Data-Efficient Visual Recognition and Localization

FANGYUN WEI

Doctor of Philosophy

Supervisor: Associate Professor Chang Xu
Associate Supervisor: Dr Siqi Ma

A thesis submitted in fulfilment of
the requirements for the degree of
Doctor of Philosophy

School of Computer Science
Faculty of Engineering
The University of Sydney
Australia

22 February 2026

Statement of Originality

This is to certify that, to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purpose.

I certify that the intellectual content of this thesis is the product of my own work and that all assistance received in preparing this thesis, as well as all sources, has been acknowledged.

No content generated by generative AI tools has been used in the preparation of this thesis.

Student Name Signature

Fangyun Wei

Authorship Attribution Statement

All content included in this thesis is based on my prior publications. Chapter 3 is derived from the work presented in Zhao et al. (2024). Chapter 4 is based on the study in Yan et al. (2025). Chapter 5 builds upon the research reported in Wei et al. (2025). Chapter 6 is adapted from the work in Wei et al. (2024). For the publications (Zhao et al., 2024; Yan et al., 2025), I prepared the initial drafts, and the experiments were conducted in collaboration with my co-authors. For the works (Wei et al., 2025, 2024), I completed all major components, including manuscript writing, code development, experiment design and implementation, and result analysis. In all cases where I am not the corresponding author of a published item, permission to include the published material in this thesis has been granted by the corresponding author.

Student Name Signature
Fangyun Wei

As the supervisor of the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Supervisor Name Signature
Chang Xu

List of Research Outcome

Research outcome covered in this thesis

(1) **Wei, Fangyun***, Jinjing Zhao*, Kun Yan, and Chang Xu. "Minimizing Labeled, Maximizing Unlabeled: An Image-Driven Approach for Video Instance Segmentation." In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19304-19314. 2025.

(2) **Wei, Fangyun***, Jinjing Zhao*, Kun Yan, Hongyang Zhang, and Chang Xu. "A large-scale human-centric benchmark for referring expression comprehension in the LMM era." *Advances in Neural Information Processing Systems* 37 (2024): 69566-69587.

(3) Yan, Kun*, **Wei, Fangyun***, Shuyu Dai, Minghui Wu, Ping Wang, and Chang Xu. "Low-shot Video Object Segmentation." *IEEE transactions on pattern analysis and machine intelligence* (2025).

(4) Zhao, Jinjing*, **Wei, Fangyun***, and Chang Xu. "Hybrid proposal refiner: Revisiting detr series from the faster r-cnn perspective." In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pp. 17416-17426. 2024.

Other published research outcome

(1) Yan, Kun, Zied Bouraoui, **Wei Fangyun**, Chang Xu, Ping Wang, Shoaib Jameel, and Steven Schockaert. "Modeling Multi-modal Cross-interaction for Multi-label Few-shot Image Classification Based on Local Feature Selection." *ACM Transactions on Multimedia Computing, Communications and Applications* 21, no. 3 (2025): 1-28.

(The star '*' indicates equal contribution.)

Abstract

Over the past decade, computer vision has advanced rapidly, evolving from visual understanding to visual generation. Visual understanding, which aims to extract semantic and geometric information from visual signals, has become a core capability behind many real-world applications, including autonomous driving and robot perception, large-scale image/video search and recommendation, medical image analysis, intelligent surveillance, and augmented/virtual reality.

Among visual understanding tasks, object detection and segmentation in images and videos have emerged as foundational research problems. Object detection focuses on localizing and recognizing object instances, typically by predicting bounding boxes and associated class labels. Segmentation provides finer-grained understanding by assigning labels to pixels: semantic segmentation categorizes each pixel into a class, while instance segmentation further separates different object instances within the same class. Extending these tasks to videos introduces additional challenges such as motion blur, occlusion, and appearance changes, and requires modeling temporal consistency; in particular, video instance segmentation seeks to detect, segment, and track object instances across frames. These problems are fundamental because they provide structured representations that serve as reusable primitives for downstream reasoning and decision-making, and they support a wide spectrum of applications such as scene understanding for robotics, content-aware video editing, visual analytics, human-computer interaction, and safety-critical perception in autonomous systems.

As model capacity and task complexity continue to grow, data has become increasingly important. Modern vision systems often rely on learning-rich representations from large-scale datasets to generalize across diverse scenes, object categories, and imaging conditions. However, acquiring high-quality labeled data remains costly and challenging. Dense annotations, such as pixel-accurate masks in videos, require substantial human effort, careful quality control, and domain expertise in certain scenarios, such as medical or industrial inspection.

Moreover, long-tail categories, rare events, and domain shifts further exacerbate the difficulty of building comprehensive labeled datasets.

To address these challenges, data-efficient learning has attracted significant attention. Data-efficient learning aims to achieve strong performance using limited labeled data by leveraging abundant unlabeled data through techniques such as semi-supervised learning, self-training with pseudo labels, consistency regularization, and representation pretraining. In this thesis, we investigate data-efficient learning for visual understanding and demonstrate its effectiveness on several representative and widely-used tasks like object detection, video instance segmentation and video object segmentation. We develop practical learning frameworks that reduce annotation dependency while maintaining competitive accuracy, providing a step toward scalable and deployable vision systems under realistic data constraints.

Acknowledgements

I would like to express my deepest gratitude to everyone who has supported and contributed to the completion of my doctoral studies. This journey has been shaped by the guidance, encouragement, and kindness of many people, and I am sincerely thankful to all of them.

I began my PhD on January 1, 2023, and have spent three wonderful years at The University of Sydney. This period has been both challenging and rewarding, and it has helped me grow not only as a researcher but also as an individual.

First and foremost, I would like to thank my supervisor, Prof. Chang Xu, for his invaluable mentorship throughout my PhD. His deep insights, rigorous academic standards, and continuous support have guided my research direction and strengthened my ability to think critically and independently. I am also grateful for his patience and encouragement, especially during difficult stages of research and writing.

I would like to extend my sincere thanks to my progress evaluation meeting chairs and members, Prof. Sasha Rubin, Prof. Zhanna Sarsenbayeva, and Prof. Simon Poon. Their thoughtful feedback and constructive suggestions have been instrumental in shaping my progress. In particular, their advice on thesis writing, research planning, and time management helped me build clearer milestones and maintain consistent momentum toward completing this thesis.

I am also grateful to my collaborators, Jinjing Zhao, Hongyang Zhang, Kun Yan, and Ping Wang, for their close collaboration and support. Working with them has been a great pleasure. Their technical discussions, shared ideas, and contributions to experiments and paper writing have significantly improved the quality of my research, and our collaboration has made this PhD experience both productive and enjoyable.

I would like to thank The University of Sydney for providing me with the opportunity to pursue my doctoral degree, for offering a supportive research environment, and for providing financial support through the University of Sydney International Stipend Scholarship and the

University of Sydney Tuition Fee Scholarship. I also appreciate the university's academic resources, research facilities, and vibrant community, which enabled me to explore ideas, collaborate broadly, and develop my research skills.

Finally, I am deeply thankful to my family for their unconditional love and support. Their understanding, encouragement, and constant belief in me have been my strongest motivation throughout this journey. This thesis would not have been possible without them.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

Computer vision has become one of the foundational pillars of modern artificial intelligence (AI), enabling machines to perceive, understand, and reason about the visual world. By extracting semantic information from images and videos, computer vision systems support a wide range of applications, including autonomous driving (Yurtsever et al., 2020), robotics (Lynch and Park, 2017), medical diagnosis (Shen et al., 2017), content moderation (Gillespie, 2020), and human–computer interaction (Preece et al., 1994). Over the past decade, advances in deep learning (LeCun et al., 2015), coupled with the availability of large-scale datasets and powerful computing resources, have fundamentally transformed the capabilities of visual recognition models, making computer vision an essential component of intelligent systems.

A central goal of computer vision is to develop algorithms capable of identifying (Zou et al., 2023b), categorizing (Rawat and Wang, 2017), and precisely localizing (Minaee et al., 2021) visual entities. This broad objective is embodied in a family of tasks collectively known as visual recognition and localization. These tasks vary in complexity and granularity, but all share the common goal of robustly interpreting visual data from images or videos.

Image classification, which focuses on predicting a semantic label for an entire image, serves as a foundational task in visual recognition and localization. Pioneering work such as AlexNet (Krizhevsky et al., 2012), VGG (Simonyan and Zisserman, 2014), ResNet (He et al., 2016), and more recent transformer-based architectures like ViT (Dosovitskiy et al., 2020) and DeiT (Touvron et al., 2021a) have significantly advanced recognition performance. *Object detection* extends classification by localizing instances within images. Classical region-based approaches like R-CNN (Girshick et al., 2014), Fast and Faster R-CNN (Girshick, 2015; Ren et al., 2016), and one-stage detectors such as YOLO (Jiang et al., 2022), SSD (Liu

et al., 2016), and RetinaNet (Lin et al., 2017) established strong baselines. More recently, end-to-end transformer models like DETR (Carion et al., 2020) and DINO (Zhang et al., 2022) have redefined the paradigm.

Despite operating on static image data, modern visual recognition and localization systems also pay increasing attention to the video modality, which is substantially more challenging than static images due to temporal dynamics, object motion, appearance changes, occlusions, and long-range dependencies across frames. Video object segmentation (VOS) and video instance segmentation (VIS) are two of the most representative tasks in this setting, as they build upon foundational image recognition and localization techniques while introducing additional mechanisms to model and leverage temporal information. VOS aims to segment a target object throughout a video, often initialized with a mask in the first frame. Representative methods include STM (Oh et al., 2019, 2020), STCN (Cheng et al., 2021c), and AOT (Yang et al., 2021d, 2023e; Yang and Yang, 2022). VIS systems must handle challenges such as occlusions, deformation, and dynamic appearance changes. VIS unifies detection, segmentation, and tracking in videos. Pioneering works such as MaskTrack R-CNN (Yang et al., 2019a), MinVIS (Huang et al., 2022), and more recent transformer-based models like TCOVIS (Li et al., 2023c) and IDOL (Wu et al., 2022) demonstrate the rapid progress of this field.

Research in visual recognition and localization has been propelled by progress across several key directions, including network architecture design, dataset construction, representation learning, domain adaptation, and data-efficient learning. Architectural innovations, ranging from convolutional networks such as ResNet (He et al., 2016) and MobileNet (Howard et al., 2017; Sandler et al., 2018) to transformer-based models such as ViT (Dosovitskiy et al., 2020) and Swin Transformer (Liu et al., 2021b), have continually redefined the capacity of visual systems. Large-scale datasets, including ImageNet (Deng et al., 2009), COCO (Lin et al., 2014), YouTube-VIS (Yang et al., 2019a, 2021b), and OVIS (Qi et al., 2022), have further enabled the development and benchmarking of increasingly sophisticated algorithms. Representation learning, particularly through self-supervised and masked modeling approaches such as MoCo (He et al., 2020), SimCLR (Chen et al., 2020a), and MAE (He et al., 2022),

has significantly reduced reliance on human annotation while producing strong transferable features. Complementing these efforts, domain adaptation and transfer learning methods, ranging from DANN (Ganin et al., 2016) to CLIP (Radford et al., 2021), aim to generalize models across diverse visual environments. Finally, data-efficient learning techniques, including semi-supervised learning methods such as Mean Teacher (Tarvainen and Valpola, 2017) and FixMatch (Sohn et al., 2020a), pseudo-labeling (Arazo et al., 2020), and active learning (Settles, 2009), play an increasingly critical role in reducing annotation costs and enabling scalable training. Together, these research directions form the foundation for advancing high-performance, generalizable, and scalable visual perception systems.

Despite remarkable progress in visual recognition and localization, most state-of-the-art models rely heavily on large-scale labeled datasets, such as ImageNet (Deng et al., 2009) and COCO (Lin et al., 2014), whose construction demands extensive human effort and incurs significant annotation costs. High-quality labels, such as object masks, instance-level correspondences, or frame-by-frame video annotations, are particularly expensive and, in many real-world domains such as medical imaging, robotics, and surveillance, can be difficult or even impractical to obtain at scale. Meanwhile, vast quantities of unlabeled images and videos are readily available on the internet or collected from sensors, representing an enormous reservoir of untapped information. These factors highlight the growing importance of data-efficient learning, which aims to minimize dependence on labeled data while effectively leveraging abundant unlabeled or weakly labeled data. By reducing annotation costs, enabling learning in label-scarce environments, and improving generalization through exposure to more diverse data, data-efficient approaches offer a scalable and practical path toward developing high-performance visual systems.

Data-efficient learning is especially valuable in scenarios where model performance may degrade or where additional supervision would otherwise be required. For example, in medical image analysis, annotated data must be labeled by experts such as radiologists, making large-scale supervision expensive and time-consuming; without sufficient data, diagnostic accuracy can drop significantly. In autonomous driving under rare weather conditions (e.g., heavy snow or fog), models trained primarily on clear-weather data may generalize poorly, and collecting

labeled edge-case data is difficult. In low-resource language processing, many languages lack large annotated corpora, causing performance degradation compared to high-resource languages unless costly annotation efforts are introduced. Similarly, in robot manipulation for novel objects, models trained on limited object categories may struggle to generalize, requiring additional demonstrations or human guidance. In all these cases, data-efficient learning aims to maintain robust performance while minimizing reliance on extensive labeled data or additional supervision.

The central objective of data-efficient learning is to achieve strong performance and robust generalization while relying on as little labeled data as possible and effectively exploiting large amounts of unlabeled or weakly labeled data. Ideally, the techniques involved in data-efficient learning should be universal and transferable, enabling the same methodology to benefit a broad set of visual recognition and localization tasks such as detection, segmentation, video understanding, and multimodal grounding. By maximizing the utility of unlabeled data and minimizing annotation requirements, data-efficient learning seeks not only to reduce labeling costs but also to support learning in domains where manual annotation is scarce, difficult, or impractical, ultimately enabling scalable, adaptable, and more accessible visual AI systems.

In this thesis, we investigate data-efficient learning across a diverse set of vision tasks, including object detection, video object segmentation, and video instance segmentation. Our studies reveal that, despite the distinct problem settings and supervision structures of these tasks, a shared set of foundational techniques, such as leveraging unlabeled data, representation learning, and efficient model design, can be effectively transferred and adapted to meet each task's unique requirements. Beyond these traditional recognition and localization tasks, we also explore the increasingly important problem of referring expression comprehension, where the goal is to identify a specific entity in an image based on a natural-language description. This task is becoming particularly critical in the era of large language models, in which language serves as a universal interface for interacting with multimodal systems. Together, these investigations highlight both the universality and adaptability of data-efficient learning principles in modern computer vision. This thesis is organized as follows:

Chapter 2 presents a comprehensive literature review of the relevant research areas, covering mainstream visual recognition and localization tasks, including object detection, video object segmentation, video instance segmentation, and referring expression comprehension, as well as key methodologies in data-efficient learning.

Chapter 3 investigates how to effectively enhance visual recognition and localization for static images under limited training data, with a focus on architecture design. Specifically, we validate our approach on object detection, one of the most representative tasks in static-image visual recognition and localization. The core idea is to combine the complementary strengths of two mainstream detector paradigms: classical two-stage detectors and DETR-style detectors. We introduce the Hybrid Proposal Refiner (HPR) (Zhao et al., 2024) and detail its design by gradually transforming a Faster R-CNN architecture into a Deformable DETR framework, highlighting several key insights uncovered along the way. Extensive experiments show that HPR can be seamlessly applied to a wide range of DETR-style detectors, consistently boosting performance and improving data utilization efficiency.

Chapter 4 focuses on visual recognition and localization in videos, which introduce additional challenges absent in static-image settings, such as the need to model long-range temporal dependencies across frames. This chapter emphasizes learning under extremely limited supervision while effectively leveraging large amounts of unlabeled video data. Specifically, we study the problem of low-shot video object segmentation (VOS), where each training video contains only one or two annotated frames. We formulate low-shot VOS as an extreme semi-supervised setting and, based on this perspective, propose a simple yet effective two-phase training paradigm (Yan et al., 2025) that fully exploits the information contained in unlabeled frames. Our approach is model-agnostic and generalizes well across diverse VOS architectures (STCN, RDE-VOS, XMem) and multiple datasets (DAVIS 2016/2017, YouTube-VOS 2018/2019, LVOS, and VOST).

Chapter 5 introduces a novel data-efficient visual recognition and localization methodology, MinMaxVIS (Wei et al., 2025), for video instance segmentation (VIS). MinMaxVIS addresses video understanding by training on a small amount of labeled static images together with a large collection of unlabeled images, in contrast to prior VIS approaches that typically (1)

rely on video data for training and (2) require dense, video-level annotations. Extensive experiments on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS demonstrate that MinMaxVIS not only achieves substantial improvements over existing image-driven baselines but also outperforms the fully supervised MinVIS, while using only 1–10% of the labeled data. This chapter demonstrates that high-quality VIS can be achieved without relying on dense video annotations.

Chapter 6 further investigates referring expression comprehension (REC), a task that requires modern visual recognition and localization systems to take natural language expressions as input in order to recognize and localize visual entities in images. Unlike traditional recognition and localization tasks that operate on a predefined closed set of categories, REC inherently works in an open-set setting, as natural language serves as a flexible and expressive interface for specifying target entities. We introduce HC-RefLoCo (Wei et al., 2024), a large-scale human-centric benchmark designed to advance referring expression comprehension in the era of large multimodal models. HC-RefLoCo contains 44,738 high-quality referring expressions for 24,129 human instances across 13,452 images. Extensive analyses show that HC-RefLoCo provides significantly richer linguistic diversity, broader image and instance-size distributions, and more uniform spatial coverage compared with existing REC benchmarks. We further introduce comprehensive evaluation protocols, including accuracy at multiple IoU thresholds, scale-aware analysis, and subject-specific assessment, and use them to benchmark 24 state-of-the-art models. These evaluations reveal several key insights that we hope will facilitate future research in referring expression comprehension.

Chapter 7 concludes the thesis by summarizing our key contributions and outlining promising directions for future research.

The task, supervision, core method, and headline results of each chapter are summarized in Table 1.1. Additionally, before introducing the main techniques in each chapter, we emphasize that while many tasks share common underlying principles, each task also requires task-specific innovations to achieve optimal performance.

Data-efficient visual recognition and localization can be generalized across multiple tasks, including object detection, video instance segmentation, video object segmentation, and referring expression comprehension. Although these tasks differ in formulation and output space, they share several fundamental principles. For example, they often rely on similar backbone architectures for visual feature extraction, emphasize efficient use of limited annotated data through strategies such as data augmentation, pretraining, or semi-supervised learning, and explore leveraging unlabeled data via self-supervised learning or pseudo-labeling. Moreover, principles such as reducing annotation redundancy, designing parameter-efficient models, and improving representation quality are broadly applicable across tasks.

At the same time, each task requires task-specific innovations. For instance, video-based tasks introduce additional challenges beyond image-based settings, such as modeling long-range temporal dependencies, ensuring cross-frame consistency, and handling object motion and occlusion. For example, in Chapter 4, we propose bidirectional inference to learn mask association across frames, which is unnecessary in image-based perception models (Chapter 3). Moreover, video settings require mechanisms such as intermediate inference to better handle object motion over time. Video instance segmentation (Chapter 5) further requires both accurate tracking and category prediction, in contrast to video object segmentation (Chapter 4). Our setting is even more challenging, as the model is trained on image data rather than video data. This requires the model to learn instance association from static images without explicit temporal supervision. To address this, we propose a contrastive learning paradigm in Chapter 5, which is not necessary for the method in Chapter 4, where training is conducted directly on video data. Therefore, while data-efficient learning is guided by shared methodological principles, it must be adapted with task-specific designs to address the unique challenges of each problem setting.

Chapter	Task	Supervision	Brief Method Description	Result
Chapter-3	Object detection	COCO dataset (18,287 images).	We combine the complementary strengths of classical two-stage and DETR-style detector paradigms to improve data efficiency under limited training data.	54.9 AP with a ResNet-50 backbone on COCO.
Chapter-4	Video object segmentation	DAVIS 2016/2017, YouTube-VOS 2018/2019, and LVOS with 1% and 2% labeled data.	We train a low-shot video object segmentation model in which each training video contains only one or two annotated frames.	83.6/83.5 $\mathcal{G}$ on YouTube-VOS 2018/2019, 90.5/91.4 $\mathcal{J}\&\mathcal{F}$ on DAVIS 2016/2017, 48.2 $\mathcal{J}\&\mathcal{F}$ on LVOS.
Chapter-5	Video instance segmentation	1% and 2% labeled data of YouTube-VIS 2019/2021, and 5% and 10% labeled data of OVIS.	We train a video instance segmentation model using a small amount of labeled static images together with a large collection of unlabeled images.	60.9/62.2 on YouTube-VIS 2019-1%/2%, 54.1/55.6 on YouTube-VIS 2021-1%/2%, and 37.5/39.2 on OVIS-5%/10%.
Chapter-6	Referring expression comprehension	Different models are trained on different datasets.	We introduce a challenging benchmark including 44,738 high-quality referring expressions across 13,452 images for modern multi-modal models.	Evaluation is conducted on 24 state-of-the-art models across seven evaluation metrics.

TABLE 1.1. Summary of each chapter.

CHAPTER 2

Literature Review

In this chapter, we survey the literature on data-efficient visual perception and summarize the key techniques that will be introduced in the subsequent chapters.

2.1 Object Detection

2.1.1 Single-Stage Detectors

Single-stage object detection approaches have become increasingly popular due to their architectural simplicity, faster inference speed, and suitability for real-time applications. Among them, the YOLO family (Redmon et al., 2016a; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020; Li et al., 2022a; Wang et al., 2023a) stands as a foundational milestone. YOLO pioneered the idea of directly predicting bounding boxes and class labels from dense grid cells in a single forward pass, eliminating the need for region proposal generation and refinement used in two-stage detectors. This design dramatically improved inference efficiency and paved the way for real-time object detection systems deployed in practical applications.

