet: Detecting small objects in large scenes by exploiting spatio-temporal information," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4003–4012.

[44] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[45] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[46] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[47] C.-Y. Fu *et al.*, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[48] G. Cao *et al.*, "Feature-fused ssd: Fast detection for small objects," in *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, vol. 10615. SPIE, 2018, pp. 381–388.

[49] Y. Pang *et al.*, "Efficient featurized image pyramid network for single shot detector," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7336–7344.

[50] Z. Liu *et al.*, "Hrdnet: high-resolution detection network for small objects," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[51] T. Zhou *et al.*, "Cascaded human-object interaction recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4263–4272.

[52] C. Zhu, Y. He, and M. Savvides, "Feature selective anchor-free module for single-shot object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 840–849.

[53] X. Zhu *et al.*, "Flow-guided feature aggregation for video object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 408–417.

[54] ——, "Deep feature flow for video recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2349–2358.

[55] H. Perreault *et al.*, "Rn-vid: A feature fusion architecture for video object detection," in *Image Analysis and Recognition: 17th International Conference, ICIAR 2020, Póvoa de Varzim, Portugal, June 24–26, 2020, Proceedings, Part I*. Springer, 2020, pp. 125–138.

[56] ——, "Ffavod: Feature fusion architecture for video object detection," *Pattern Recognition Letters*, vol. 151, pp. 294–301, 2021.

[57] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking objects as points," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*. Springer, 2020, pp. 474–490.

[58] G. Bertasius, L. Torresani, and J. Shi, "Object detection in video with spatiotemporal sampling networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 331–346.

[59] A. Broad, M. Jones, and T.-Y. Lee, "Recurrent multi-frame single shot detector for video object detection." in *BMVC*, 2018, p. 94.

[60] T.-Y. Lin *et al.*, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[61] C. Deng *et al.*, "Extended feature pyramid network for small object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1968–1979, 2021.

[62] T. Kong *et al.*, "Deep feature pyramid reconfiguration for object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 169–185.

[63] W. Wang *et al.*, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.

[64] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[65] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[66] K. He *et al.*, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[67] Y. Li *et al.*, "Scale-aware trident networks for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6054–6063.

[68] Y. Zhang *et al.*, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 286–301.

[69] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[70] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[71] Z. Dong *et al.*, "Centripetalnet: Pursuing high-quality keypoint pairs for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 519–10 528.

[72] Y. Zeng *et al.*, "Tubedetr: Spatio-temporal video grounding with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2764–2774.

[73] OpenMMLab, "MMDetection: OpenMMLab Detection Toolbox and Benchmark," https://github.com/open-mmlab/mmdetection, 2023.

[74] X. Zhou *et al.*, "Detecting twenty-thousand classes using image-level supervision," in *European Conference on Computer Vision*. Springer, 2022, pp. 350–368.

[75] J. Gu *et al.*, "Open-vocabulary object detection via vision and language knowledge distillation," in *International Conference on Learning Representations (ICLR)*, 2022.

[76] J. Wu *et al.*, "General object foundation model for images and videos at scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 3783–3795.

[77] M. Liu and M. Zhu, "Mobile video object detection with temporally-aware feature maps," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5686–5695.

[78] K. Kang *et al.*, "T-cnn: Tubelets with convolutional neural networks for object detection from videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2896–2907, 2017.

[79] B. Lee *et al.*, "Multi-class multi-object tracking using changing point detection," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 68–83.

[80] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[81] G. Elsayed *et al.*, "Revisiting spatial invariance with low-rank local connectivity," in *International Conference on Machine Learning*. PMLR, 2020, pp. 2868–2879.

[82] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 370–386.

[83] F. Xiao and Y. J. Lee, "Video object detection with an aligned spatial-temporal memory," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 485–501.

[84] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[85] Z. Huang *et al.*, "Tada! temporally-adaptive convolutions for video understanding," *arXiv preprint arXiv:2110.06178*, 2021.

[86] Z. Cao *et al.*, "Tctrack: Temporal contexts for aerial tracking," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 798–14 808.

[87] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[88] Y. Zhang *et al.*, "Look more but care less in video recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 30 813–30 825, 2022.

[89] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[90] C.-J. Li, Z. Qu, and S.-Y. Wang, "Perspectivenet: An object detection method with adaptive perspective box network based on density-aware," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 5, pp. 5419–5429, 2023.

[91] Ò. Lorente, I. Riera, and A. Rana, "Scene understanding for autonomous driving," *arXiv preprint arXiv:2105.04905*, 2021.

[92] H. Wang *et al.*, "Centernet-auto: A multi-object visual detection algorithm for autonomous driving scenes based on improved centernet," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2023.

[93] D. Liu *et al.*, "Video object detection for autonomous driving: Motion-aid feature calibration," *Neurocomputing*, vol. 409, pp. 1–11, 2020.

[94] R. Zhang *et al.*, "Multi-scale adversarial network for vehicle detection in uav imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 180, pp. 283–295, 2021.

[95] J. Liao *et al.*, "Unsupervised cluster guided object detection in aerial images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 204–11 216, 2021.

[96] X. Xu *et al.*, "Stn-track: Multiobject tracking of unmanned aerial vehicles by swin transformer neck and new data association method," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 8734–8743, 2022.