Following YOLO, SSD (Liu et al., 2016) extended the single-stage paradigm by introducing multi-scale feature extraction through a hierarchy of convolutional layers, enabling more accurate localization of objects with large scale variations. Despite their computational advantages, early single-stage detectors often lagged behind two-stage and multi-stage counterparts in accuracy due to challenges such as severe foreground–background imbalance and limited representational capacity.

To close this performance gap, RetinaNet (Lin et al., 2017) introduced the influential Focal Loss, which down-weights easy negative samples while focusing training on hard, informative examples. This innovation substantially improved single-stage detection accuracy, establishing RetinaNet as a strong competitor to two-stage detectors. Building on this momentum, researchers explored simplifying object detection further through anchor-free models (Huang et al., 2015; Kong et al., 2020; Law and Deng, 2018; Tian et al., 2019b; Wei et al., 2020; Zhou et al., 2019). These methods dispense with predefined anchor boxes and instead learn to localize objects directly from points, corners, or centers. Anchor-free detectors reduce design complexity, eliminate extensive anchor hyperparameter tuning, and offer improved adaptability across datasets.

Subsequently, ATSS (Zhang et al., 2020c) provided a unified perspective by examining the inconsistency between anchor-based and anchor-free methods and introducing an adaptive training sample selection mechanism. By dynamically determining positive and negative samples based on statistical characteristics, ATSS effectively bridges the two paradigms and improves training stability and overall detection accuracy. Together, these advancements illustrate the rapid evolution of single-stage detectors toward architectures that are not only efficient but also competitive with, and sometimes surpassing, their multi-stage counterparts.

2.1.2 R-CNN Series

The R-CNN family of detectors has played a foundational role in shaping the landscape of modern object detection. R-CNN (Girshick et al., 2014) first introduced the two-stage detection paradigm, in which region proposals are generated using external algorithms and subsequently classified and refined by a convolutional network. Fast R-CNN (Girshick, 2015) streamlined this pipeline by enabling end-to-end training with shared feature extraction, dramatically improving efficiency and laying a solid foundation for future advancements.

Building upon these ideas, Faster R-CNN (Ren et al., 2015) introduced the Region Proposal Network (RPN), which generates high-quality proposals directly from feature maps, enabling a fully end-to-end trainable two-stage detector. This innovation not only improved accuracy

and speed but also established a unified architecture that remains influential in contemporary research.

Following Faster R-CNN, a wave of architectural enhancements (Cai and Vasconcelos, 2018; Sun et al., 2021; Wei et al., 2021; Lu et al., 2019; Zhang et al., 2020b; Du et al., 2022; Yang et al., 2022b) further strengthened the two-stage detection family. Cascade R-CNN (Cai and Vasconcelos, 2018) introduced a multi-stage cascade of progressively stricter classifiers and regressors, effectively improving localization quality and mitigating detector overfitting at high IoU thresholds. Grid R-CNN (Lu et al., 2019) refined bounding box localization by predicting structured grids, while Dynamic R-CNN (Zhang et al., 2020b) adaptively adjusted IoU thresholds during training to improve the alignment between classification and localization.

More recently, Sparse R-CNN (Sun et al., 2021) fundamentally rethought two-stage detection by replacing dense region proposals with a fixed set of learnable object queries. This design dramatically reduces computational overhead and avoids the complexities associated with anchor boxes and proposal generation. Additional innovations such as AlignDet (Wei et al., 2021), Learning to Align Proposals (Du et al., 2022), and factorized interaction models (Yang et al., 2022b) continue to refine how proposals are generated, aligned, and interacted with features.

Collectively, the R-CNN series not only established the canonical two-stage framework but also inspired a long line of architectural innovations that pushed detection accuracy, efficiency, and robustness to new levels. These methods continue to serve as strong baselines and conceptual pillars in both academic research and industrial applications.

2.1.3 DETR Series

DETR (Carion et al., 2020) has emerged as a transformative approach in object detection, introducing a fully end-to-end paradigm built upon the Transformer architecture (Vaswani et al., 2017b) and bipartite matching via the Hungarian algorithm (Kuhn, 1955). By discarding many hand-crafted components, such as anchor design, proposal generation, and

Non-Maximum Suppression (NMS), DETR establishes a remarkably simple yet conceptually elegant framework. Its formulation casts object detection as a direct set prediction problem and inspires a rich line of follow-up research aimed at improving convergence speed, accuracy, and architectural flexibility (Meng et al., 2021a; Liu et al., 2022; Wang et al., 2022b; Li et al., 2022b; Lin et al., 2023a; Wang et al., 2021b).

A key milestone in the DETR family is Deformable DETR (Zhu et al., 2020), which integrates multi-scale feature representations and introduces deformable attention. Instead of attending densely to all spatial locations, the deformable attention module focuses on a sparse set of dynamically predicted key points around reference anchors, greatly enhancing computational efficiency. This modification not only accelerates training by an order of magnitude but also substantially improves performance on small objects, which was one of the main deficiencies of the original DETR.

Building on this progress, a variety of advanced designs have further expanded the DETR ecosystem (Cai et al., 2023; Jia et al., 2023; Yao et al., 2021; Chen et al., 2022, 2023d; Zong et al., 2023; Zhang et al., 2023d). For example, DINO (Zhang et al., 2022) introduces an effective denoising training strategy and contrastive query learning to strengthen the stability and discriminative power of decoder queries. Hybrid matching strategies, as adopted in $\mathcal{H}$-DETR (Jia et al., 2023) and Group DETR (Chen et al., 2023d), combine one-to-one matching with auxiliary one-to-many assignments to enrich positive supervision and improve optimization. $\mathcal{C}$o-DETR (Zong et al., 2023) further enhances this direction through collaborative hybrid assignment mechanisms that better balance classification and localization learning.

Other works provide deeper insights into the behavior of DETR queries. DDQ (Zhang et al., 2023d) highlights the importance of queries being both dense enough to cover potential objects and unique enough to avoid redundancy under the one-to-one assignment constraint. Align DETR (Cai et al., 2023) improves localization quality by introducing a localization-precision-aware classification loss and a prime sample weighting mechanism to suppress noisy or misleading samples during training. Moreover, efficient variants (Yao et al., 2021; Chen

et al., 2022) explore architectural refinements that reduce computation while maintaining or improving accuracy.

Collectively, these developments establish the DETR series as a vibrant and evolving research direction, reshaping object detection with principled set prediction, streamlined pipelines, and increasingly powerful Transformer-based architectures.

2.2 Video Object Segmentation

2.2.1 Architecture

Existing Video Object Segmentation (VOS) methods can be broadly categorized into online-learning-based and offline-learning-based approaches. Online learning methods (Caelles et al., 2017; Luiten et al., 2018; Perazzi et al., 2017; Voigtlaender and Leibe, 2017; Xiao et al., 2018; Maninis et al., 2018; Cheng et al., 2017) rely on fine-tuning the segmentation model at test time using the ground-truth mask from the first frame. This process customizes the network to the specific appearance of the target object, often leading to strong segmentation accuracy—especially in challenging scenarios with substantial appearance variations. However, test-time fine-tuning introduces significant computational overhead and latency, making these methods impractical for real-time applications or large-scale deployment. Moreover, their performance can be sensitive to fine-tuning hyperparameters, reducing robustness across datasets.

In contrast, offline-learning-based methods (Mao et al., 2021; Yang et al., 2021d; Hu et al., 2021c; Zhang et al., 2020d; Ge et al., 2021; Lu et al., 2020) aim to segment videos directly at inference time without any online adaptation. These methods learn generic representations during training and operate in a fully feed-forward manner at test time. Offline VOS techniques generally fall into two categories: propagation-based or matching-based. Propagation-based approaches (Chen et al., 2020b; Li and Loy, 2018; Oh et al., 2018; Johnander et al., 2019) segment each frame by relying on the predicted mask of the previous frame. While this

allows efficient temporal forwarding, it often suffers from error accumulation, where small inaccuracies grow progressively throughout the video.

Matching-based approaches (Cheng et al., 2021c; Yang et al., 2020; Oh et al., 2019; Wang et al., 2021a) mitigate this issue by maintaining a memory bank that stores key-value features from past frames. During inference, the model retrieves and matches these features to assist in segmenting the current frame. This design enables long-term temporal reasoning and robust performance under occlusion, fast motion, or appearance changes.

A representative breakthrough in this category is STM (Oh et al., 2019, 2020), which introduced a spatiotemporal memory network that stores both image features and corresponding masks from previous frames. STM's memory-based matching strategy significantly advanced offline VOS performance and inspired a series of subsequent improvements. Kernelized attention and hierarchical memory mechanisms (Seong et al., 2020, 2021) enhance memory retrieval efficiency. XMem (Cheng and Schwing, 2022) further improves performance by incorporating a multi-scale, long-range feature memory, setting new state-of-the-art results.

Several models refine efficiency and scalability. STCN (Cheng et al., 2021c) reduces redundancy by avoiding repeated encoding of object-specific mask features, enabling faster inference without sacrificing accuracy. RDE-VOS (Li et al., 2022c) introduces a recurrent dynamic embedding mechanism that maintains a fixed-size memory bank while preserving strong temporal consistency. The AOT family (Yang et al., 2021d, 2023e; Yang and Yang, 2022) is designed to handle multiple objects simultaneously in a single inference pass, greatly improving scalability for multi-object scenarios.

Together, these advances highlight the evolution from computationally intensive online learning toward highly efficient and robust offline architectures. The progression from simple mask propagation to sophisticated memory-matching frameworks underscores the importance of long-term temporal modeling, memory design, and scalable representation learning in modern VOS research.

2.2.2 Segment Anything Model

Recent breakthroughs in foundation segmentation models (Kirillov et al., 2023c; Zou et al., 2023a; Wang et al., 2023d) have ushered in a new era for image and video segmentation, enabling unprecedented levels of generalization and flexibility. Among them, the Segment Anything Model (SAM) (Kirillov et al., 2023c) stands out as a landmark contribution. Trained on billions of masks, SAM exhibits remarkable zero-shot segmentation capabilities and supports a wide range of prompt types, including points, bounding boxes, and coarse masks, making it highly adaptable across diverse domains and tasks.

As SAM gained traction, researchers began exploring its integration into Video Object Segmentation (VOS) and other specialized segmentation tasks. These efforts generally fall into two categories: augmenting SAM with learnable components and building SAM-driven task-specific pipelines. Methods such as HQ-SAM (Ke et al., 2023) introduce a trainable high-quality token to enhance mask precision, while PerSAM (Zhang et al., 2023c) adapts SAM for personalized segmentation using one-shot learning to capture object-specific appearance cues.

Other approaches develop customized frameworks that incorporate SAM as a core segmentation module. SAM-Track (Cheng et al., 2023b) employs SAM interactively to extract high-quality masks from key frames before propagating them through the video. The Tracking Anything Model (TAM) (Yang et al., 2023a) integrates SAM with XMem (Cheng and Schwing, 2022): SAM is used interactively to initiate segmentation in reference frames, whereas XMem handles temporal propagation for subsequent frames. Several other works (Cheng et al., 2023b; Zhu et al., 2023b; Cheng et al., 2023a; Zhang et al., 2023g) expand SAM for video tracking, unsupervised VOS, and long-term segmentation through architectural adaptations or task-specific training.

2.3 Instance Segmentation

2.3.1 Image Instance Segmentation

Image instance segmentation, which aims to predict both the category and the pixel-level mask of each object in an image, has advanced rapidly alongside progress in object detection (Ren et al., 2016; Tian et al., 2019a; Redmon et al., 2016b; Carion et al., 2020). A major milestone in this field was Mask R-CNN (He et al., 2017), which extended Faster R-CNN (Ren et al., 2016) by introducing a parallel mask prediction branch. This simple yet powerful design established a strong baseline for many subsequent approaches. Building on this foundation, later works improved the quality of bounding box detection (Cai and Vasconcelos, 2019; Chen et al., 2019; Liu et al., 2018) and enhanced mask precision through refined feature aggregation, boundary-aware learning, and point-based prediction (Cheng et al., 2020; Kirillov et al., 2020; Tang et al., 2021; Huang et al., 2019; Zhang et al., 2021).

To overcome the limitations of two-stage pipelines, particularly the reliance on region-of-interest (RoI) operations, a set of one-stage instance segmentation methods emerged (Bolya et al., 2019; Cheng et al., 2022b; Xie et al., 2020a), often built upon fast single-shot detectors (Tian et al., 2019a; Redmon et al., 2016b). YOLACT (Bolya et al., 2019), for example, generates a set of global prototype masks and instance-specific coefficients to efficiently assemble instance masks. Meanwhile, SOLO (Wang et al., 2020a) and SOLOv2 (Wang et al., 2020b) reframe instance segmentation as a pure segmentation task by directly predicting object-specific regions based on spatial position and object center cues, completely bypassing bounding box prediction.

A more recent trend is the rise of query-based instance segmentation methods (Hu et al., 2021b; Cheng et al., 2022a; Dong et al., 2021; He et al., 2023; Li et al., 2023b; Zhang et al., 2024b; Zhao et al., 2024), inspired by DETR (Carion et al., 2020), the first end-to-end Transformer-based object detector. These methods formulate instance segmentation as a set prediction problem and use learnable queries to simultaneously reason about objects and their

masks. Mask2Former (Cheng et al., 2022a) introduces masked attention, restricting cross-attention to predicted foreground regions, thereby improving spatial precision. MaskDINO (Li et al., 2023b) extends DINO (Zhang et al., 2022) by jointly predicting object classes, bounding boxes, and masks within a unified Transformer framework. Other variants further explore efficient query representations, language-guided instance segmentation, and hybrid matching strategies to achieve stronger performance.

Despite their success in static images, most of these instance segmentation methods do not natively handle temporal consistency, object re-identification, or long-term tracking, which are essential capabilities for video instance segmentation (VIS). Without mechanisms to associate objects across frames, these image-based approaches cannot be directly extended to VIS, motivating the development of specialized models that incorporate temporal cues, memory structures, and inter-frame matching strategies.

Overall, the evolution of image instance segmentation, from two-stage region-based frameworks to one-stage architectures and fully end-to-end query-based models, has laid a strong foundation for more advanced tasks. However, bridging the gap between image-based segmentation and temporally consistent video instance segmentation remains an open and active research frontier.

2.3.2 Video Instance Segmentation

Video Instance Segmentation (VIS) was first introduced in (Yang et al., 2019a) as a unified task that requires detecting, segmenting, and tracking object instances across all frames of a video. Unlike image instance segmentation, VIS introduces additional challenges such as object motion, long-term occlusion, appearance variations, and persistent identity assignment across frames. Current VIS methods can be broadly categorized into offline and online approaches (Athar et al., 2020; Cheng et al., 2021a; Lin et al., 2021; Heo et al., 2022; Hwang et al., 2021; Wu et al., 2021; Wang et al., 2021c; Yang et al., 2022a, 2019a; Cao et al., 2020; Liu et al., 2021a; Wu et al., 2022; Yang et al., 2021c; Heo et al., 2023; Huang et al., 2022; Li et al., 2023c; Kim et al., 2024).

Offline VIS methods operate in a video-in, video-out manner: they process the entire video (or a long clip) in a single pass and simultaneously predict segmentation masks and instance identities. This allows the models to capture global temporal structure but makes them unsuitable for real-time or streaming applications.

VisTR (Wang et al., 2021c) pioneered the use of Transformers (Vaswani et al., 2017a) for video instance segmentation by jointly modeling spatial and temporal dependencies across frames. SeqFormer (Wu et al., 2021) improves this design by dynamically generating a per-frame mask sequence for each instance and aggregating temporal context into robust video-level instance features. VITA (Heo et al., 2022) further advances offline video instance segmentation by associating frame-level object tokens into coherent instance trajectories without requiring explicit spatio-temporal backbones.

Although these offline models achieve strong accuracy by leveraging global context, their high memory consumption and full-video processing design make them impractical for latency-sensitive or long-duration videos.

Online VIS methods process incoming frames sequentially, making them naturally suited for streaming, real-time, and long-horizon settings. The task is decomposed into predicting masks for the current frame and associating them with previously tracked instances.

MaskTrack R-CNN (Yang et al., 2019a), an extension of Mask R-CNN (He et al., 2017), is the seminal online VIS baseline, introducing a tracking branch to associate object instances across frames. Subsequent works (Cao et al., 2020; Liu et al., 2021a; Yang et al., 2021c) improved frame-level segmentation accuracy and identity matching strategies using attention mechanisms, feature propagation, and cross-frame interaction.

IDOL (Wu et al., 2022) boosts association robustness by learning discriminative instance embeddings through contrastive learning, built on top of Deformable DETR (Zhu et al., 2020). TCOVIS (Li et al., 2023c) improves temporal consistency by combining global instance assignment with spatio-temporal enhancement modules. VISAGE (Kim et al., 2024) strengthens object distinction by incorporating stronger appearance cues during tracking. The DVIS family (Zhang et al., 2023e,f; Zhou et al., 2024) proposes a versatile segmentation

framework that flexibly supports both online and offline regimes, making it adaptable to different application needs.

2.4 Referring Expression Comprehension

2.4.1 Benchmarks

Referring expression comprehension (REC) aims to localize a specific object instance in an image based on a natural language description. As a multimodal task bridging vision and language, REC requires models to jointly interpret linguistic semantics, visual cues, spatial relationships, and fine-grained appearance details. Existing human-centric REC benchmarks are largely derived from general-purpose REC datasets such as RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016), all of which originate from the COCO2014 images (Lin et al., 2014).

RefCOCO includes approximately 50,000 referring expressions across 19,994 images and is characterized by short, succinct descriptions often relying on spatial cues, such as “Right guy,” “Far left man,” or “Guy on left.” RefCOCO+, with 49,856 expressions over a similar image set, intentionally removes explicit locational terms (e.g., “left,” “right”) to emphasize appearance-based reasoning. Expressions like “Man with light hat” or “Guy in white” require the model to rely more heavily on visual attributes and object properties. RefCOCOg provides significantly richer annotations, typically longer and more descriptive. Examples such as “A person in a hat on a wooden bench” or “A man in white playing Frisbee” demonstrate its focus on detailed contextual and relational reasoning.

To enhance referring expression comprehension model performance on these benchmarks, large-scale multimodal datasets such as GRIT (Peng et al., 2023), GranD (Rasheed et al., 2023), and RecapD (Guo et al., 2024) are often used for pre-training or auxiliary supervision. In addition, datasets not originally designed for referring expression comprehension, such as Flickr30k Entities (Plummer et al., 2017; Young et al., 2014) and Visual Genome (Krishna

et al., 2017), are widely adopted due to their rich region-level descriptions and dense annotations.

2.4.2 LMMs for Visual Grounding

Recent advancements in large multimodal models (LMMs), including Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023d), MiniGPT-4 (Zhu et al., 2023a; Chen et al., 2023a), InstructBLIP (Chen et al., 2023a), mPLUG-Owl (Ye et al., 2023), and LLaVA (Liu et al., 2023a), have dramatically strengthened the integration of visual and linguistic modalities. By capitalizing on the rapid progress of large language models (LLMs) such as GPT-4 (OpenAI, 2023a,b), Gemini (Team et al., 2023), and the LLaMA family (Touvron et al., 2023a,b; Chiang et al., 2023), these systems exhibit powerful image understanding and visual reasoning abilities, often achieving near-human performance in tasks such as image captioning, visual question answering, and multimodal dialogue.

Despite these impressive gains, instance-level localization remains a substantial challenge for LMMs. Unlike global understanding tasks, localization requires the model to precisely map linguistic expressions to specific visual regions and generate accurate bounding boxes or segmentation masks. This demands not only semantic alignment between modalities but also fine-grained spatial reasoning, object disambiguation, and grounding of complex descriptions, which are capabilities that current LMMs only partially possess. As a result, referring expression comprehension (REC) plays a vital role as a diagnostic benchmark for evaluating an LMM's grounding and localization abilities.

To address REC, many pioneering LMMs, such as KOSMOS-2 (Peng et al., 2023), Shikra (Chen et al., 2023b), GroundingGPT (Li et al., 2024), Qwen-VL (Bai et al., 2023), and the SPHINX series (Lin et al., 2023b; Gao et al., 2024), adopt auto-regressive causal Transformers that output tokenized bounding box coordinates. By treating localization as a language generation problem, these models seamlessly integrate grounding into the LLM pipeline. However, bounding boxes offer limited precision, especially for fine-grained or non-rectangular objects.

Motivated by the success of segmentation foundation models like SAM (Kirillov et al., 2023a), recent approaches advocate for pixel-level mask predictions to achieve more accurate localization. Models such as LISA (Lai et al., 2023; Yang et al., 2023b), PixelLLM (Zhongwei Ren, 2023), PSALM (Zhang et al., 2024a), and GlaMM (Rasheed et al., 2023) extend the LMM architecture to generate segmentation masks guided by natural language. These mask-based approaches capture spatial detail more effectively than bounding boxes, making them promising directions for high-resolution grounding.

A closely related field is open-vocabulary object detection and segmentation (Gu et al., 2021; Du et al., 2022; Xu et al., 2022a, 2023), in which models detect and classify arbitrary objects using free-form category names. Although sharing conceptual similarities with REC, the two tasks differ fundamentally: open-vocabulary detection relies on short labels or simple phrases, whereas REC requires grounding based on long, descriptive, context-rich expressions that may include relationships, attributes, actions, and fine-grained distinctions. Thus, REC poses a substantially more demanding challenge, requiring deeper language comprehension and more precise visual grounding.

In summary, while LMMs have rapidly advanced multimodal understanding, REC remains a crucial benchmark for measuring the fine-grained localization and grounding capabilities essential for reliable real-world multimodal intelligence.

2.5 Data-Efficient Learning

2.5.1 Semi-Supervised Learning

Semi-supervised learning leverages a small amount of labeled data together with a large pool of unlabeled samples to significantly boost model performance. This paradigm has proven highly effective across a wide range of computer vision tasks, including image classification (Sohn et al., 2020a; Tarvainen and Valpola, 2017), semantic segmentation (Hu et al., 2021a; Ke et al., 2020), object detection (Sohn et al., 2020b; Xu et al., 2021), and action recognition (Xu et al., 2022b). By reducing reliance on costly annotations and fully

utilizing accessible unlabeled data, semi-supervised learning has become a crucial technique for data-efficient model training.

Most existing semi-supervised methods fall into two major categories: consistency-based approaches and pseudo-label-based approaches. Consistency-based methods (Laine and Aila, 2016; Tarvainen and Valpola, 2017; Berthelot et al., 2019b; Sajjadi et al., 2016; French et al., 2019; Chen et al., 2021b) encourage the model to produce stable predictions under a variety of perturbations. These perturbations may include model perturbations such as parameter noise (Bachman et al., 2014), input perturbations through data augmentation (Xie et al., 2020b; Berthelot et al., 2019a), or adversarial perturbations (Miyato et al., 2018). The underlying idea is that a well-regularized model should be invariant to these variations, thereby improving generalization.

On the other hand, pseudo-labeling methods (Lee et al., 2013; Sohn et al., 2020b; Zoph et al., 2020; Xie et al., 2020c; Grandvalet and Bengio, 2004) generate hard one-hot labels for unlabeled samples based on the model's confident predictions. These pseudo labels are then treated as ground truth during training, allowing the model to iteratively refine itself. Pseudo-labeling has proven particularly effective when high-confidence predictions correlate strongly with true labels, enabling efficient learning even in low-annotation regimes.

2.5.2 Self-Supervised Learning

Unsupervised feature learning has played a foundational role in the development of deep neural networks. Early approaches such as auto-encoders (Vincent et al., 2010) and Deep Boltzmann Machines (Salakhutdinov and Hinton, 2009) learned hierarchical representations by reconstructing input pixels from compressed latent features. These models were often used to initialize deep architectures, followed by supervised fine-tuning on downstream tasks such as handwritten digit classification. Although effective for their time, these early unsupervised methods were typically limited by the simplicity of their reconstruction objectives and the small scale of available datasets.

More recently, self-supervised learning (SSL) has emerged as a powerful paradigm for unsupervised representation learning. Modern SSL formulations are designed primarily for transfer learning, where pretraining and downstream tasks are performed on different datasets with distinct objectives. A wide range of pretext tasks have been developed to encourage models to learn high-level semantic structure: colorization (Larsson et al., 2017) teaches models to infer object semantics from grayscale images; context prediction (Doersch et al., 2015) and inpainting (Pathak et al., 2016) require understanding spatial relationships; and rotation prediction (Gidaris et al., 2018) forces the model to learn object orientation and appearance cues.

The most influential line of work in SSL is contrastive learning, especially the instance discrimination paradigm (Wu et al., 2018), where augmented views of the same image are encouraged to map to nearby embeddings while different images are pushed apart. Contrastive methods such as MoCo (He et al., 2020), SimCLR (Chen et al., 2020a,c), BYOL (Grill et al., 2020), Barlow Twins (Zbontar et al., 2021), and SwAV (Caron et al., 2021) have achieved exceptional transfer performance on downstream tasks, in some cases surpassing fully supervised ImageNet pretraining (He et al., 2020). These breakthroughs highlight the ability of SSL to learn rich, high-level semantics without labels.

With further innovations, such as clustering-based learning of assignments in SwAV (Caron et al., 2020), large-scale contrastive learning has been successfully applied to massive, uncurated image corpora (Caron et al., 2019; Goyal et al., 2021). While linear evaluation accuracy on ImageNet has steadily improved, early SSL methods often struggled to match supervised features on dense prediction tasks such as semantic segmentation or object detection. These tasks require spatially or regionally aligned representations rather than solely image-level embeddings.

To bridge this gap, many recent works explore pixel-level or region-level self-supervised pretraining. Methods such as VADer (Pinheiro et al., 2020), PixPro (Xie et al., 2020d), and DenseCL (Wang et al., 2020c) learn dense representations by enforcing consistency between pixel features that correspond to the same physical point across augmentations. InsLoc (Yang et al., 2021a) extends this idea to region-level features using composite imagery, enabling

more meaningful instance-level representations. DetCon (Hénaff et al., 2021) improves dense contrastive learning by leveraging hierarchical segmentation masks from MCG (Arbeláez et al., 2014) to enforce intra-segment consistency and inter-segment discrimination.

CHAPTER 3

Data-Efficient Learning through Network Architecture Design

In this chapter, we investigate how to effectively enhance visual recognition and localization for static images under limited training data, with a particular focus on architecture design. Specifically, we validate our approach on object detection, one of the most representative tasks in static-image visual recognition and localization. Section 3.1 presents the problem formulation of object detection. Section 3.2 introduces the motivation and analyzes the complementary strengths of two major detector families, namely the Faster R-CNN series and the DETR series, showing that integrating their advantages can lead to more effective data utilization. Based on this analysis, we propose the Hybrid Proposal Refiner (HPR) (Zhao et al., 2024), and detail the proposed method in Section 3.3. Section 3.4 reports extensive experimental results, and Section 3.5 concludes the chapter with a summary.

3.1 Problem Formulation

Object detection is a fundamental task in computer vision that aims to identify what objects are present in an image (or video frame) and where they are located. Concretely, a detector predicts a set of object instances, each typically represented by a bounding box (e.g. (x, y, w, h)) and a category label (e.g., person, car, dog). Compared with image classification which outputs a single label for the whole image, object detection provides a structured, instance-level understanding of the scene, making it a key building block for higher-level perception tasks.

For example, in a street scene, an object detector can locate multiple objects simultaneously, such as cars, pedestrians, traffic lights, bicycles, and road signs. In an indoor environment, it may detect objects like cups, keyboards, books, and monitors on a desk. In each case, the

output is not only the object type but also its spatial extent, enabling downstream systems to reason about object relationships, prioritize attention, and make decisions based on the scene layout.

Object detection underpins a wide range of real-world applications, including:

(1) Autonomous driving and intelligent transportation: detecting vehicles, pedestrians, traffic signs/lights, and obstacles for safe planning and navigation.
(2) Robotics and embodied AI: enabling robots to locate and interact with target objects for grasping, manipulation, navigation, and task execution.
(3) Human–computer interaction and AR/VR: detecting hands, faces, or everyday objects to support interaction, overlay, and scene-aware rendering.
(4) Retail and logistics: inventory checking, package detection, defect detection, and automated checkout.
(5) Medical imaging: locating lesions, tumors, organs, or abnormalities to assist diagnosis and treatment planning.

Overall, object detection provides a compact yet expressive representation of the visual world, and its performance directly affects many downstream tasks such as instance segmentation, tracking, and scene understanding.

3.2 Motivation

The attention-based Transformer architecture introduced by Vaswani et al. (Vaswani et al., 2017b) has become a general-purpose backbone for sequence modeling. After redefining the state of the art in natural language processing, it was soon adapted to visual tasks, leading to strong results in image classification (Dosovitskiy et al., 2020; Chen et al., 2021a; Liu et al., 2021b; Touvron et al., 2021b; d'Ascoli et al., 2021; Han et al., 2021) and, more recently, object detection (Carion et al., 2020; Liu et al., 2022; Meng et al., 2021a; Zhu et al., 2020; Zhang et al., 2022, 2023d; Lin et al., 2023a; Jia et al., 2023; Zong et al., 2023; Wang et al., 2022b). A milestone in this transition is DETR (Carion et al., 2020), which reformulates detection as a set

Model	AP
Faster R-CNN (ResNet-50, FPN, 12-epoch)	36.5
+ Class-Agnostic RPN→Class-Aware RPN	36.1 (-0.4)
+ FPN→Deformable Encoder	44.0 (+7.9)
+ IoU Matching→Hungarian Matching (RPN)	32.7 (-11.3)
+ IoU Matching→Hungarian Matching (R-CNN)	32.2 (-0.5)
+ RoI Feature→Object Feature	41.2 (+9.0)
+ Object Feat.→Object Feat. + RoI Feat.	41.7 (+0.5)
+ Object Feat. + RoI Feat.→Deformable Attention	44.2 (+2.5)
+ 6× Deformable Attention	46.2 (+2.0)

TABLE 3.1. Step by step, we transform the Faster R-CNN (Ren et al., 2015) into the Deformable DETR (Zhu et al., 2020). We report AP on COCO benchmark. Object feature denotes RPN's point feature extracted by the neck network.

prediction problem using an encoder–decoder Transformer. Instead of relying on hand-crafted anchors and post-processing heuristics, DETR introduces learnable query embeddings that attend to CNN feature maps to produce object categories and bounding boxes in a unified pipeline. Its end-to-end property is enabled by one-to-one bipartite assignment via Hungarian matching. While elegant, DETR is widely known to converge slowly and may underperform without careful design choices, which triggered a series of works that improve optimization and accuracy while retaining the overall encoder–decoder template. Among them, Deformable DETR (Zhu et al., 2020) stands out by replacing dense attention with sparse, sampling-based deformable attention, substantially accelerating training and strengthening performance.

Despite these rapid iterations, it is still not fully clear which ingredients are responsible for the practical advantage of modern DETR variants. In contrast, earlier detectors, most notably Faster R-CNN (Ren et al., 2015), follow a modular design: a backbone extracts features, a neck aggregates multi-scale information, an RPN generates candidate boxes, and a downstream head refines them, possibly across multiple refinement stages (Ren et al., 2015; He et al., 2017; Cai and Vasconcelos, 2018; Chen et al., 2019). This classical "propose-then-refine" pipeline provides a transparent interpretation of where localization and classification capacity originate.

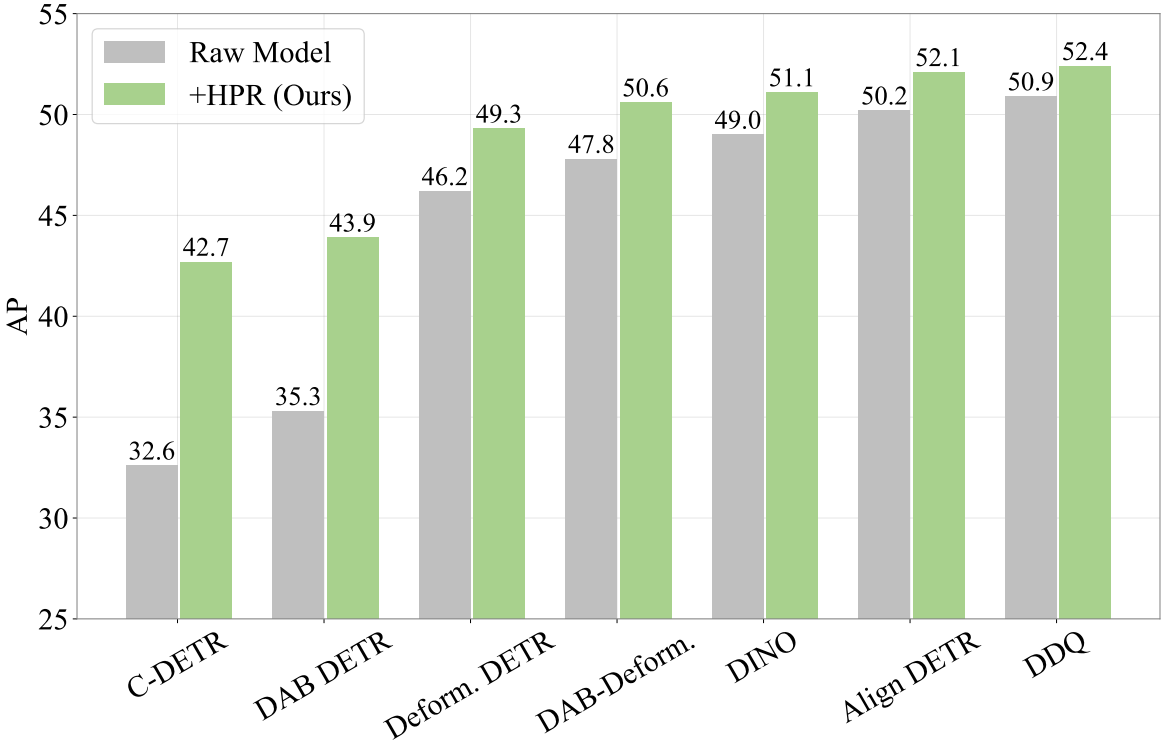


FIGURE 3.1. Applying Hybrid Proposal Refiner (HPR) to the DETR series including Conditional DETR (Meng et al., 2021a), DAB DETR (Liu et al., 2022), Deformable DETR (Zhu et al., 2020), DAB-Deformable DETR (Liu et al., 2022), DINO (Zhang et al., 2022), Align DETR (Cai et al., 2023) and DDQ (Zhang et al., 2023d) on COCO dataset. All models use a ResNet-50 backbone and a 12-epoch training schedule. For efficiency, we use 300 queries for DDQ (Zhang et al., 2023d) and DDQ equipped with HPR.

Motivated by this, we re-examine DETR-style detectors through the lens of Faster R-CNN. Our key hypothesis is that the DETR encoder–decoder can be understood as a modernized counterpart of the proposal-generation and proposal-refinement paradigm: the encoder acts as a stronger feature aggregator that prepares object-centric evidence, while the decoder plays the role of an iterative refiner operating on a fixed-size set of object representations. To make this comparison concrete, we use Deformable DETR (Zhu et al., 2020) with a ResNet-50 backbone as the primary reference model and progressively morph Faster R-CNN toward it by inserting the corresponding components one by one (Table 3.1). The modifications cover multiple dimensions, including: making the proposal generator class-aware, upgrading the neck from FPN to a deformable encoder, replacing RoI Align-style refinement with attention-based refinement (e.g., deformable attention), extending refinement from two stages to multiple

stages, and switching the training assignment from IoU-driven one-to-many matching to one-to-one Hungarian matching.

This controlled transformation reveals three important takeaways. (1) Directly introducing Hungarian matching into Faster R-CNN can noticeably hurt accuracy. A major reason is that one-to-one assignment tends to produce sharper, more localized activations, and when RoI Align pools from these maps, the resulting region features may contain excessive irrelevant context. (2) If the refinement head consumes object representations derived from the neck rather than pooled RoI features, the degradation caused by Hungarian matching is greatly reduced, suggesting that end-to-end, one-to-one training can also be viable in a modified Faster R-CNN-style framework. (3) The empirical gap between Faster R-CNN and Deformable DETR is largely explained by two upgrades: a more capable feature aggregation module (deformable encoder vs. FPN) and a more expressive refinement mechanism (attention-based refinement vs. RoI Align-based heads).

These observations naturally shift the focus to proposal refinement. In practice, a detector typically uses one feature aggregation module but may benefit from richer refinement. We therefore systematically study a family of refinement operators, including RoI Align, dynamic convolution, cross-attention, deformable attention, global attention, and object-feature refinement, and find that many of them provide complementary benefits and can be combined effectively. Based on this, we propose a Hybrid Proposal Refiner (HPR) that integrates multiple refinement operators and enables interaction among their features. As illustrated in Figure 3.1, HPR is lightweight, broadly compatible with existing DETR variants, and consistently improves their vanilla performance.

3.3 Methodology

Section 3.3.1 traces how Faster R-CNN can be progressively transformed into Deformable DETR. Motivated by the insight that DETR's encoder–decoder pipeline can be interpreted

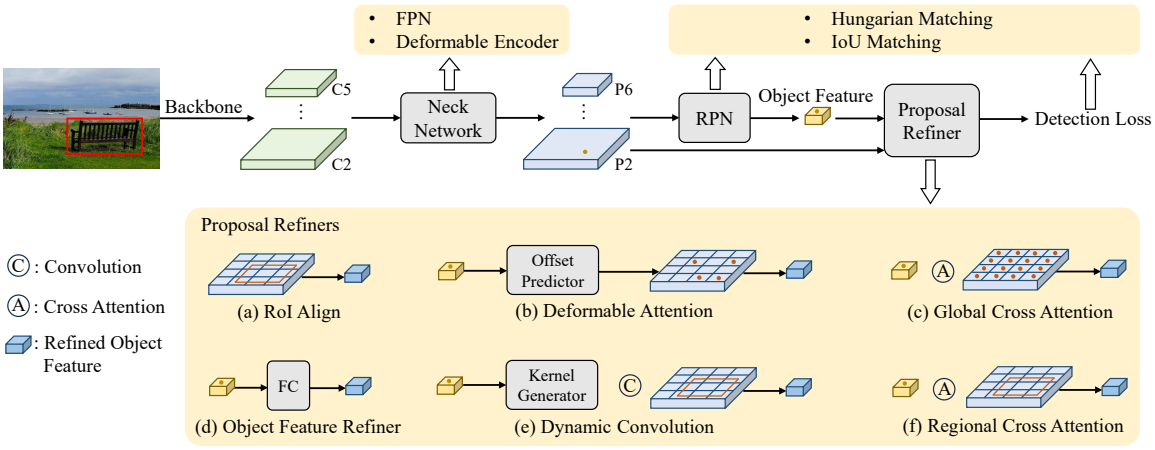


FIGURE 3.2. We regard the *encoder-decoder* structure employed by the DETR series as a refined version of the *RPN-refiner* paradigm utilized in Faster R-CNN. We investigate various elements (highlighted by yellow) that contribute to the transition from Faster R-CNN to Deformable DETR. Our hybrid proposal refiner (HPR) is predicated on exploring a multitude of proposal enhancement strategies that operate on different levels: regional (a, b, e, f), global (c), and point level (d).

as an enhanced counterpart of Faster R-CNN's proposal generation and refinement framework, Section 3.3.2 presents the Hybrid Proposal Refiner (HPR) and details how it can be incorporated into a range of DETR-based detectors.

3.3.1 From Faster R-CNN to Deformable DETR

As shown in Figure 3.2, we examine the key design choices that bridge Faster R-CNN and Deformable DETR, including the proposal generator (RPN), the feature aggregation module (neck), the refinement head, the number of refinement iterations, and the training assignment strategy. The accuracy of each step in this transition is summarized in Table 3.1.

Faster R-CNN baseline. We start from a standard Faster R-CNN equipped with a ResNet-50 backbone and an FPN neck, trained under a 12-epoch schedule. Region features are obtained via RoI Align. This baseline reaches 36.5 AP on the COCO `val` set.

From class-agnostic to class-aware proposals. To align the proposal stage with Deformable DETR, we replace the class-agnostic RPN in Faster R-CNN with a class-aware variant. This

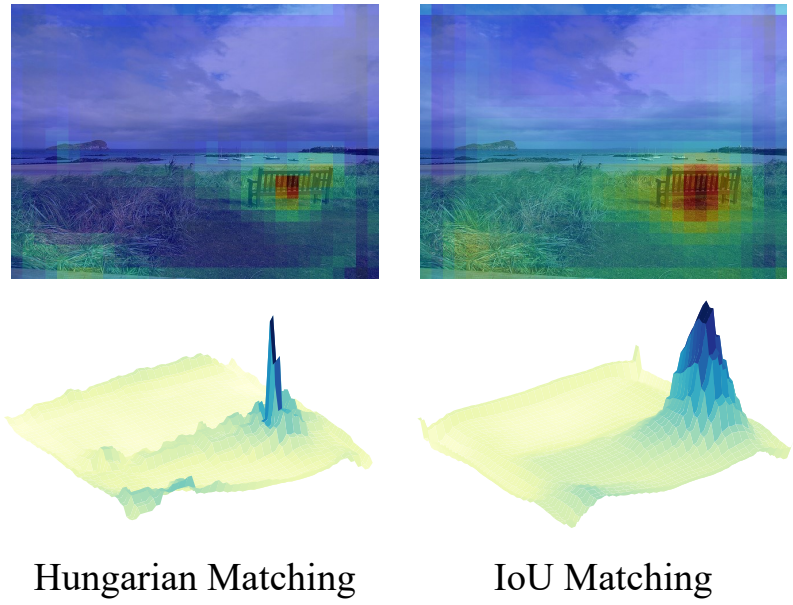


FIGURE 3.3. Visualization of two activation maps generated by variants of Faster R-CNN using either Hungarian matching or IoU matching.

change leads to a minor decrease in AP (36.5 $\rightarrow$ 36.1), indicating that proposal classification alone is not the primary driver of performance differences.

Upgrading the neck network. A major architectural distinction lies in the feature aggregation module. Replacing the FPN with the deformable encoder used in Deformable DETR substantially strengthens the detector, improving AP from 36.1 to 44.0. This highlights the importance of a more expressive, attention-based neck for building object-centric representations.

IoU-based assignment vs. Hungarian assignment. End-to-end DETR-style training relies on one-to-one bipartite assignment, whereas Faster R-CNN conventionally uses IoU-based one-to-many matching. When we switch the RPN assignment from IoU matching to Hungarian matching in our Faster R-CNN variant, performance drops sharply (44.0 $\rightarrow$ 32.7). Applying Hungarian matching at the R-CNN stage further reduces AP by 0.5. We attribute the large degradation mainly to an incompatibility between one-to-one assignment and RoI Align.

Our hypothesis is that Hungarian matching produces more peaked and localized feature activations, but RoI Align still pools a relatively large spatial region, thereby introducing

Matching Strategy	AP	AP_l	AP_m	AP_s
IoU	38.3	51.2	43.2	21.1
Hungarian	38.4	48.6	42.2	24.2

TABLE 3.2. The performance of the improved class-aware RPNs with different positive sample matching strategies.

substantial irrelevant context into the region features. We validate this in two ways. First, Figure 3.3 visualizes activation maps for two variants, showing that Hungarian matching yields noticeably sharper responses than IoU matching. Second, we retrain a detector that consists of a class-aware RPN with a deformable encoder under Hungarian matching, effectively operating as a strengthened single-stage model. As reported in Table 3.2, Hungarian matching does not harm this configuration, suggesting that the point-wise object features used by the proposal stage are already sufficient for accurate classification and localization.

Together, these quantitative and qualitative results support our conclusion: when Hungarian matching is introduced into Faster R-CNN, the RoI Align pooling becomes a bottleneck by gathering excessive non-essential information, which in turn leads to a pronounced performance drop.