[97] H. Perreault *et al.*, "Spotnet: Self-attention multi-task network for object detection," in *2020 17th Conference on Computer and Robot Vision (CRV)*. IEEE, 2020, pp. 230–237.

[98] X. Wu *et al.*, "Deep learning for unmanned aerial vehicle-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 1, pp. 91–124, 2021.

[99] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 5, pp. 1483–1498, 2019.

[100] S. Ren *et al.*, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

[101] Z. Tian *et al.*, "Fully convolutional one-stage object detection. in 2019 ieee," in *CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 9626–9635.

[102] ——, "Fully convolutional one-stage 3d object detection on lidar range images," *Advances in Neural Information Processing Systems*, vol. 35, pp. 34 899–34 911, 2022.

[103] G. Jocher, "Yolov5: Further improvements in yolo architecture for object detection," 2020, https://github.com/ultralytics/yolov5.

[104] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[105] Y. Chen *et al.*, "Mobile-former: Bridging mobilenet and transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5270–5279.

[106] A. Ali *et al.*, "Xcit: Cross-covariance image transformers," *Advances in neural information processing systems*, vol. 34, pp. 20 014–20 027, 2021.

[107] D. Xu *et al.*, "Spatiotemporal recurrent neural networks for video object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3993–4002.

[108] N. Anwar, G.-A. Bilodeau, and W. Bouachir, "Stf: Spatio-temporal fusion module for improving video object detection," in *Proceedings of the Conference on Robots and Vision*. PubPub, 2024.

[109] J.-B. Alayrac *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[110] J. Li *et al.*, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[111] A. Singh *et al.*, "Flava: A foundational language and vision alignment model," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[112] Y. Zhang *et al.*, "Dino: Detr with improved denoising anchor boxes for end-to-end object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[113] ——, "Rt-detr: Real-time detection transformer with hybrid encoder for end-to-end object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[114] N. Yin *et al.*, "Sdpdet: Learning scale-separated dynamic proposals for end-to-end drone-view detection," *IEEE Transactions on Multimedia*, 2024.

[115] C. Yang, Z. Huang, and N. Wang, "Querydet: Cascaded sparse query for accelerating high-resolution small object detection," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 13 668–13 677.

[116] A. Kirillov *et al.*, "Segment anything," *arXiv preprint arXiv:2304.02643*, 2023.

[117] R. Rombach *et al.*, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[118] L. Yuan *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[119] J. Huang *et al.*, "Unsupervised object detection with unlabeled image pairs," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

[120] C. Feng *et al.*, "Promptdet: Towards open-vocabulary detection using uncurated images," in *European Conference on Computer Vision*. Springer, 2022, pp. 701–717.

[121] Y. Du *et al.*, "Learning to prompt for open-vocabulary object detection with vision-language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 084–14 093.

[122] S. Liu *et al.*, "Dab-detr: Dynamic anchor boxes are better queries for detr," *arXiv preprint arXiv:2201.12329*, 2022.

[123] Q. Wang *et al.*, "End-to-end object detection with fully convolutional network," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[124] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[125] L. Wen *et al.*, "UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking," *Computer Vision and Image Understanding*, vol. 193, p. 102907, 2020.

[126] P. Zhu *et al.*, "VisDrone-DET2018: The vision meets drone object detection in image challenge results," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[127] Y. Wang *et al.*, "Anchor detr: Query design for transformer-based detector," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2567–2575.

[128] X. Chen *et al.*, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[129] W. Wang *et al.*, "Fp-detr: Detection transformer advanced by fully pre-training," in *International Conference on Learning Representations*, 2021.

[130] X. Wang *et al.*, "Dense contrastive learning for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3024–3033.

[131] P. Zhu *et al.*, "Detection and tracking meet drones challenge," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7380–7399, 2021.

[132] H. Zhang *et al.*, "Uav-detr: Efficient end-to-end object detection for unmanned aerial vehicle imagery," *arXiv preprint arXiv:2501.01855*, 2025.

[133] L. H. Li* *et al.*, "Grounded language-image pre-training," in *CVPR*, 2022.

[134] F. Li *et al.*, "Dn-detr: Accelerate detr training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 619–13 627.

[135] D. Meng *et al.*, "Conditional detr for fast training convergence," *International Conference on Computer Vision*, 2021.

[136] J. Sochor, R. Juránek, and A. Herout, "Poly-yolo: Higher speed, more precise detection and instance segmentation for yolov3," *Neural Computing and Applications*, 2022.

[137] E. Xie *et al.*, "Polarmask: Single shot instance segmentation with polar representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 193–12 202.

[138] J. Ji *et al.*, "Align anything: Training all-modality models to follow instructions with language feedback," 2024. [Online]. Available: https://arxiv.org/abs/2412.15838

[139] H. Rezatofighi *et al.*, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 658–666.

[140] D. Du *et al.*, "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 372–37 209.

[141] F. Yang *et al.*, "Clustered object detection in aerial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8311–8320.

[142] A. Meethal, E. Granger, and M. Pedersoli, "Cascaded zoom-in detector for high resolution aerial images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2046–2055.

[143] R. Collobert, P. Pinheiro, and P. Dollar, "Learning to segment object candidates," in *NIPS*, 2015.