RoI features vs. object features. Motivated by the analysis above, we replace the conventional second-stage head that relies on RoI Align, i.e., RoI Align $\rightarrow$ region feature $\rightarrow$ CNN $\rightarrow$ FC", with a lighter refiner that directly operates on proposal point features, namely object feature $\rightarrow$ FC". This change substantially improves performance under Hungarian assignment, boosting AP from 32.2 to 41.2. Moreover, Table 3.1 shows that fusing RoI-level cues into the object features yields an additional +0.5 AP gain. These results suggest that, under one-to-one matching, refinement operators that act on compact object representations are better aligned than RoI Align, which tends to introduce unnecessary context.

Stronger proposal refinement. Deformable DETR further upgrades refinement via deformable attention: each proposal representation is enhanced by sampling and aggregating features from a sparse set of informative points on the neck feature maps. Following this design, we replace the refiner "object feature + RoI feature $\rightarrow$ FC" with a deformable decoder,

enabling structured interaction between proposal embeddings and the deformable-encoder features. As reported in Table 3.1, this replacement brings a further +2.5 AP improvement. Finally, stacking $6\times$ deformable decoder layers yields 46.2 AP, completing our step-by-step transformation from Faster R-CNN to Deformable DETR.

3.3.2 Hybrid Proposal Refiner

The above study indicates that the advantage of Deformable DETR over Faster R-CNN mainly comes from two upgrades: a more expressive neck and a more capable refinement head. In practice, detectors typically have a single neck, but the refinement stage can be repeated and enriched in many ways. Hence, before introducing our Hybrid Proposal Refiner (HPR), we investigate alternative refinement operators beyond RoI Align (Figure 3.2.a) and deformable attention (Figure 3.2.b), as summarized in Figure 3.2.

Notations. Let the input resolution be $H \times W$. The backbone outputs multi-stage feature maps $\mathcal{C}_l \in \mathbb{R}^{H/2^l \times W/2^l \times d_l}$, where l indexes stages and d_l is the channel dimension. The neck network[1] produces encoded features $\mathcal{P}_l \in \mathbb{R}^{H/2^l \times W/2^l \times D}$ with unified dimension D. For the i-th proposal with box $\boldsymbol{b}_i = (x_i, y_i, w_i, h_i)$, we denote its proposal (point) representation as $\boldsymbol{p}_i \in \mathbb{R}^D$. The RoI feature associated with $\boldsymbol{b}_i$ is denoted by $\boldsymbol{r}_i \in \mathbb{R}^{7\times7\times D}$, obtained via RoI Align.

Global cross-attention (Figure 3.2.c). This operator follows the original DETR (Carion et al., 2020): a fixed set of learnable queries collects information from a neck feature map (e.g., $\mathcal{P}_5$) through cross-attention. While effective, global attention incurs high computation due to its dense interactions over the spatial tokens.

Object feature refinement (Figure 3.2.d). As analyzed in Section 3.3.1, one way to improve proposal quality is to update proposals directly in the object-feature space. Concretely, an object feature refiner takes the proposal embeddings $\boldsymbol{p}_i$ as input and outputs refined features, which are then used to update the RPN proposals.

[1]We use "neck" as an umbrella term for feature-enhancement modules placed between backbone and heads, including FPN, Transformer encoders, and deformable encoders.

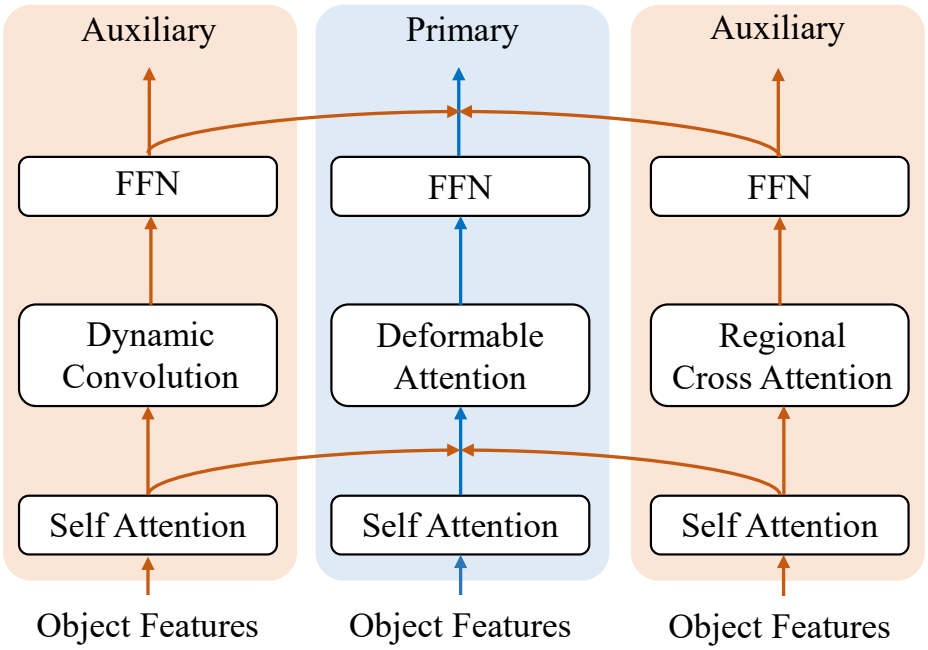


FIGURE 3.4. Illustration of the HPR module. The auxiliary refiners inject implicit information into the intermediate features of the primary refiner. We use $6\times$ HPRs by default.

Dynamic convolution (Figure 3.2.e). Dynamic convolution (Sun et al., 2021) strengthens each proposal embedding by coupling it with its associated RoI feature. Specifically, the object feature $\boldsymbol{p}_i$ is first fed through fully-connected layers to produce instance-specific convolution kernels. These kernels are then applied to the RoI tensor $\boldsymbol{r}_i$; the resulting responses are mapped through additional layers (e.g., convolution and FC) to yield an updated object representation for $\boldsymbol{p}_i$.

Regional cross-attention (Figure 3.2.f). Another way to fuse $\boldsymbol{p}_i$ with its RoI feature $\boldsymbol{r}_i$ is attention-based aggregation. We treat $\boldsymbol{p}_i$ as the query and use the spatial tokens within $\boldsymbol{r}_i$ as keys and values, allowing the proposal embedding to selectively attend to informative regions inside the RoI. We refer to this refinement operator as regional cross-attention.

Hybrid Proposal Refiner (HPR). Thus far, we have examined a spectrum of proposal refinement operators that act at different granularities: scene-level aggregation (global cross-attention), region-level fusion (RoI Align, deformable attention, dynamic convolution, and regional cross-attention), and proposal-level updates (object feature refinement). As discussed in Section 3.3.1, RoI Align is poorly suited to one-to-one Hungarian assignment, often introducing mismatched or redundant context. Meanwhile, our empirical results consistently indicate that effective end-to-end detectors rely on tightly coupling proposal embeddings with their corresponding localized visual evidence. This motivates HPR, which is designed to explicitly strengthen and unify the interaction between object features and region-derived signals while avoiding the limitations of RoI Align under Hungarian matching.

Unlike prior DETR variants that rely on a single refinement operator, our Hybrid Proposal Refiner (HPR) combines several complementary region-level refiners, including deformable attention, dynamic convolution, and regional cross-attention. Although all three aim to distill foreground-relevant cues for each proposal, they construct local evidence in fundamentally different ways. Deformable attention aggregates a sparse set of sampled point features around each proposal. In contrast, dynamic convolution and regional cross-attention both exploit RoI features, but they couple RoI features with proposal embeddings differently: dynamic convolution converts the object feature into instance-specific kernels, whereas regional cross-attention treats the object feature as the attention query to selectively read out informative RoI tokens.

As illustrated in Figure 3.4, HPR is designed to capitalize on the unique advantages of each refiner by assigning one module as the *primary* refiner and using the others as *auxiliary* refiners. The auxiliary branches provide additional refinement signals by injecting their intermediate representations into the corresponding blocks of the primary branch. Concretely, we fuse the self-attention outputs and FFN outputs from auxiliary refiners into those of the primary refiner through weighted summation, where the fusion weights are learnable. In Section 3.4.2, we further evaluate alternative fusion strategies. Each refiner is trained with its own detection objective; unless otherwise stated, we set the loss weights for the primary and

the two auxiliary refiners to 1.0, 0.5, and 0.5, respectively, and adopt the same loss formulation as Deformable DETR.

Using HPR within the DETR family. Similar to standard DETR-style decoders, HPR modules can be stacked to progressively improve proposal quality. By default, we stack six HPR layers. Moreover, HPR can be integrated into a wide range of DETR-based detectors that originally contain only a single refiner: we keep the original refiner as the primary branch and attach auxiliary refiners in parallel. As shown in Figure 3.1, this plug-and-play design yields consistent gains across multiple DETR variants.

Data re-augmentation. We additionally propose a simple yet effective augmentation recipe, termed *data re-augmentation*. Starting from a batch processed by standard (weak) augmentations, we duplicate the weakly-augmented samples and apply stronger transformations, such as color jitter and more aggressive geometric perturbations, to the duplicates. The final batch therefore contains paired weakly-augmented and strongly-augmented views. Compared with batch augmentation (Hoffer et al., 2020), our approach differs in two aspects: (i) it replicates already-augmented samples rather than raw inputs, and (ii) it uses stronger and distinct augmentation operators on the replicated views. Empirically, we find that data re-augmentation synergizes well with HPR and further improves detection performance.

3.4 Experiment

Dataset and evaluation. All experiments are performed on the COCO dataset (Lin et al., 2014). COCO provides 118,287 annotated training images covering 80 categories, and a validation split of 5,000 images. We use COCO *AP* on the `val` split as the primary metric.

Implementation details. Our implementation is based on the MMDetection framework (Carion et al., 2020). Unless stated otherwise, we adopt an ImageNet-1K pre-trained ResNet-50 backbone (He et al., 2016; Deng et al., 2009) and train for 12 epochs. We use 900 object queries by default. Optimization is done with AdamW (Loshchilov and Hutter, 2017) with a learning rate of 1×10^{-4}. We follow standard DETR-style augmentations used in recent works (Zong et al.,

Method	Backbone	#Queries	#Epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
Conditional DETR (Meng et al., 2021a)	R-50	300	108	43.0	64.0	45.7	22.7	46.7	61.5
Anchor DETR (Wang et al., 2022b)		300	50	42.1	63.1	44.9	22.3	46.2	60.0
Efficient DETR (Yao et al., 2021)		300	50	45.1	63.1	49.1	28.3	48.4	59.0
DAB DETR (Liu et al., 2022)		900	50	45.7	66.2	49.0	26.1	49.4	63.1
Deformable DETR (Zhu et al., 2020)		300	50	46.9	65.6	51.0	29.6	50.1	61.6
DN-Deformable DETR (Li et al., 2022b)		900	50	48.6	67.4	52.7	31.0	52.0	63.7
$\mathcal{H}$-Deformable DETR (Jia et al., 2023)		300	12	48.7	66.4	52.9	31.2	51.5	63.5
$\mathcal{H}$-Deformable DETR (Jia et al., 2023)		300	36	50.0	-	-	32.9	52.7	65.3
DINO (Zhang et al., 2022)		900	12	49.4	66.9	53.8	32.3	52.5	63.9
DINO (Zhang et al., 2022)		900	36	51.2	69.0	55.8	35.0	54.3	65.3
Group DETR (Chen et al., 2023d)		900	12	50.1	-	-	32.4	53.2	64.7
Align DETR (Cai et al., 2023)		900	12	50.2	67.8	54.4	32.9	53.3	65.0
Align DETR (Cai et al., 2023)		900	24	51.3	68.2	56.1	35.5	55.1	65.6
DETA (Ouyang-Zhang et al., 2022)		900	12	50.5	67.6	55.3	33.1	54.7	65.2
DETA (Ouyang-Zhang et al., 2022)		900	24	51.6	69.0	56.7	34.0	55.8	66.5
DDQ (Zhang et al., 2023d)		900	12	51.3	68.6	56.4	33.5	54.9	65.9
DDQ (Zhang et al., 2023d)		900	24	52.0	69.5	57.2	35.2	54.9	65.9
Deformable DETR with HPR		900	12	50.6	68.7	55.5	34.4	53.9	63.5
Deformable DETR with HPR		900	24	51.9	70.0	57.0	35.3	55.0	65.3
DINO with HPR		900	12	51.1	68.6	55.7	34.6	54.5	64.9
DINO with HPR		900	24	51.9	69.7	56.8	34.9	55.0	65.8
Align DETR with HPR		900	12	52.1	69.6	56.9	35.6	55.4	66.6
Align DETR with HPR		900	24	52.7	69.8	57.2	35.8	56.0	66.4
Align DETR with HPR†		900	12	52.4	70.3	57.2	35.9	56.3	68.5
Align DETR with HPR†		900	24	54.2	72.1	58.8	37.8	57.9	70.0
DDQ with HPR		300	12	52.4	69.9	57.5	35.9	55.5	66.7
DDQ with HPR		300	24	52.5	69.8	57.6	35.4	55.5	67.0
DDQ with HPR†		300	12	53.0	70.6	58.0	35.3	56.3	68.6
DDQ with HPR†		300	24	54.2	72.0	59.6	37.3	57.8	69.1
DDQ with HPR†		300	36	**54.9**	**72.4**	**60.3**	**37.7**	**58.9**	**69.6**

TABLE 3.3. Comparison with state-of-the-art DETR models on the COCO `val` set utilizing a ResNet-50 backbone. Considering GPU memory utilization, we use 300 queries when applying HPR to DDQ (Zhang et al., 2023d). †: the application of large-scale jitter data augmentation.

2023; Jia et al., 2023; Chen et al., 2023d; Cai et al., 2023; Zhang et al., 2022), and additionally apply our data re-augmentation strategy. For comparisons against stronger settings, we employ a larger Swin-L backbone (Liu et al., 2021b) pre-trained on ImageNet-22K (Deng et al., 2009), train with extended schedules (24 or 36 epochs), and augment the default pipeline with large-scale jitter and Copy-Paste (Ghiasi et al., 2021).

Method	Backbone	#Queries	#Epochs	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
HTC (Chen et al., 2019)		900	36	57.1	75.6	62.5	42.4	60.7	71.1
Group-DINO (Chen et al., 2023d)		900	36	58.4	-	-	41.0	62.5	73.9
DETA (Ouyang-Zhang et al., 2022)		900	24	58.5	76.5	64.4	38.5	62.6	73.8
DINO (Zhang et al., 2022)		900	12	57.5	-	-	-	-	-
DINO (Zhang et al., 2022)		900	36	58.5	77.0	64.1	41.5	62.3	74.0
DDQ (Zhang et al., 2023d)		900	36	58.7	76.8	64.5	41.6	62.9	74.3
Mask DINO (Li et al., 2023b)		300	50	59.0	-	-	-	-	-
$\mathcal{H}$-Deformable DETR (Jia et al., 2023)		900	12	55.9	-	-	39.1	59.9	72.2
$\mathcal{H}$-Deformable DETR (Jia et al., 2023)		900	36	57.1	-	-	39.7	61.4	73.4
$\mathcal{H}$-DINO (Jia et al., 2023)		900	36	59.4	77.8	65.4	43.1	63.1	74.2
DDQ with HPR	Swin-L	300	12	58.7	76.7	64.5	41.5	62.5	74.6
DDQ with HPR†		300	12	58.4	76.8	64.3	41.2	62.5	75.1
DDQ with HPR†		300	24	59.3	77.6	65.0	43.1	63.4	75.5
AlignDETR with HPR		900	12	58.6	76.8	64.0	40.9	62.7	75.4
AlignDETR with HPR		900	24	59.3	77.5	64.7	41.9	63.7	75.2
AlignDETR with HPR†		900	12	58.5	76.7	63.7	41.6	62.8	76.6
AlignDETR with HPR†		900	24	59.6	77.9	64.5	42.6	64.0	**76.9**
AlignDETR with HPR†		900	36	**60.0**	**78.0**	**65.5**	**43.8**	**64.5**	76.6

TABLE 3.4. Comparison with other DETR models on the COCO `val` set utilizing a Swin-L backbone pre-trained on ImageNet-22K. Considering GPU memory utilization, we use 300 queries when applying HPR to DDQ (Zhang et al., 2023d). †: the utilization of large-scale jitter.

Proposal Refiner	AP	AP_l	AP_m	AP_s
Global Cross Attention	42.3	56.3	45.4	26.8
RoI Align	32.2	37.3	36.9	23.6
Deformable Attention	47.8	62.0	51.2	30.6
Dynamic Convolution	48.3	62.7	51.2	32.3
Regional Cross Attention	47.6	61.5	50.6	31.7
Object Feature Refiner	41.2	51.7	44.8	26.9

TABLE 3.5. Performance comparison of different proposal refiners, including image-level aggregation (global cross-attention), region-level refinement (RoI Align, deformable attention, dynamic convolution, and regional cross-attention), and proposal-level updates (object feature refinement).

3.4.1 Main Results

As illustrated in Figure 3.1, HPR is readily integrated into a broad set of DETR-style detectors, including Conditional DETR (Meng et al., 2021a), DAB-DETR (Liu et al., 2022), Deformable DETR (Zhu et al., 2020), DAB-Deformable DETR (Liu et al., 2022), DINO (Zhang et al.,

2022), Align DETR (Cai et al., 2023), and DDQ (Zhang et al., 2023d). Across all these backbones, adding HPR consistently improves over the corresponding baseline models, with gains ranging from +1.5 to +10.1 AP.

Table 3.3 further reports comparisons under the ResNet-50 setting. In particular, equipping the strong DDQ baseline (Zhang et al., 2023d) with HPR yields 54.9 AP using a 36-epoch schedule. We also evaluate stronger backbones in Table 3.4. With a Swin-L backbone, our method reaches 59.3 AP on DDQ (Zhang et al., 2023d) and 60.0 AP on AlignDETR (Cai et al., 2023).

3.4.2 Ablation Studies

Unless stated otherwise, all ablations are conducted on the DINO-enhanced Deformable DETR baseline (Zhang et al., 2022). This baseline uses a ResNet-50 backbone with standard data augmentation, 300 object queries, and a 12-epoch schedule, reaching 47.8 AP.

Proposal Refiner Variants. Section 3.3.2 describes a set of refinement operators operating at different granularities: image-level aggregation (global cross-attention), region-level refinement (RoI Align, deformable attention, dynamic convolution, and regional cross-attention), and proposal-level updates (object feature refinement). Table 3.5 reports the performance of each choice. Except for RoI Align and the object-feature refiner, we evaluate these refiners with six refinement stages. Consistent with the findings in Section 3.3.1, RoI Align interacts poorly with Hungarian assignment and therefore yields inferior accuracy. Global cross-attention, while conceptually simple, is computationally costly and is less convenient for leveraging multi-scale features that are essential in modern DETR variants. Among the remaining options, deformable attention (DA), dynamic convolution (DC), and regional cross-attention (RCA) consistently perform best. We attribute their advantage to explicitly coupling proposal embeddings with localized visual evidence. For this reason, DA, DC, and RCA form the core components of our HPR.

Feature Integration. As illustrated in Figure 3.4, HPR fuses intermediate representations from the auxiliary refiners into the corresponding blocks of the primary refiner. Each refiner

SA	Dedicated Module	FFN	AP
✓			48.0
	✓		46.5
		✓	49.2
✓	✓		48.6
	✓	✓	48.6
✓		✓	**49.3**
✓	✓	✓	49.1

TABLE 3.6. Ablation study examining which types of auxiliary features contribute most when injected into the primary branch, including outputs from the self-attention (SA) layer, the feed-forward network (FFN), and the refinement module itself.

consists of a self-attention (SA) layer, a task-specific refinement module (deformable attention, dynamic convolution, or regional cross-attention), and a feed-forward network (FFN). Table 3.6 evaluates which types of auxiliary features contribute most when injected into the primary branch, including SA outputs, FFN outputs, or the outputs of the refinement module itself. Our experiments show that combining SA and FFN features provides the strongest gains, making this the default fusion strategy in HPR.

We further analyze how the fusion weights should be configured. Let $\boldsymbol{f}_p$, $\boldsymbol{f}_{a1}$, and $\boldsymbol{f}_{a2}$ denote the SA or FFN outputs from the primary refiner and the two auxiliary refiners, respectively. The fused representation is computed as $\boldsymbol{f}'_p = w_p\boldsymbol{f}_p + w_{a1}\boldsymbol{f}_{a1} + w_{a2}\boldsymbol{f}_{a2}$. Table 3.7 evaluates several design choices for the weight parameters, including: (1) whether the weights w_p, w_{a1}, w_{a2} are fixed constants or learnable parameters; (2) whether each weight is a single scalar or a feature-wise vector matching the dimensionality of $\boldsymbol{f}_p$; and (3) the effect of different initialization values for these weights.

Performance Enhancement. Section 3.3.2 introduced our data re-augmentation strategy. Table 3.8 summarizes its impact, along with increasing the number of object queries from 300 to 900. Data re-augmentation is performed by first duplicating normally augmented samples, then applying stronger transformations to the duplicates, and finally training on the union of the weakly augmented and strongly augmented batches. To assess its effectiveness, Table 3.9

Weight	Type	Initialization	AP
Fixed	Scalar	1:1:1	48.9
Fixed	Scalar	2:1:1	49.1
Learnable	Scalar	1:1:1	48.9
Learnable	Scalar	2:1:1	48.8
Learnable	Vector	1:1:1	**49.3**
Learnable	Vector	2:1:1	49.0

TABLE 3.7. Ablation study on the integration weights, where the weights correspond to the primary refiner and the two auxiliary refiners, respectively.

HPR	Data Re-Augmentation	900 Queries	AP
			47.8
✓			49.3
✓		✓	49.8
✓	✓		50.3
✓	✓	✓	**50.6**

TABLE 3.8. Study on data re-augmentation and more object queries. We verify the data re-augmentation strategy introduced in Section 3.3.2 and increasing the number of object queries from 300 to 900.

Augmentation Strategy	AP
Normal Augmentation	49.3
Strong Augmentation	48.4
Batch Augmentation (Hoffer et al., 2020)	49.6
Data Re-Augmentation	**50.3**

TABLE 3.9. Comparison among standard data augmentation (the first and second rows), batch augmentation (Hoffer et al., 2020) (the third row), and data re-augmentation (the last row).

compares data re-augmentation against standard weak/strong augmentation pipelines and against batch augmentation (Hoffer et al., 2020).

Method	Backbone	#Queries	#Epochs	AP
DiffusionDet	R50	500	-	46.8
Lite DETR	R50	900	36	49.5
Decoupled DETR	R50	300	50	47.0
Plain DETR	Swin-T	300	12	50.9
$\mathcal{C}$o-DETR	R50	900	12	52.1
Ours (w/ DDQ)	R50	300	12	**53.0**

TABLE 3.10. Comparison with the latest models (Chen et al., 2023e; Li et al., 2023a; Zhang et al., 2023a; Lin et al., 2023a; Zong et al., 2023). When paired with DDQ, HPR attains an AP of 53.0, outperforming all competing models.

Deformable Att.	Dynamic Conv.	Regional C.A.	AP
✓			47.8
✓	✓		48.9
✓		✓	48.4
✓	✓	✓	**49.3**

TABLE 3.11. Study on the integration of auxiliary proposal refiners (dynamic convolution and regional cross attention) into the primary proposal refiner (deformable attention). Refer to the supplementary materials for more results.

3.4.3 Analysis

Comparison with More Latest Models. Table 3.10 further benchmarks HPR against several recent state-of-the-art approaches (Chen et al., 2023e; Li et al., 2023a; Zhang et al., 2023a; Lin et al., 2023a; Zong et al., 2023). When paired with DDQ, HPR attains an AP of 53.0, outperforming all competing models included in the comparison.

Integration of Auxiliary Object Refiners into Primary Object Refiner. We use deformable attention as the primary proposal refiner. Table 3.11 reports the gains obtained when augmenting it with dynamic convolution, regional cross-attention, or both as auxiliary branches. Each auxiliary refiner provides additional benefits, consistently boosting the performance beyond that of using the primary refiner alone.

Ablation Study on Primary Object Refiners. Figure 3.4 showcases the design where deformable attention serves as the primary refiner, with dynamic convolution and regional

Primary	Auxiliary-1	Auxiliary-2	AP
Deformable Att.	-	-	47.8
Deformable Att.	Deformable Att.	Deformable Att.	48.5
Regional CA	Dynamic Conv.	Deformable Att.	48.9
Dynamic Conv.	Regional CA	Deformable Att.	48.8
Deformable Att.	Regional CA	Dynamic Conv.	**49.3**

TABLE 3.12. Ablation study on primary object refiners. Att.: attention. CA: cross attention. Conv.: convolution.

Loss Weight	AP	AP_l	AP_m	AP_s
1:1:1	49.1	63.8	51.7	32.5
2:1:1	**49.3**	62.8	52.4	32.6

TABLE 3.13. Ablation study the effect of varying loss weights assigned to the primary and auxiliary refiners.

cross-attention acting as auxiliary branches. Table 3.12 further explores alternative configurations by assigning the primary role to regional cross-attention or dynamic convolution, and comparing them against the default setup where deformable attention remains the primary refiner.

We also include a baseline that uses deformable attention as the sole primary refiner supported by two auxiliary branches of the same type. Across all comparisons, HPR consistently outperforms these baselines, demonstrating that combining heterogeneous regional refinement operators offers clear advantages over relying on a single refinement mechanism.

Ablation Study on Loss Weight. We further investigate the effect of varying loss weights assigned to the primary and auxiliary refiners. As shown in Table 3.13, a weight configuration of 2:1:1 yields the best performance, reaching 49.3 AP.

Examination of Encoder and Decoder Number Variations. We analyze how different combinations of encoder and decoder depths influence overall performance. Specifically, in a setup with six encoders, we vary the number of decoders from 1 to 6, and apply the same

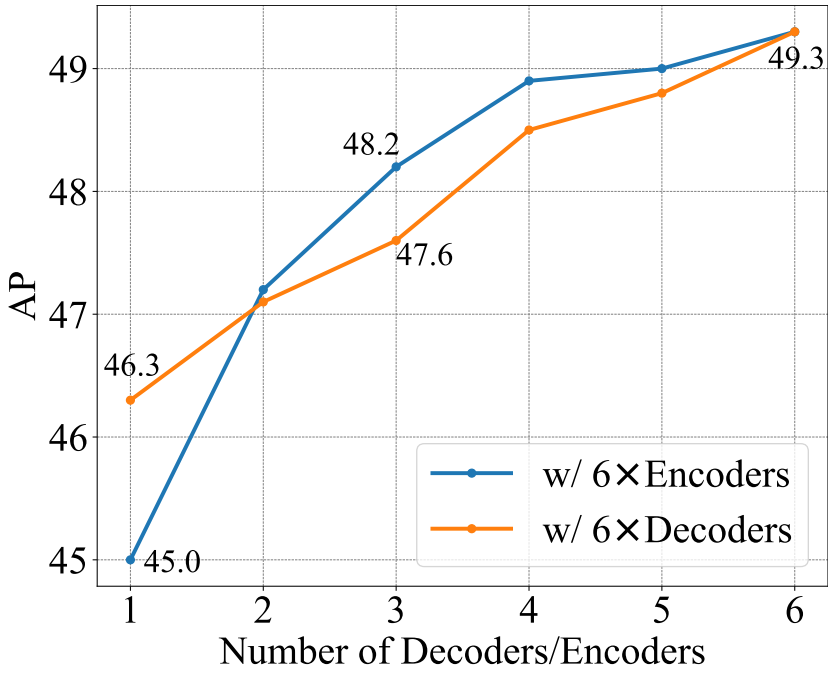


FIGURE 3.5. Ablation study on variations in the number of encoders (deformable encoders) and decoders (HPRs). Blue line: variation in the number of decoders within a model with 6× encoders. Orange line: variation in the number of encoders within a model with 6× decoders.

variation in a model configured with six decoders. The corresponding results are summarized in Figure 3.5.

Operational Mechanisms of Various Proposal Refinement Strategies. In terms of how object features are leveraged, deformable attention, dynamic convolution, and regional cross-attention behave quite differently. Deformable attention samples a sparse set of keypoints around each proposal and aggregates their features. Dynamic convolution converts the proposal embedding into instance-specific kernels, which are then applied over the corresponding RoI feature map to produce an updated object representation. Regional cross-attention fuses proposal and RoI features by treating the proposal embedding as the query and the RoI tokens as keys and values. Figure 3.6 visualizes the activation patterns of these refiners, revealing that each mechanism attends to distinct spatial regions and semantic cues of the target object.

We further compute cosine-similarity statistics between the features produced by pairs of proposal refiners across all object queries and all images in the COCO `val` set. This yields

FIGURE 3.6. Visualizations of the activation maps for deformable attention (the second row), dynamic convolution (the third row), and regional cross attention (the last row).

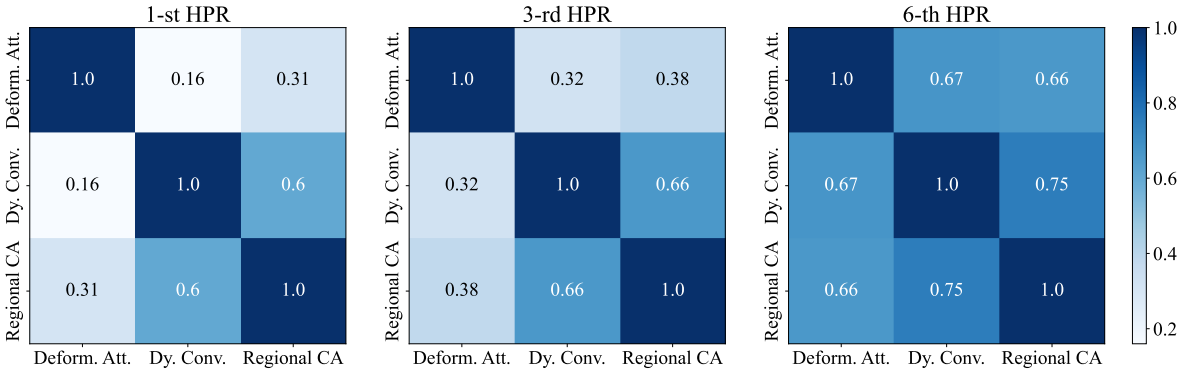


FIGURE 3.7. Visualizations for cosine similarities of various proposal refiners in distinct HPR stages.

an average similarity score s, which reflects how closely two refiners encode proposal features—higher values indicate more aligned representations. For this analysis, we extract features from the first stage, an intermediate stage (stage 3), and the final stage of HPR. The resulting similarity matrices are shown in Figure 3.7. As illustrated, the refiners produce

FIGURE 3.8. Visualizations of the activation maps generated by variants of Faster R-CNN using either IoU matching (the second row) or Hungarian matching (the third row).

noticeably different representations in the early and mid stages, whereas their outputs become substantially more alike at the final stage, suggesting that HPR gradually guides different refiners toward a shared latent representation space.

Qualitative Study on Positive Sample Matching Strategies. In Figure 3.3, we compare activation maps produced by Faster R-CNN variants trained with either Hungarian matching or IoU matching. Additional examples are provided in Figure 3.8, offering a more comprehensive visualization of the differences between the two matching strategies.

Training Curve Analysis. Figure 3.9 presents a comparison of training dynamics for Align DETR (Cai et al., 2023) enhanced with our HPR, its original counterpart, and two additional DETR variants including DINO (Zhang et al., 2022) and Deformable DETR (Zhu et al., 2020). Incorporating HPR leads to noticeably faster convergence.

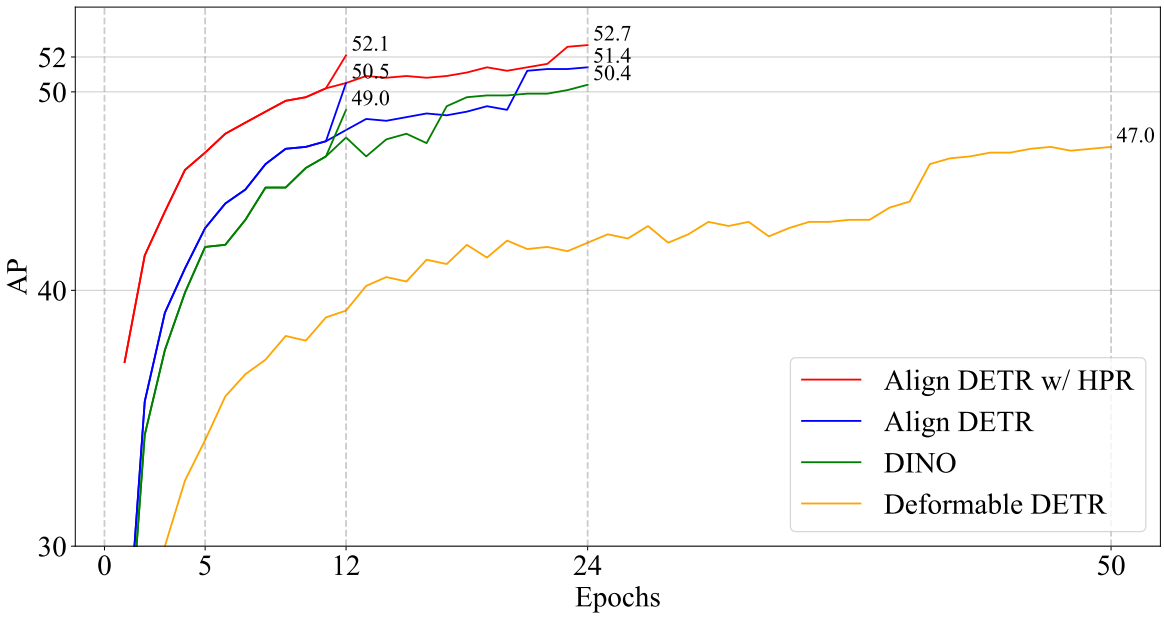


FIGURE 3.9. Training curves for AlignDETR equipped with our HPR, the original AlignDETR, DINO, and Deformable DETR.

#Epochs	Data Re-aug.	LSJ	AP
12			49.3
	✓		50.3
		✓	49.3
	✓	✓	50.4
24			50.5
	✓		51.3
		✓	51.6
	✓	✓	52.8

TABLE 3.14. Ablation study on data re-augmentation and large-scale jitter (LSJ) augmentation.

Ablation Study on Data Augmentations. Table 3.14 evaluates the impact of our proposed data re-augmentation and large-scale jitter augmentation. The results show that the two strategies are complementary and can be applied together effectively.

3.5 Chapter Summary

This chapter presented a comprehensive study of object detection through the lens of modern Transformer-based architectures, culminating in the development of the Hybrid Proposal Refiner (HPR). We began by outlining the fundamentals of object detection, emphasizing its role in identifying and localizing instances within images and its importance across a wide range of real-world applications such as autonomous driving, robotics, AR/VR, retail automation, and medical imaging.

We then revisited the evolution of detection models, contrasting the classical two-stage Faster R-CNN pipeline with the more recent DETR family. By progressively transforming Faster R-CNN into Deformable DETR, we gained insights into the architectural components that drive performance differences. Our analysis revealed three key findings: (1) one-to-one Hungarian matching conflicts with RoI Align and leads to degraded performance; (2) replacing region-pooled features with compact object features largely mitigates this issue; and (3) the performance advantage of Deformable DETR comes primarily from its stronger feature aggregation (deformable encoder) and more expressive refinement head (deformable attention).

Building on these observations, we explored a broad set of proposal refinement modules including RoI Align, dynamic convolution, regional cross-attention, deformable attention, global attention, and object-level refinement, and showed that many operate at different granularities and capture complementary cues. This inspired the design of HPR, which integrates multiple refiners into a unified framework. HPR designates one refiner as the primary branch and incorporates the others as auxiliary branches that inject their intermediate features through learnable fusion mechanisms. This design strengthens proposal updates while avoiding the limitations of RoI Align under Hungarian matching.

Extensive experiments demonstrated that HPR can be seamlessly applied to a wide spectrum of DETR-style detectors, such as Conditional DETR, DAB-DETR, Deformable DETR, DINO, AlignDETR, and DDQ, yielding consistent improvements (up to +10.1 AP). We further introduced a data re-augmentation strategy that pairs weakly augmented samples with

strongly augmented counterparts, providing additional performance gains and synergizing well with HPR. Ablation studies validated the importance of integrating self-attention and FFN features, the benefits of learnable fusion weights, and the complementary strengths of deformable attention, dynamic convolution, and regional cross-attention.

Overall, this chapter established a unified perspective that connects classical two-stage detectors with modern Transformer-based models and introduced a general, effective proposal refinement framework that advances the state of the art in end-to-end object detection.

CHAPTER 4

Video Perception under Extremely Sparse Annotations

This chapter focuses on visual recognition and localization in videos, which introduce additional challenges absent in static-image settings (e.g., object detection in Chapter 3), such as the need to model long-range temporal dependencies across frames. The chapter emphasizes learning under extremely limited supervision while effectively leveraging large amounts of unlabeled video data. Specifically, we study the problem of low-shot video object segmentation (VOS), where each training video contains only one or two annotated frames. Section 4.1 presents the problem formulation of VOS. Section 4.2 discusses the motivation, highlighting that annotating videos is time-consuming and costly, and shows that analyzing training under sparse video annotations provides a feasible foundation for effective model learning. In Section 4.3, we formulate low-shot VOS as an extreme semi-supervised learning setting and, from this perspective, propose a simple yet effective two-phase training paradigm (Yan et al., 2025) that fully exploits the information contained in unlabeled frames. Section 4.4 demonstrates that the proposed approach is model-agnostic and generalizes well across diverse VOS architectures (STCN, RDE-VOS, XMem) and multiple datasets (DAVIS 2016/2017, YouTube-VOS 2018/2019, LVOS, and VOST). Finally, Section 4.5 summarizes the chapter.

4.1 Problem Formulation

Video Object Segmentation. Video Object Segmentation (VOS) aims to generate a pixel-level mask that consistently follows a target object throughout a video as it moves, changes shape, or undergoes variations in lighting, occlusion, or camera viewpoint. In practice, VOS is typically initialized by providing the object's segmentation in the first frame, either manually

or via an automatic detector, and the objective is to propagate this mask to all subsequent frames in the video, producing temporally coherent and spatially accurate masks across the sequence.

Semi-Supervised Learning. In semi-supervised VOS, only a subset of frames have ground-truth annotations. Let $L \subset \{1, \ldots, T\}$ denote the indices of the labeled frames with corresponding masks $\{M_t^{\text{gt}}\}_{t \in L}$. The remaining frames are unlabeled, and the model must infer their segmentation masks by leveraging the annotated frames, temporal continuity, and learned appearance priors. This setting reflects realistic constraints where dense video annotations are prohibitively expensive.

Low-Shot Video Object Segmentation. Low-shot VOS is an even more challenging regime of semi-supervised segmentation. Instead of annotating only the first frame, the annotator provides pixel-level masks for only *one or two* frames in the entire video, i.e., $|L| \in \{1, 2\}$. These labeled frames may be temporally distant from one another, requiring the model to propagate semantic information across long temporal gaps. Key challenges include: (1) long-range temporal propagation, (2) handling dramatic appearance changes and occlusions, and (3) learning robust object representations from extremely sparse supervision.

Applications. Video object segmentation is a fundamental problem with broad applications across video editing and post-production (e.g., object cutout, background replacement, and compositing), autonomous driving and robotics (e.g., perception, tracking, and manipulation), AR/VR systems (e.g., real-time interactive effects), medical video analysis (e.g., surgical instrument or anatomical structure segmentation), and sports or surveillance analytics. Its pivotal role across domains underscores the need for efficient semi-supervised and low-shot VOS approaches.

4.2 Motivation

Video object segmentation (VOS) aims to segment a target object throughout a video given its annotation in the reference (typically the first) frame (Oh et al., 2019; Yang et al., 2020; Seong

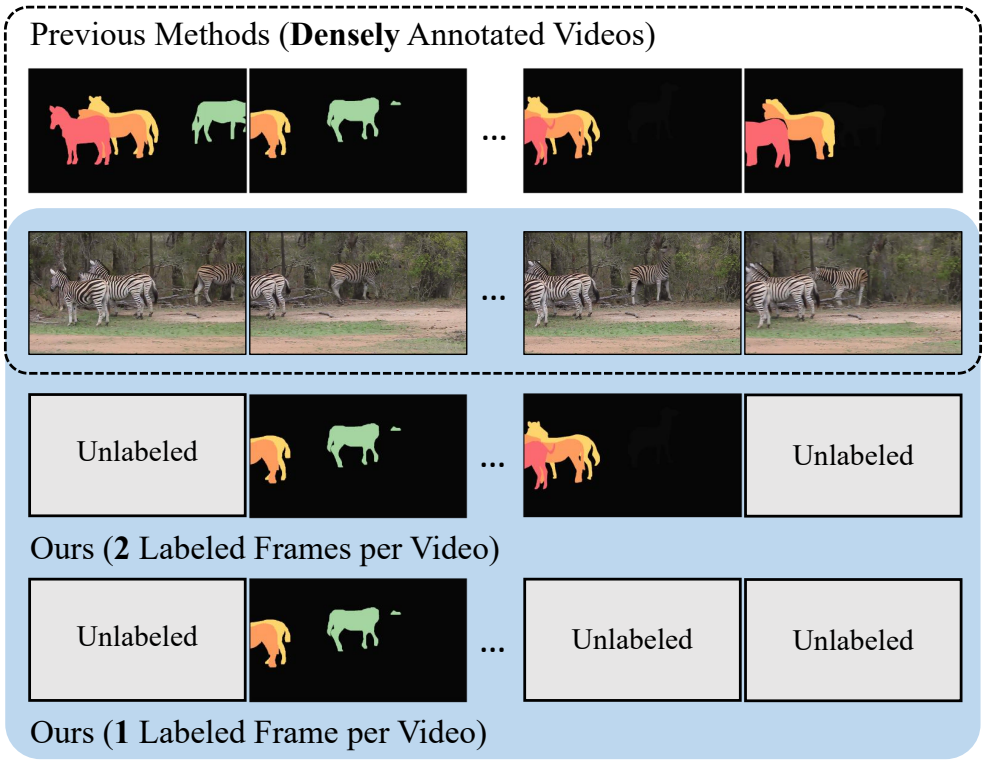


FIGURE 4.1. Previous works on video object segmentation rely on densely annotated videos, while we only require one or two labeled frames per video.

et al., 2021; Fan et al., 2023). State-of-the-art VOS models (Oh et al., 2019; Yang et al., 2020; Seong et al., 2021; Cheng et al., 2021c; Xie et al., 2021; Li et al., 2022c; Cheng and Schwing, 2022) are commonly trained on densely annotated datasets such as DAVIS (Perazzi et al., 2016; Pont-Tuset et al., 2017) and YouTube-VOS (Xu et al., 2018), where each video provides tens or even hundreds of pixel-level masks. However, such dense annotations are extremely labor-intensive and expensive to acquire. For instance, the DAVIS benchmark contains 60 videos with an average of 70 labeled frames per video, while YouTube-VOS labels every fifth frame solely to reduce annotation cost. These constraints highlight the importance of developing *data-efficient* VOS models that can operate effectively with significantly fewer labeled frames.

Practicality of Low-Shot VOS. In this work, we explore whether strong VOS models can be trained with only *one or two* annotated frames per training video (Figure 4.1). We refer to this setting as N-shot VOS, where N denotes the number of annotated frames per video.

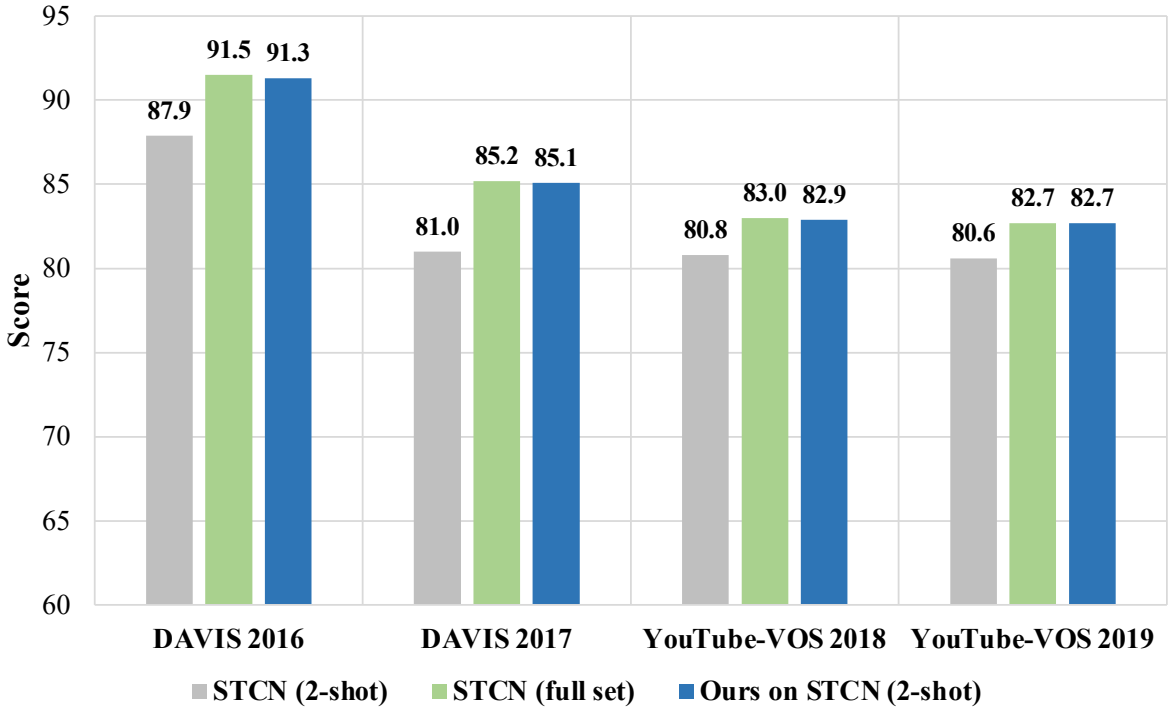


FIGURE 4.2. Comparison under 2-shot setting. The naive 2-shot STCN exhibits only a modest performance drop compared to its full-set counterpart (e.g., -2.1% on YouTube-VOS 2019), indicating that low-shot VOS is more feasible than previously believed. Our approach enables 2-shot STCN to achieve performance nearly identical to a fully supervised model.

Using STCN (Cheng et al., 2021c) as a baseline, we first train a naive 2-shot model on YouTube-VOS and DAVIS. Surprisingly, as shown in Figure 4.2, the naive 2-shot STCN exhibits only a modest performance drop compared to its full-set counterpart (e.g., -2.1% on YouTube-VOS 2019), indicating that low-shot VOS is more feasible than previously believed.

Two-Shot VOS via Semi-Supervised Learning. Despite the encouraging results, existing 2-shot training does not fully exploit the abundant unlabeled frames available in each video. Semi-supervised learning, which leverages a small amount of labeled data together with numerous unlabeled samples, has proven effective across image classification (Sohn et al., 2020a; Berthelot et al., 2019b), object detection (Sohn et al., 2020b; Xu et al., 2021), and semantic segmentation (Hu et al., 2021a; Ke et al., 2020). Inspired by this paradigm, we propose to enhance low-shot VOS by generating reliable pseudo labels for unlabeled frames and jointly optimizing the model on labeled and pseudo-labeled data (Figure 4.1).

To illustrate our design, we again consider STCN. During training, each STCN iteration samples a triplet of frames in which supervision is applied only to the last two frames, as the first frame serves as the reference. This suggests a natural strategy: use the ground-truth first frame to avoid early error accumulation, while allowing the last two frames to be either labeled frames or unlabeled frames with high-quality pseudo labels. This constitutes our *phase-1* training, which already improves upon naive 2-shot training.

To fully exploit unlabeled data, we remove the restriction that the starting frame must be labeled. Using the phase-1 model, we generate pseudo labels for all unlabeled frames and store them in a pseudo-label bank via an *intermediate inference* step. We then retrain the VOS model (*phase-2*) on a mixture of ground-truth labels and pseudo labels, updating the pseudo-label bank as predictions become more accurate. As shown in Figure 4.2, this approach enables 2-shot STCN to achieve performance nearly identical to a fully supervised model (e.g., 85.2% vs. 85.1% on DAVIS 2017, 82.7% vs. 82.7% on YouTube-VOS 2019), while using only 7.3% and 2.9% of the labeled data.

One-Shot VOS. The strong results in the 2-shot setting naturally raise the question: *Can we train an effective VOS model with only a single annotated frame per video?* However, naively reducing labeled frames harms the model's generalization ability and degrades pseudo labels. To address this, we incorporate SAM (Kirillov et al., 2023c) during intermediate inference as a complementary universal segmentation model capable of producing per-frame masks from box or point prompts. We employ lightweight fine-tuning and point-prompt augmentation to mitigate domain shift, and introduce a mask assessment module to select the better prediction between SAM and the one-shot VOS model for each frame. This collaboration significantly improves pseudo-label quality, enabling strong one-shot VOS performance (Figure 4.3) while keeping inference cost unchanged, as SAM is not used during inference.

Our main contributions are threefold:

- We present the first systematic study demonstrating the feasibility of *low-shot video object segmentation*.

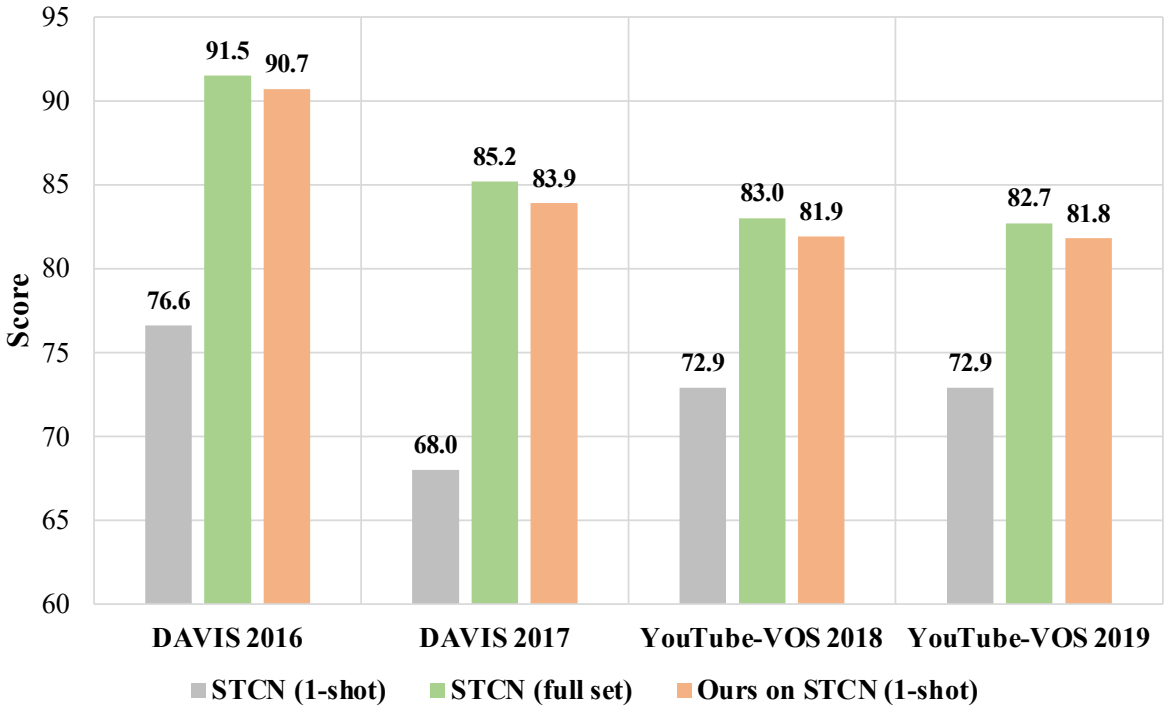


FIGURE 4.3. Comparison under 1-shot setting. Our approach enables strong one-shot VOS performance while keeping inference cost unchanged.

- We propose a simple yet effective semi-supervised training paradigm that unlocks the potential of unlabeled frames and is compatible with various VOS models, including STCN (Cheng et al., 2021c), RDE-VOS (Li et al., 2022c), and XMem (Cheng and Schwing, 2022).
- Despite using only a small portion of labeled data (e.g., 7.3% for YouTube-VOS and 2.9% for DAVIS in the two-shot setting; 3.7% and 1.4% in the one-shot setting), our method achieves performance on par with fully supervised models, as depicted in Figures 4.2 and 4.3.

4.3 Methodology

We begin by revisiting the fundamentals of VOS in Section 4.3.1. Section 4.3.2 then formalizes the low-shot VOS problem and presents an overview of our approach. The training

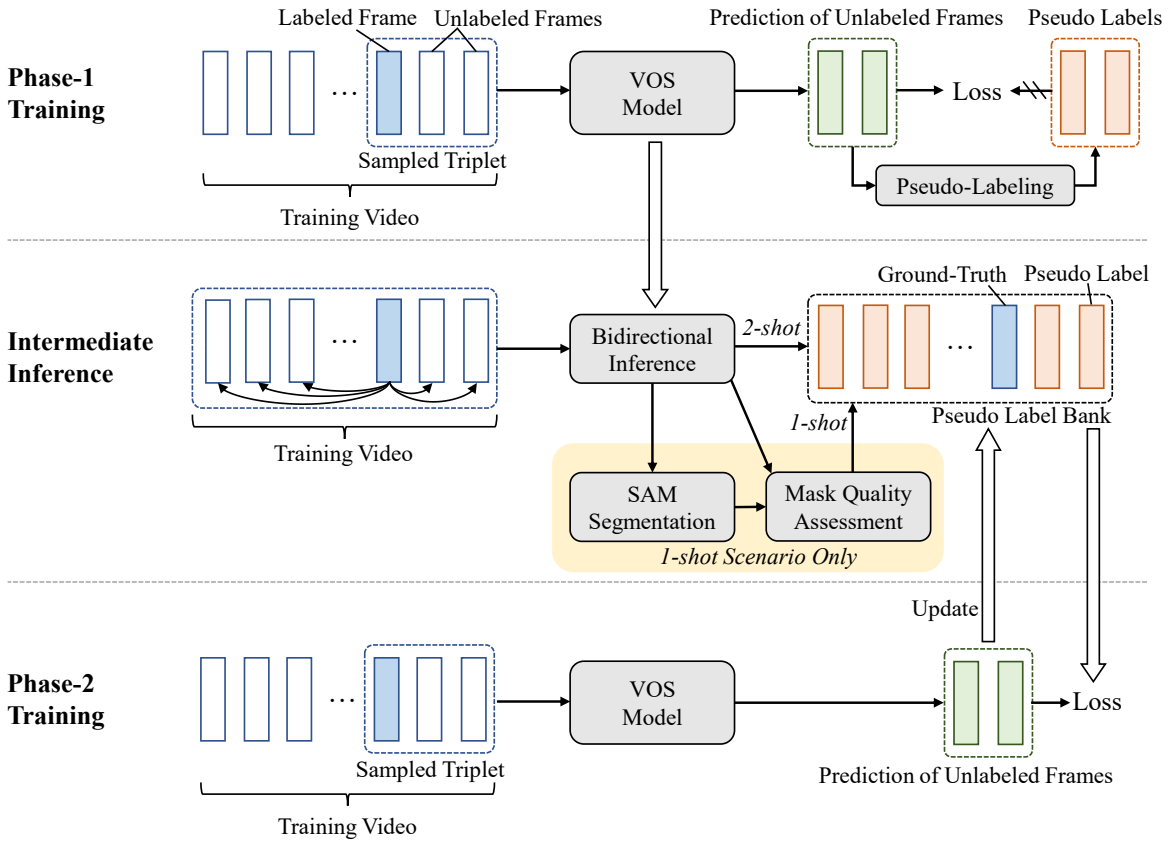


FIGURE 4.4. Overview of our methodology. During phase-1 training (top), we optimize a VOS model (*i.e.* STCN) which takes a triplet of frames as input on a low-shot VOS dataset in a semi-supervised manner. We constrain the reference (first) frame to be a labeled frame to ease the learning. The remaining frames can be either labeled or unlabeled. Then we perform an intermediate inference (middle) to generate pseudo labels for unlabeled frames by the VOS model trained in phase-1, and construct a pseudo-label bank to store the pseudo labels in addition to the ground-truth. During phase-2 training (bottom), we re-train a VOS model, which could be most models, on the combination of labeled and pseudo-labeled data without any restrictions on the first frame. The pseudo-label bank is dynamically updated once more reliable pseudo labels are identified during phase-2 training. Note that the SAM segmentation module along with the mask quality assessment module are specifically designed for the one-shot VOS training.

procedures for two-shot and one-shot VOS models are detailed in Section 4.3.3 and Section 4.3.4, respectively. Finally, Section 4.3.5 demonstrates that our training paradigm is broadly applicable to a wide range of VOS architectures.

4.3.1 Preliminary

Conventional VOS methods are trained on densely annotated videos, where the ground-truth mask of the first frame is used to supervise the segmentation of all subsequent frames. The standard training objective is therefore to maximize the accuracy of mask predictions from the second frame onward. For example, STM (Oh et al., 2019) and STCN (Cheng et al., 2021c) process triplets of frames during training, while RDE-VOS (Li et al., 2022c) and XMem (Cheng and Schwing, 2022) extend this paradigm by modeling longer temporal contexts using 5-frame and 8-frame sequences, respectively.

In our setting, only one or two labeled frames are available for each training video. Without loss of generality, we use STCN (Cheng et al., 2021c) to illustrate our low-shot training strategy, though the same paradigm can be applied to any VOS model as discussed in Section 4.3.5. For a given training video, STCN samples a triplet of frames and predicts the mask of the second frame using the ground-truth annotation of the first frame. It then predicts the mask of the third frame using both the first-frame ground truth and the predicted mask of the second frame. A standard segmentation loss is applied to supervise each of these two predictions.

4.3.2 Overview

Problem formulation. Given a VOS dataset $\mathcal{D}$, each training video $\mathcal{V} = [\boldsymbol{V}_1, \ldots, \boldsymbol{V}_T] \in \mathcal{D}$ consists of T frames ($T \gg 2$) with corresponding ground-truth masks $\mathcal{Y} = [\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_T]$. For the one-shot and two-shot settings, we randomly select one or two frames as labeled data, respectively, while treating all remaining frames as unlabeled. Our goal is to train a VOS model that effectively leverages both the limited labeled frames and the abundant unlabeled frames during learning.

Overview. Figure 4.4 provides an overview of our low-shot VOS framework. We begin with *phase-1 training*, where a VOS model is trained in a semi-supervised fashion while enforcing that the reference frame is always labeled. Next, we conduct an *intermediate inference* step:

the phase-1 model is used to generate pseudo labels for all unlabeled frames, and these pseudo labels are stored in a pseudo-label bank for efficient access.

In *phase-2 training*, we retrain the VOS model using both labeled and pseudo-labeled frames, without imposing any constraints on which frame serves as the reference. The pseudo-label bank is continually updated as more accurate pseudo labels are produced during training.

For the one-shot setting, we additionally incorporate a SAM-based segmentation module and a mask quality assessment module to enhance pseudo-label quality. The designs of these modules are detailed in Section 4.3.4.

4.3.3 Two-Shot VOS Training

Phase-1 training. We adopt STCN (Cheng et al., 2021c) as our base architecture, which processes a triplet of frames as input. However, in the two-shot setting, each training video contains only two labeled frames, insufficient for forming a supervised triplet. To address this, we employ a semi-supervised training strategy that augments the labeled frames with pseudo-labeled frames, enabling valid triplet construction.

Since STCN requires the annotated reference (first) frame to segment the target object in subsequent frames, we always sample a labeled frame as the reference during phase-1 training to mitigate early-stage error propagation. The remaining two frames in the triplet can be either labeled or unlabeled. In practice, we assign a 50% probability that both frames are unlabeled, and a 50% probability that one is labeled and one is unlabeled. All frames are drawn from the same training video, with unlabeled frames randomly sampled in each iteration.

Training in the two-shot setting closely mirrors full-set training, except that each triplet now consists of a mixture of ground-truth–labeled frames and pseudo-labeled frames. For a sampled triplet in which the last two frames contain N_1 labeled frames and N_2 unlabeled frames ($N_1 = 1, N_2 = 1$ or $N_1 = 0, N_2 = 2$), the total loss is defined as:

$$\mathcal{L} = \mathcal{L}_S + \mathcal{L}_U.$$

Supervised Loss. The supervised loss $\mathcal{L}_S$ is a standard pixel-wise segmentation loss applied to the labeled frames:

$$\mathcal{L}_S = \frac{1}{HWN_1} \sum_{n=1}^{N_1} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathcal{H}(\boldsymbol{Y}_n^{(i,j)}, \boldsymbol{P}_n^{(i,j)}), \tag{4.1}$$

where H and W denote the spatial resolution, $\mathcal{H}(\cdot, \cdot)$ is the cross-entropy loss, $\boldsymbol{P}_n^{(i,j)}$ is the predicted probability at pixel (i, j), and $\boldsymbol{Y}_n^{(i,j)}$ is the corresponding ground-truth label.

Unsupervised Loss. The unsupervised loss $\mathcal{L}_U$ applies the same segmentation loss to unlabeled frames using filtered pseudo labels:

$$\mathcal{L}_U = \frac{1}{HWN_2} \sum_{n=1}^{N_2} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbb{1}_{[\max(\boldsymbol{P}_n^{(i,j)}) \geq \tau_1]} \mathcal{H}(\hat{\boldsymbol{Y}}_n^{(i,j)}, \boldsymbol{P}_n^{(i,j)}), \tag{4.2}$$

where $\mathbb{1}_{[\cdot]}$ is an indicator function filtering out predictions whose maximum confidence is below a threshold τ_1, and $\hat{\boldsymbol{Y}}_n^{(i,j)} = \mathrm{argmax}(\boldsymbol{P}_n^{(i,j)})$ denotes the corresponding one-hot pseudo label. We set $\tau_1 = 0.9$ by default to ensure pseudo-label reliability.

As training progresses, an increasing number of high-quality pseudo labels are produced, allowing the model to gradually absorb meaningful information from unlabeled frames and improve its segmentation performance.

Discussion on phase-1 training. In phase-1 training, we restrict the reference (first) frame to be a labeled frame, as the quality of subsequent predictions heavily depends on the accuracy of the reference mask. Using an unlabeled frame with a pseudo label as the reference at this stage would amplify error propagation during early training.

To fully leverage the unlabeled data, we introduce phase-2 training, which removes this constraint and allows the reference frame to be either labeled or pseudo-labeled. The key idea is to first use the reasonably strong VOS model obtained from phase-1 to generate pseudo labels for all unlabeled frames. These pseudo-labeled frames are then organized into a pseudo-label bank, enabling efficient retrieval when constructing training triplets in which the reference frame may be a pseudo-labeled one.

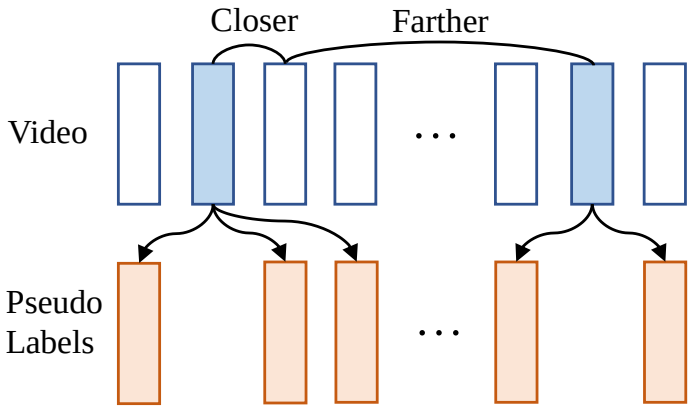


FIGURE 4.5. Illustration of bidirectional inference. Two reference frames are denoted by blue rectangles. A pre-trained VOS model infers unlabeled frames from the inference frame to the end frame and, in a reverse manner, from the inference frame to the beginning frame. We pick the prediction inferred by the labeled frame that is closest to the unlabeled frame.

Intermediate inference and pseudo-label bank. Before commencing phase-2 training, we conduct an intermediate inference step. Since VOS inference requires the annotation of a reference (first) frame, a challenge arises in the two-shot setting where only two labeled frames are available. To generate reliable pseudo labels for all unlabeled frames, we adopt a bidirectional inference strategy inspired by (Lee et al., 2022; Miao et al., 2021), as illustrated in Figure 4.5.

Concretely, for each of the two labeled frames, the phase-1 VOS model uses that frame as the reference to propagate predictions forward, from the reference frame to the end of the video, and backward, from the reference frame to the beginning. Consequently, each unlabeled frame obtains two predicted masks, one from each labeled reference frame. For every unlabeled frame, we select the prediction generated by the labeled frame that is temporally closest to it. All resulting pseudo labels are then stored in a pseudo-label bank for efficient access during phase-2 training.

Phase-2 training. The training procedure in phase-2 mirrors that of phase-1, with the key difference that the reference (first) frame is no longer restricted to labeled frames; it may instead be an unlabeled frame equipped with a pseudo label retrieved from the pseudo-label bank.

Pseudo-label bank update. As training progresses, the model's predictions gradually improve, yielding increasingly reliable pseudo labels. To further enhance phase-2 training, we dynamically update the pseudo-label bank throughout the process. Specifically, at each iteration, given the prediction $\boldsymbol{P}$ for an unlabeled frame, let $\boldsymbol{P}^{(i,j)}$ denote the predicted probability vector at pixel (i, j). Whenever a pixel satisfies $\max\Big(\boldsymbol{P}^{(i,j)}\Big) \geq \tau_2$, where τ_2 is a predefined confidence threshold, we update its pseudo label in the pseudo-label bank using

$$\hat{\boldsymbol{Y}}^{(i,j)} = \arg\max\Big(\boldsymbol{P}^{(i,j)}\Big).$$

We set $\tau_2 = 0.99$ by default to ensure that only highly confident predictions are used for bank updates.

4.3.4 One-Shot VOS Training

Intermediate inference of one-shot VOS training. Recall that two-shot VOS training consists of three stages: phase-1 training, intermediate inference, and phase-2 training. To effectively tackle the more challenging one-shot setting, we incorporate SAM (Kirillov et al., 2023c) into the intermediate inference stage. In particular, we introduce: (1) a lightweight SAM fine-tuning scheme with a point-prompt augmentation strategy tailored to VOS, and (2) a mask assessment module that selects the higher-quality mask between the one-shot VOS model and the fine-tuned SAM model. The collaboration between these two models substantially improves pseudo-label quality, as illustrated in Figure 4.6. Both phase-1 and phase-2 training follow the same procedures as in the two-shot setting, with the only difference being that each training video contains a single labeled frame in the one-shot scenario.

SAM fine-tuning, point sampling, and point-prompt augmentation. The SAM model (Kirillov et al., 2023c) has ushered in a new era of image segmentation. It can segment foreground objects given a point prompt, which consists of a set of reference points located inside the target object. More accurate point prompts generally yield higher-quality masks. In our framework, the initial point prompt for each unlabeled frame is extracted from the mask predicted by the phase-1 VOS model (see Figure 4.6). However, these predicted masks may contain noise, causing the extracted reference points to deviate from the actual ground-truth

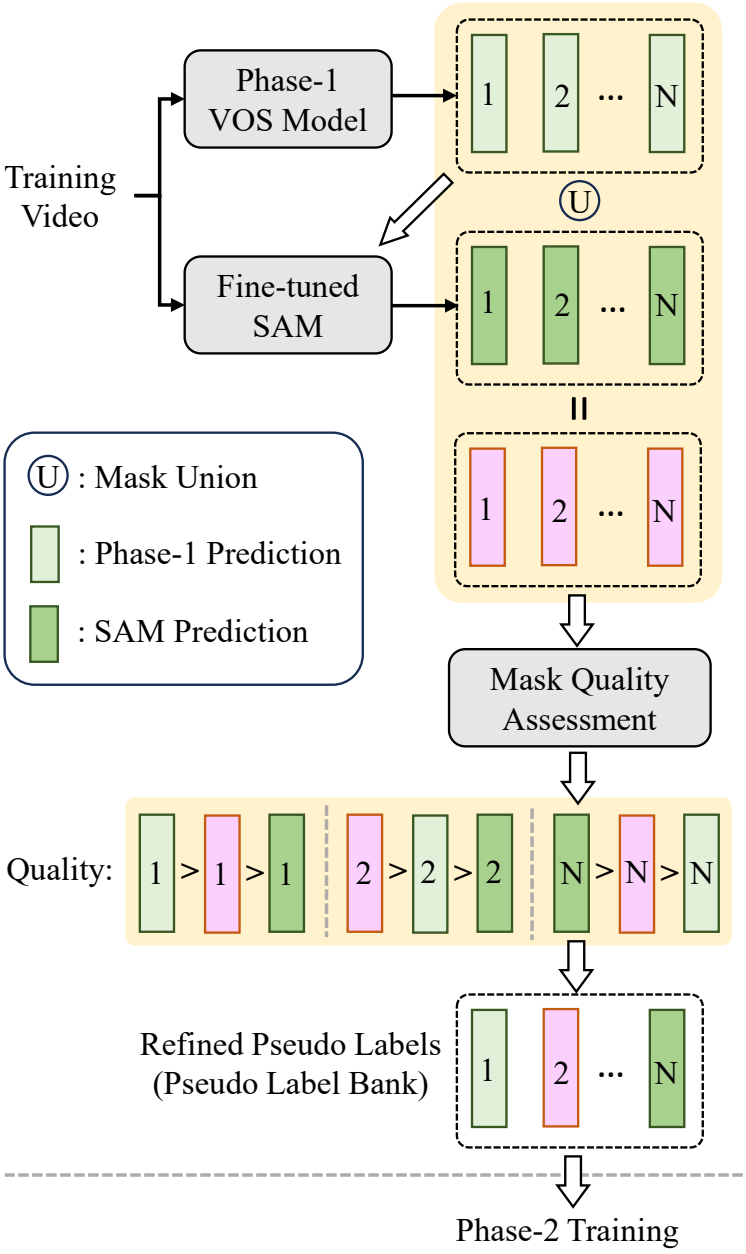


FIGURE 4.6. Overview of the intermediate inference stage of the one-shot VOS training. We fine-tune a SAM model with a point-prompt augmentation strategy. A mask quality assessment module is proposed to select the best mask for each frame. This selection is made from the predictions of the phase-1 model, the output of the fine-tuned SAM model, and the combined mask derived from both the phase-1 and SAM models. See Figure 4.7 for SAM fine-tuning and Figure 4.8 for mask quality assessment.

object. To address this challenge and to better adapt SAM to the one-shot VOS scenario, where only a single labeled frame is available per training video, we propose a lightweight

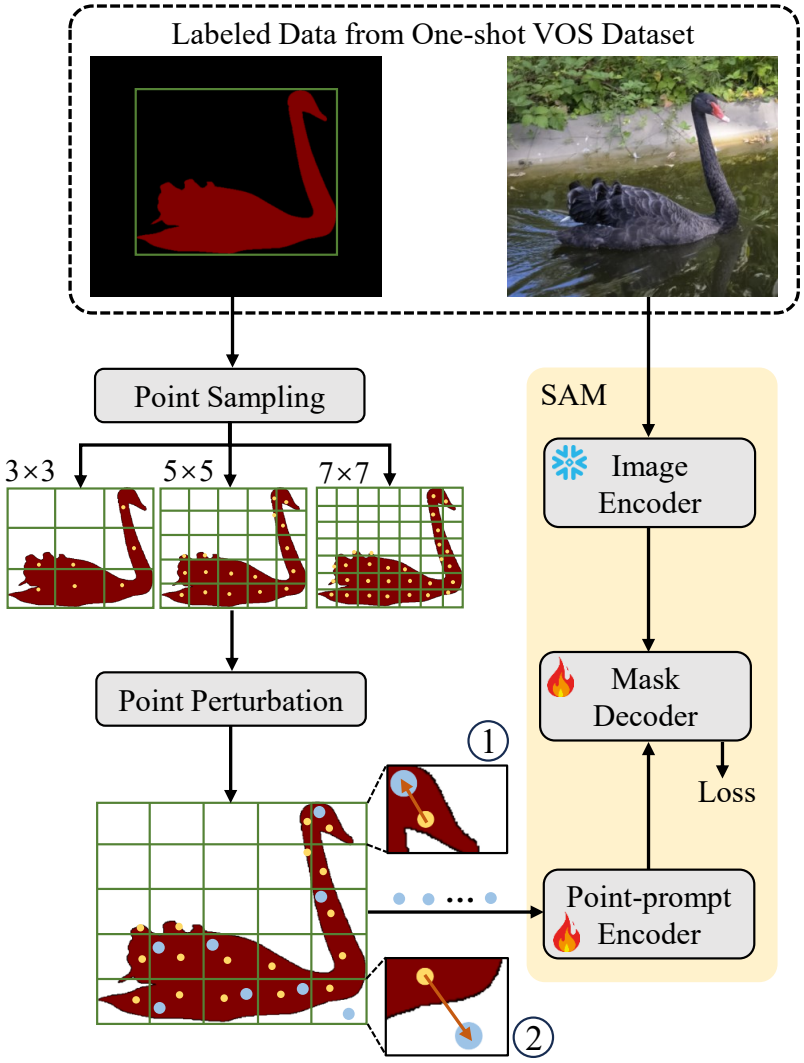


FIGURE 4.7. Illustration of the SAM fine-tuning. Given a labeled frame from our one-shot VOS dataset, we perform a point sampling to yield the corresponding point prompt, which is essentially a set of reference points. Subsequently, each of these reference points undergoes a point perturbation process to enhance the variety of the point prompts. Only the mask decoder and the point-prompt encoder of the SAM model are fine-tuned. Notably, the perturbation point could either be within the foreground area (①) or outside (②) it.

fine-tuning strategy coupled with a point-prompt augmentation mechanism, as illustrated in Figure 4.7.

Given a labeled frame, we first compute the bounding rectangle of the ground-truth mask and uniformly divide it into grids of sizes $\{D \times D\}_{D=3}^{8}$. For each grid, we perform point

sampling: within each cell, we collect all foreground pixels, sort them by their x-coordinate and then their y-coordinate, and select the median pixel as the reference point. Cells without foreground pixels do not contribute reference points. Among all candidate grids, we select the one whose number of reference points is closest to M (default $M = 16$), ensuring that the foreground object is represented by approximately M reference points. For simplicity, we continue to denote the actual number of selected points by M.

Using the extracted reference points directly as SAM point prompts is the most straightforward approach for fine-tuning. During pseudo-label refinement, however, the initial point prompts derived from noisy phase-1 predictions may be inaccurate. To mitigate this issue, we introduce *point-prompt augmentation*, which applies small perturbations ("jitters") to each reference point. Let $\{(x_i, y_i)\}_{i=1}^{M}$ be the original point prompt. For each reference point (x_i, y_i), we generate a perturbed point $(\bar{x}_i, \bar{y}_i)$ via:

$$\bar{x}_i = x_i + D \cdot L, \tag{4.3}$$

$$\bar{y}_i = y_i + D \cdot L, \tag{4.4}$$

where $D \in \{-1, 0, 1\}$ specifies the perturbation direction and $L \in [0, 20]$ controls the perturbation magnitude.

Given the augmented point prompt $\{(\bar{x}_i, \bar{y}_i)\}_{i=1}^{M}$ and the raw image, we fine-tune SAM using the supervised loss in Eq. 4.1 together with the Dice loss (Milletari et al., 2016). Following prior work, we fine-tune only the mask decoder and the point-prompt encoder, as shown in Figure 4.7. The resulting fine-tuned SAM model provides improved pseudo masks for unlabeled frames during intermediate inference.

Mask quality assessment. As illustrated in Figure 4.6, we employ a mask quality assessment module, trained with a rank loss, to select the optimal pseudo mask for each unlabeled frame. The module evaluates three candidate masks: (1) the prediction from the phase-1 VOS model, denoted by $\hat{\boldsymbol{Y}}$; (2) the prediction from the fine-tuned SAM model, denoted by $\boldsymbol{Y}_S$; and (3) the mask union $\boldsymbol{Y}_U$, obtained by merging $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}_S$.

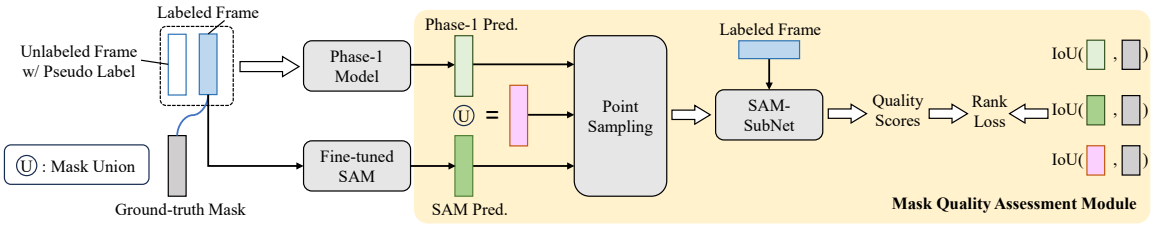


FIGURE 4.8. Illustration of the training process of the mask quality assessment module. The module is trained on our one-shot dataset. We use margin ranking loss as the objective function.

The mask quality assessment module is trained using all labeled frames from the one-shot benchmark, as illustrated in Figure 4.8. Given a labeled frame $\boldsymbol{V}_L$ with its ground-truth mask $\boldsymbol{Y}$, we compute the Intersection over Union (IoU) between $\boldsymbol{Y}$ and each candidate mask in $\{\hat{\boldsymbol{Y}}, \boldsymbol{Y}_S, \boldsymbol{Y}_U\}$. Based on these IoU scores, we derive a rank R indicating the relative quality among the three candidate masks.

To obtain $\hat{\boldsymbol{Y}}$, we randomly select one of the three preceding unlabeled frames and use its pseudo label as the reference (first) frame. This frame is processed by the phase-1 VOS model, producing the prediction $\hat{\boldsymbol{Y}}$. To generate $\boldsymbol{Y}_S$, we directly feed $\boldsymbol{V}_L$ and an initial point prompt into the fine-tuned SAM model. The point prompt is constructed from $\hat{\boldsymbol{Y}}$ using the reference-point sampling strategy shown in Figure 4.7. The union mask $\boldsymbol{Y}_U$ is then obtained by combining $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}_S$.

The assessment module is derived from the SAM architecture. It consists of the SAM image encoder, the SAM point-prompt encoder, the first part of the SAM mask decoder up to the image-to-token-attention (I2TA) layer (Kirillov et al., 2023c), and an additional two-layer MLP followed by a sigmoid activation. The SAM-related components are initialized from our fine-tuned SAM model, while only the newly added MLP is trained.

As shown in Figure 4.8, the labeled frame $\boldsymbol{V}_L$, the three candidate masks $\{\hat{\boldsymbol{Y}}, \boldsymbol{Y}_S, \boldsymbol{Y}_U\}$, and the ground-truth rank R are fed into the module. The final MLP produces quality scores $\{\hat{s}, s_S, s_U\}$ corresponding to $\{\hat{\boldsymbol{Y}}, \boldsymbol{Y}_S, \boldsymbol{Y}_U\}$. We optimize the module using a margin ranking loss (MRL) to encourage consistency with the ground-truth ranking:

$$\mathcal{L} = \mathrm{MRL}(\hat{s}, s_S) + \mathrm{MRL}(\hat{s}, s_U) + \mathrm{MRL}(s_S, s_U), \tag{4.5}$$

where MRL is defined as

$$\mathrm{MRL}(i,j) = \max\left(0, -\Omega_{[i,j]}(i-j) + \beta\right), \tag{4.6}$$

and $\Omega_{[i,j]} = 1$ if $i \geq j$ and -1 otherwise. The margin hyper-parameter is set to $\beta = 0.005$ by default.

4.3.5 Generalization Capability

Our low-shot training paradigm can be seamlessly integrated with a wide range of VOS models, regardless of their architectural designs or input requirements. To maintain generality, we employ the phase-1 STCN model to construct a pseudo-label bank, after which any VOS model can apply the unified phase-2 training strategy to support low-shot learning. To empirically validate the generalization ability of our approach, we adopt it not only for STCN (Cheng et al., 2021c) but also for RDE-VOS (Li et al., 2022c) and XMem (Cheng and Schwing, 2022), demonstrating consistent improvements across models.

4.4 Experiment

4.4.1 Experimental Setup

We conduct experiments on six widely used VOS benchmarks: DAVIS 2016/2017 (Perazzi et al., 2016; Pont-Tuset et al., 2017), YouTube-VOS 2018/2019 (Xu et al., 2018), LVOS (Hong et al., 2023), and VOST (Tokmakov et al., 2023).

DAVIS 2016/2017. DAVIS 2017 extends DAVIS 2016 to the multi-object setting. It contains 60 training videos with 138 annotated objects and 30 validation videos with 59 objects.

YouTube-VOS 2018/2019. YouTube-VOS is a large-scale multi-object benchmark consisting of 3,471 videos from 65 categories in the training set. Annotations are provided every five frames. The 2018 and 2019 validation splits consist of 474 and 507 videos, respectively.

LVOS. LVOS is a long-term video segmentation dataset comprising 220 videos with a total duration of 421 minutes (averaging 1.59 minutes and 574 frames per video at 6 FPS). It spans 27 categories and is divided into 120 training videos, 50 validation videos, and 50 testing videos. We follow two evaluation protocols: (1) *Without fine-tuning*, where models are trained on YouTube-VOS 2019 and DAVIS 2017 and directly evaluated on LVOS; and (2) *With fine-tuning*, where models are further fine-tuned on the LVOS training set prior to evaluation.

VOST. VOST focuses on object transformations where appearance-based cues are unreliable, requiring strong spatio-temporal modeling. It contains 713 clips covering 51 transformation types and 155 object categories, with over 175,000 annotated masks and an average video duration of 21.2 seconds. The dataset is split into 572 training videos, 70 validation videos, and 71 testing videos. We follow the official evaluation protocol.

Low-Shot Settings. In the *two-shot* setting, we randomly sample two labeled frames per video, treating all remaining frames as unlabeled. Relative to full supervision, this corresponds to using only 7.3% of labeled data on YouTube-VOS, 2.9% on DAVIS, 0.37% on LVOS, and 1.91% on VOST.

In the more challenging *one-shot* setting, we sample only one labeled frame per video. This further reduces the proportion of labeled data to 3.7% (YouTube-VOS), 1.4% (DAVIS), 0.18% (LVOS), and 0.95% (VOST).

Evaluation Metrics. Following common practice (Oh et al., 2019; Cheng et al., 2021c; Cheng and Schwing, 2022), for the DAVIS benchmarks, we report region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$. For the YouTube-VOS datasets, we evaluate $\mathcal{J}$ and $\mathcal{F}$ on both seen and unseen categories, as well as their averaged score $\mathcal{G}$. For LVOS, we report $\mathcal{J}\&\mathcal{F}$, along with $\mathcal{J}$ and $\mathcal{F}$ individually. For VOST, we provide $\mathcal{J}_{tr}$ and $\mathcal{J}$ following the official protocol.

Implementation Details. Our method is implemented in PyTorch (Paszke et al., 2017). Pretraining on static image datasets is a common practice in video object segmentation. Following prior works (Cheng et al., 2021c; Cheng and Schwing, 2022; Oh et al., 2019;

Yang et al., 2021d), we first pre-train the STCN (Cheng et al., 2021c) model on static image datasets (Wang et al., 2017; Shi et al., 2015; Zeng et al., 2019) with synthetic augmentations before phase-1 training. For consistency and fairness, all models considered in this study, including STCN (Cheng et al., 2021c), XMem (Cheng and Schwing, 2022), and RDE-VOS (Li et al., 2022c), are pretrained on the same static image datasets for the one-shot, two-shot, and full-set settings.

During training, the parameter K for random frame skipping is gradually increased from 5 to 25 using a curriculum strategy. We set the confidence thresholds to $\tau_1 = 0.9$ and $\tau_2 = 0.99$. Our low-shot VOS paradigm is model-agnostic and can be applied to any architecture in the phase-2 training stage. We validate this generality using STCN (Cheng et al., 2021c), RDE-VOS (Li et al., 2022c), and XMem (Cheng and Schwing, 2022).

For SAM, we adopt the default ViT-H model with the `multimask_output`[1] parameter set to `False`. We fine-tune the SAM model for 200 epochs using the Adam optimizer with an initial learning rate of 0.001.

The mask quality assessment module is trained for 200 epochs with an Adam optimizer and a cosine annealing learning rate schedule, starting from 0.001 and decaying to 0.0001.

4.4.2 Main Results

Results on YouTube-VOS and DAVIS. We evaluate our low-shot training paradigm on STCN (Cheng et al., 2021c), RDE-VOS (Li et al., 2022c), and XMem (Cheng and Schwing, 2022), and compare the results against: (1) their fully supervised counterparts trained on the full datasets; (2) naive low-shot versions trained without using unlabeled data (i.e., repeatedly sampling the few labeled frames to meet the model's input requirements); and (3) other strong fully supervised models. The validation results on YouTube-VOS and DAVIS are reported in Tables 4.1 and 4.2.

[1]The `multimask_output` option addresses ambiguity in single-point prompts by producing three candidate masks representing whole, part, and subpart regions of the object. Since our approach uses multiple point prompts, this ambiguity is significantly reduced. We therefore set `multimask_output` to "False" to obtain a single mask prediction.

Method	Labeled data	YouTube-VOS 2018					YouTube-VOS 2019				
		$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
STM (Oh et al., 2019)	100%	79.4	79.7	84.2	72.8	80.9	-	-	-	-	-
MiVOS (Cheng et al., 2021b)	100%	80.4	80.0	84.6	74.8	82.4	80.3	79.3	83.7	75.3	82.8
CFBI (Yang et al., 2020)	100%	81.4	81.1	85.8	75.3	83.4	81.0	80.6	85.1	75.2	83.0
RDE-VOS (Li et al., 2022c)	100%	-	-	-	-	-	81.9	81.1	85.5	76.2	84.8
HMMN (Seong et al., 2021)	100%	82.6	82.1	87.0	76.8	84.6	82.5	81.7	86.1	77.3	85.0
JOINT (Mao et al., 2021)	100%	83.1	81.5	85.9	78.7	86.5	82.7	81.1	85.4	78.2	85.9
STCN (Cheng et al., 2021c)	100%	83.0	81.9	86.5	77.9	85.7	82.7	81.1	85.4	78.2	85.9
R50-AOT-L (Yang et al., 2021d)	100%	84.1	83.7	88.5	78.1	86.1	84.1	83.5	88.1	78.4	86.3
XMem (Cheng and Schwing, 2022)	100%	85.7	84.6	89.3	80.2	88.7	85.5	84.3	88.6	80.3	88.6
STCN* (Cheng et al., 2021c)	100%	83.0	82.0	86.5	77.8	85.8	82.7	81.2	85.4	78.2	86.0
2-shot STCN* (Cheng et al., 2021c)	**7.3**%	80.8	79.5	83.9	75.9	84.0	80.6	79.5	83.8	75.6	83.4
2-shot STCN w/ Ours	**7.3**%	82.9$_{+2.1}$	81.6$_{+2.1}$	86.3$_{+2.4}$	77.7$_{+1.8}$	86.0$_{+2.0}$	82.7$_{+2.1}$	80.9$_{+1.4}$	85.1$_{+1.3}$	78.3$_{+2.7}$	86.6$_{+3.2}$
1-shot STCN* (Cheng et al., 2021c)	**3.7**%	72.9	71.7	74.4	68.7	76.9	72.9	71.3	73.6	69.5	77.3
1-shot STCN w/ Ours	**3.7**%	81.9$_{+9.0}$	81.0$_{+9.3}$	85.7$_{+11.3}$	76.2$_{+7.5}$	84.7$_{+7.8}$	81.8$_{+8.9}$	80.4$_{+9.1}$	84.7$_{+11.1}$	76.8$_{+7.3}$	84.9$_{+7.6}$
RDE-VOS* (Li et al., 2022c)	100%	-	-	-	-	-	82.1	81.3	85.7	76.2	85.0
2-shot RDE-VOS* (Li et al., 2022c)	**7.3**%	-	-	-	-	-	78.4	77.2	81.3	73.4	81.7
2-shot RDE-VOS w/ Ours	**7.3**%	-	-	-	-	-	82.1$_{+3.7}$	80.4$_{+3.2}$	84.8$_{+3.5}$	77.3$_{+3.9}$	85.8$_{+4.1}$
1-shot RDE-VOS* (Li et al., 2022c)	**3.7**%	-	-	-	-	-	72.6	70.9	74.5	68.6	76.6
1-shot RDE-VOS w/ Ours	**3.7**%	-	-	-	-	-	81.4$_{+8.8}$	79.6$_{+8.7}$	83.8$_{+9.3}$	77.2$_{+8.6}$	85.7$_{+9.1}$
XMem* (Cheng and Schwing, 2022)	100%	85.5	84.4	89.1	80.0	88.3	85.3	84.0	88.2	80.4	88.4
2-shot XMem* (Cheng and Schwing, 2022)	**7.3**%	79.2	77.5	81.9	74.5	82.9	79.1	77.6	81.5	74.5	82.7
2-shot XMem w/ Ours	**7.3**%	84.8$_{+5.6}$	83.6$_{+6.1}$	88.5$_{+6.6}$	79.2$_{+4.7}$	87.7$_{+4.8}$	84.5$_{+5.4}$	83.5$_{+5.9}$	88.0$_{+6.5}$	79.1$_{+4.6}$	87.3$_{+4.6}$
1-shot XMem* (Cheng and Schwing, 2022)	**3.7**%	70.7	69.4	72.5	65.7	75.3	70.2	67.7	69.6	67.5	75.9
1-shot XMem w/ Ours	**3.7**%	83.6$_{+12.9}$	82.6$_{+13.2}$	87.1$_{+14.6}$	77.9$_{+12.2}$	86.6$_{+11.3}$	83.5$_{+13.3}$	83.0$_{+15.3}$	87.5$_{+17.9}$	77.6$_{+10.1}$	85.9$_{+10.0}$

TABLE 4.1. Comparison of various methods on YouTube-VOS 2018 and 2019 validation sets. The subscripts S and U denote seen and unseen categories respectively. The symbol * indicates results that are reproduced using open-source code. With only 7.3% of labeled data (equivalent to 2 labeled frames for each training video) from the YouTube-VOS benchmark, our method performs on par with its counterpart that is trained on the entire dataset. When utilizing just 3.7% labeled data (or 1 labeled frame per training video), VOS models that incorporate our training methodology significantly surpass their 1-shot counterparts by a substantial margin.

Three observations emerge: (1) Even with only two labeled frames per video, naive 2-shot models already perform competitively. For example, the 2-shot STCN reaches 80.8% on YouTube-VOS 2018, only -2.2% below the full-set model (83.0%). (2) In the 2-shot setting, using merely 7.3% (YouTube-VOS) and 2.9% (DAVIS) of the labeled data, our method yields performance very close to the full-set models while markedly outperforming naive low-shot variants. For instance, 2-shot STCN with our approach achieves 85.1% / 82.7% on DAVIS 2017 / YouTube-VOS 2019, improving over the naive 2-shot STCN by +4.1% / +2.1%, and falling only -0.1% / -0.0% short of the full-set STCN. (3) In the more challenging 1-shot scenario (3.7% and 1.4% labeled data on YouTube-VOS and DAVIS), our training strategy leads to substantial improvements over naive 1-shot models. For example, naive 1-shot XMem

Method	Labeled data	DAVIS 2016			DAVIS 2017		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
STM (Oh et al., 2019)	100%	89.3	88.7	89.9	81.8	78.2	84.3
CFBI (Yang et al., 2020)	100%	89.4	88.3	90.5	81.9	79.1	84.6
JOINT (Mao et al., 2021)	100%	-	-	-	83.5	80.8	86.2
RDE-VOS (Li et al., 2022c)	100%	91.1	89.7	92.5	84.2	80.8	87.5
MiVOS (Cheng et al., 2021b)	100%	91.0	89.6	92.4	84.5	81.7	87.4
HMMN (Seong et al., 2021)	100%	90.8	89.6	92.0	84.7	81.9	87.5
R50-AOT-L (Yang et al., 2021d)	100%	91.1	90.1	92.1	84.9	82.3	87.5
STCN (Cheng et al., 2021c)	100%	91.6	90.8	92.5	85.4	82.2	88.6
XMem (Cheng and Schwing, 2022)	100%	91.5	90.4	92.7	86.2	82.9	89.5
STCN* (Cheng et al., 2021c)	100%	91.5	90.7	92.3	85.2	81.9	88.5
2-shot STCN* (Cheng et al., 2021c)	**2.9**%	87.9	87.1	88.7	81.0	77.7	84.3
2-shot STCN w/ Ours	**2.9**%	$91.3_{+3.4}$	$90.6_{+3.5}$	$92.0_{+3.3}$	$85.1_{+4.1}$	$81.7_{+4.0}$	$88.4_{+4.1}$
1-shot STCN* (Cheng et al., 2021c)	**1.4**%	76.6	76.3	77.0	68.0	65.4	70.6
1-shot STCN w/ Ours	**1.4**%	$90.7_{+14.1}$	$89.9_{+13.6}$	$91.5_{+14.5}$	$83.9_{+15.9}$	$80.8_{+15.4}$	$86.9_{+16.3}$
RDE-VOS* (Li et al., 2022c)	100%	91.0	89.5	92.4	84.2	80.7	87.7
2-shot RDE-VOS* (Li et al., 2022c)	**2.9**%	87.6	86.6	88.8	79.4	75.6	83.1
2-shot RDE-VOS w/ Ours	**2.9**%	$90.8_{+3.2}$	$90.0_{+3.4}$	$92.0_{+3.2}$	$83.9_{+4.5}$	$80.4_{+4.8}$	$87.3_{+4.2}$
1-shot RDE-VOS* (Li et al., 2022c)	**1.4**%	78.7	77.7	78.6	68.9	65.2	72.6
1-shot RDE-VOS w/ Ours	**1.4**%	$89.7_{+11.0}$	$88.5_{+10.8}$	$91.0_{+12.4}$	$82.8_{+13.9}$	$79.7_{+14.5}$	$85.9_{+13.3}$
XMem* (Cheng and Schwing, 2022)	100%	91.3	90.3	92.4	86.2	82.8	89.7
2-shot XMem* (Cheng and Schwing, 2022)	**2.9**%	88.1	87.1	89.0	81.7	78.2	85.1
2-shot XMem w/ Ours	**2.9**%	$91.3_{+3.2}$	$90.3_{+3.2}$	$92.3_{+3.3}$	$85.6_{+3.9}$	$82.1_{+3.9}$	$89.1_{+4.0}$
1-shot XMem* (Cheng and Schwing, 2022)	**1.4**%	75.8	75.4	76.2	66.7	63.3	70.2
1-shot XMem w/ Ours	**1.4**%	$90.5_{+14.7}$	$89.7_{+14.3}$	$91.4_{+15.2}$	$84.1_{+17.4}$	$80.8_{+17.5}$	$87.4_{+17.2}$

TABLE 4.2. Comparison of various methods on DAVIS 2016 and 2017 validation sets. The symbol * indicates results that are reproduced using open-source code.

scores only 70.2% on YouTube-VOS 2019 and 66.7% on DAVIS 2018, whereas our method boosts these results to 83.5% and 84.1%, respectively.

Results on LVOS. We train STCN, RDE-VOS, and XMem using our 2-shot and 1-shot methods under both LVOS evaluation protocols and compare them with their full-set versions in Table 4.3. Key findings include: (1) Under the *without fine-tuning* setting, our 2-shot models generalize well to LVOS, with only modest drops of -0.8, -0.9, and -1.2 in $\mathcal{J}\&\mathcal{F}$ for STCN, RDE-VOS, and XMem, respectively. (2) Under the *with fine-tuning* setting, 2-shot models remain close to the full-set models, declining by only -2.0, -0.7, and -1.8 in $\mathcal{J}\&\mathcal{F}$ while using just 0.37% of the labeled data. (3) Across all settings (one-shot/two-shot,

Method	Labeled data	Without fine-tuning			With fine-tuning		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CFBI (Yang et al., 2020)	100%	50.0	45.0	55.1	51.5	46.2	56.7
LWL (Bhat et al., 2020)	100%	54.1	49.6	58.6	56.4	51.8	60.9
AOT-B (Yang et al., 2021d)	100%	56.9	51.8	61.9	58.9	53.5	64.2
AOT-L (Yang et al., 2021d)	100%	59.4	53.6	65.2	60.9	55.1	66.8
AFN-URR (Liang et al., 2020)	100%	34.8	31.3	38.2	36.2	33.1	39.3
STCN (Cheng et al., 2021c)	100%	45.8	41.1	50.5	48.9	43.9	54.0
RDE-VOS (Li et al., 2022c)	100%	52.9	47.7	58.1	53.7	48.3	59.2
XMem (Cheng and Schwing, 2022)	100%	50.0	45.5	54.4	52.9	48.1	57.7
STCN* (Cheng et al., 2021c)	100%	45.8	41.2	50.3	48.8	43.8	53.8
2-shot STCN* (Cheng et al., 2021c)	**0.37**%	32.6	28.7	36.4	35.1	31.9	38.3
2-shot STCN w/ Ours	**0.37**%	$45.0_{+12.4}$	$39.6_{+10.9}$	$50.4_{+14.0}$	$46.8_{+11.7}$	$41.9_{+10.0}$	$51.6_{+13.3}$
1-shot STCN* (Cheng et al., 2021c)	**0.18**%	24.3	21.8	26.8	26.9	24.5	29.2
1-shot STCN w/ Ours	**0.18**%	$43.3_{+19.0}$	$38.2_{+16.4}$	$48.5_{+21.7}$	$44.6_{+17.7}$	$40.6_{+16.1}$	$48.6_{+19.4}$
RDE-VOS* (Li et al., 2022c)	100%	53.0	47.6	58.4	53.5	49.3	51.3
2-shot RDE-VOS* (Li et al., 2022c)	**0.37**%	31.3	27.5	35.0	34.0	30.2	37.7
2-shot RDE-VOS w/ Ours	**0.37**%	$52.1_{+20.8}$	$48.1_{+20.6}$	$56.1_{+21.1}$	$52.8_{+18.8}$	$48.4_{+18.2}$	$57.1_{+19.4}$
1-shot RDE-VOS* (Li et al., 2022c)	**0.18**%	23.5	19.1	27.9	26.2	21.2	31.2
1-shot RDE-VOS w/ Ours	**0.18**%	$50.7_{+27.2}$	$45.5_{+26.4}$	$55.8_{+27.9}$	$51.3_{+25.1}$	$47.1_{+25.9}$	$55.6_{+24.4}$
XMem* (Cheng and Schwing, 2022)	100%	50.1	45.9	54.3	52.8	48.5	57.2
2-shot XMem* (Cheng and Schwing, 2022)	**0.37**%	32.1	27.0	37.2	33.7	28.9	38.6
2-shot XMem w/ Ours	**0.37**%	$48.9_{+16.8}$	$43.6_{+16.6}$	$54.3_{+17.1}$	$51.0_{+17.3}$	$46.5_{+17.6}$	$55.4_{+16.8}$
1-shot XMem* (Cheng and Schwing, 2022)	**0.18**%	21.8	17.0	26.5	23.1	20.5	25.6
1-shot XMem w/ Ours	**0.18**%	$47.3_{+25.5}$	$42.7_{+25.7}$	$51.8_{+25.3}$	$48.2_{+25.1}$	$43.7_{+23.2}$	$52.7_{+27.1}$

TABLE 4.3. Comparison of various methods on the LVOS validation set. The evaluation is performed under two settings: (1) "Without fine-tuning", where models are trained on a combination of YouTube-VOS 2019 and DVIS 2017, and then evaluated directly on LVOS; and (2) "With fine-tuning", where models are initially trained on YouTube-VOS 2019 and DVIS 2017, followed by fine-tuning on the LVOS training set before evaluation. The symbol * indicates results that are reproduced. "Labeled data" indicates the percentage of labeled frames used in the LVOS fine-tuning setting. It is worth noting that, in the "without fine-tuning" setting, our model is trained on the two-shot or one-shot YouTube-VOS 2019+DVIS 2017 datasets, where only two or one frames per video are annotated.

with/without fine-tuning), our low-shot models consistently and significantly outperform their counterparts trained without our method.

Results on VOST. We further evaluate 2-shot and 1-shot models on VOST using only 1.91% and 0.95% of the labeled data, respectively, and compare them with fully trained versions in Table 4.4. Our 2-shot STCN, RDE-VOS, and XMem models achieve $\mathcal{J}_{tr}$ scores of 30.7, 32.6,

Method	Labeled data	$\mathcal{J}_{tr}$	$\mathcal{J}$
OSMN Match (Yang et al., 2018)	100%	7.0	8.7
OSMN Tune (Yang et al., 2018)	100%	17.6	23.0
CRW (Jabri et al., 2020)	100%	13.9	23.7
CFBI (Yang et al., 2020)	100%	32.0	45.0
CFBI+ (Yang et al., 2021e)	100%	32.6	46.0
AOT (Yang et al., 2021d)	100%	36.4	48.7
XMem (Cheng and Schwing, 2022)	100%	33.8	44.1
HODOR Img (Athar et al., 2022)	100%	13.9	24.2
HODOR Vid (Athar et al., 2022)	100%	25.4	37.1
STCN* (Cheng et al., 2021c)	100%	31.6	42.6
2-shot STCN* (Cheng et al., 2021c)	**1.91**%	27.5	39.5
2-shot STCN w/ Ours	**1.91**%	30.7$_{+3.2}$	42.0$_{+2.5}$
1-shot STCN* (Cheng et al., 2021c)	**0.95**%	15.2	27.2
1-shot STCN w/ Ours	**0.95**%	29.0$_{+13.8}$	41.2$_{+14.0}$
RDE-VOS* (Li et al., 2022c)	100%	33.4	43.2
2-shot RDE-VOS* (Li et al., 2022c)	**1.91**%	26.0	38.7
2-shot RDE-VOS w/ Ours	**1.91**%	32.6$_{+6.6}$	42.6$_{+3.9}$
1-shot RDE-VOS* (Li et al., 2022c)	**0.95**%	14.0	24.3
1-shot RDE-VOS w/ Ours	**0.95**%	30.5$_{+16.5}$	41.4$_{+17.1}$
XMem* (Cheng and Schwing, 2022)	100%	33.9	44.8
2-shot XMem* (Cheng and Schwing, 2022)	**1.91**%	24.9	38.1
2-shot XMem w/ Ours	**1.91**%	33.1$_{+8.2}$	43.2$_{+5.1}$
1-shot XMem* (Cheng and Schwing, 2022)	**0.95**%	14.1	24.8
1-shot XMem w/ Ours	**0.95**%	31.2$_{+17.1}$	42.0$_{+17.2}$

TABLE 4.4. Comparison of various methods on the VOST validation set. The symbol * indicates results that are reproduced.

and 33.1, which are only -0.9, -0.8, and -0.8 below their fully supervised counterparts, despite using less than 2% of the labeled data. Furthermore, both our 1-shot and 2-shot models substantially outperform their naive low-shot baselines across all evaluated metrics.

4.4.3 Ablation Studies for Two-Shot VOS Training

In this section, we systematically verify the effectiveness of our two-shot VOS training method applying to STCN (Cheng et al., 2021c) on Youtube-VOS 2019.

Components	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Baseline	80.6	79.5	83.8	75.7	83.4
+Phase-1	$81.6_{+1.0}$	79.3	83.5	77.7	86.0
+Phase-2	$82.7_{+1.1}$	80.9	85.1	78.3	86.6

TABLE 4.5. Ablation study on the effectiveness of each phase. The naive 2-shot STCN is adopted as the baseline.

Sampling	Naive 2-shot	Phase-1	Phase-2
Random	80.6	81.6	82.7
A: 0%, B: 100%	80.3	80.8	82.2
A: 25%, B: 75%	80.3	81.0	82.3
A: 33%, B: 66%	80.4	81.2	82.4
A: 49%, B: 51%	80.1	80.6	82.0

TABLE 4.6. Ablation study on sampling strategies for labeled data. A% and B% indicate that the two labeled frames are sampled from the first A% and the last B% portions of each video, respectively.

Effects of each phase. The results are presented in Table 4.5. Starting from a naive 2-shot STCN baseline achieving 80.6%, phase-1 training increases the score to 81.6%. Building upon this, phase-2 training further boosts performance to 82.7%, effectively matching the performance of the fully supervised STCN.

Sampling strategies for labeled data. Suppose the two labeled frames are sampled from the first A% and the last B% of each video. Table 4.6 reports the results for different combinations of A and B. We observe that the best performance is achieved when the two labeled frames are selected at random.

Thresholds of pseudo-labeling. The pseudo-labeling procedures in phase-1 and phase-2 are governed by two hyper-parameters, τ_1 and τ_2, respectively. Figure 4.9 illustrates the performance curves obtained by varying these thresholds. Higher thresholds yield pseudo labels of better quality but reduce the number of usable pseudo-labeled samples, and vice

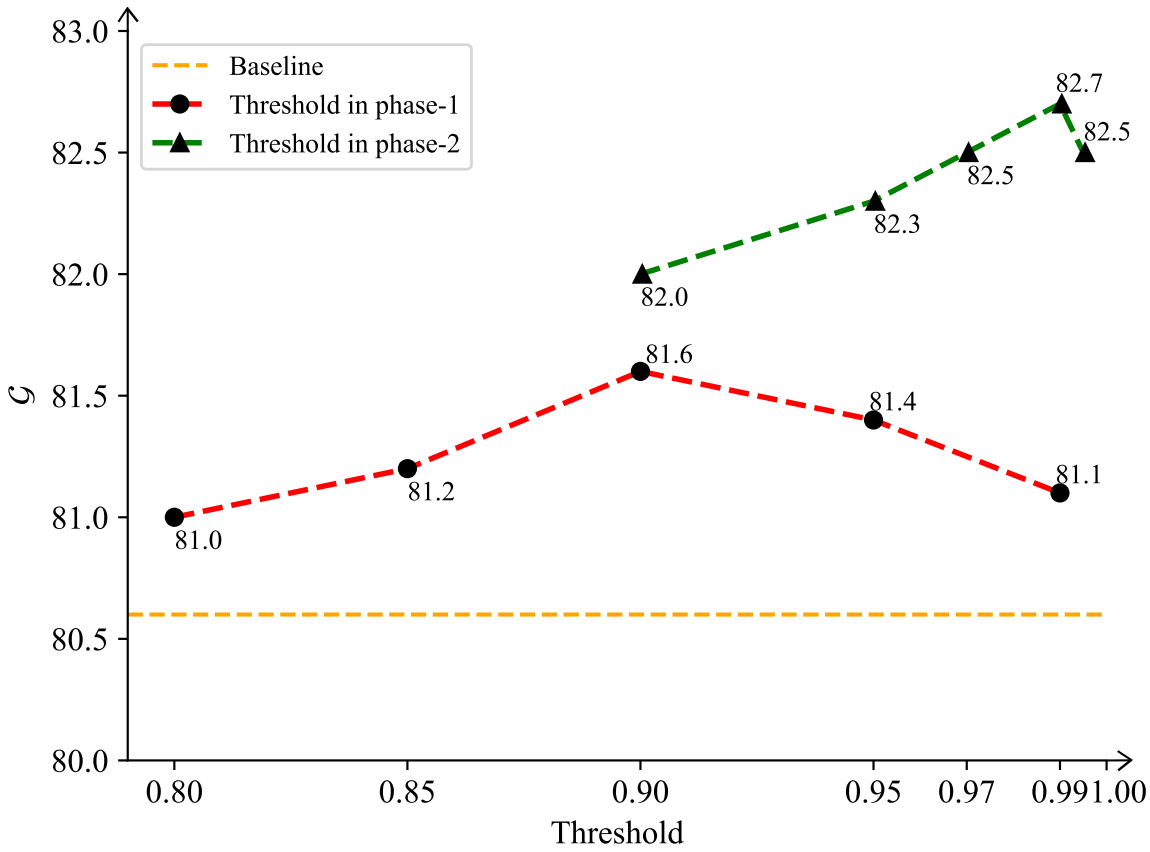


FIGURE 4.9. Study on hyper-parameters τ_1 and τ_2 for phase-1 and phase-2 pseudo-labeling. We adopt a higher threshold in phase-2 training since the predictions in phase-2 are more accurate than that in phase-1. By default, we set $\tau_1 = 0.9$ and $\tau_1 = 0.99$.

versa. Since predictions in phase-2 are generally more reliable, we adopt a higher threshold in this stage. Empirically, the best performance is achieved with $\tau_1 = 0.9$ and $\tau_2 = 0.99$.

Different pseudo labelers. In Table 4.7, we investigate the effect of different pseudo-labelers used during phase-1 training. We evaluate two variants: (1) using the STCN model itself as the pseudo-labeler, and (2) using STCN equipped with a Mean Teacher (MT) framework (Tarvainen and Valpola, 2017). The key idea of the MT approach is to update the teacher model parameters via an exponential moving average (EMA) of the student model parameters at each iteration:

$$\theta'_t = \alpha\theta'_{t-1} + (1-\alpha)\theta_t,$$

Pseudo-labeler	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
-	80.6	79.5	83.8	75.7	83.4
STCN	$81.2_{+0.6}$	79.2	83.5	77.2	84.9
MT-STCN	$81.6_{+0.4}$	79.3	83.5	77.7	86.0

TABLE 4.7. Ablation study of different pseudo-labelers in phase-1. MT-STCN: the parameters of STCN is updated by a Mean Teacher (Tarvainen and Valpola, 2017) strategy.

α	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
0.990	81.2	79.4	83.8	76.9	84.5
0.995	81.6	79.3	83.5	77.7	86.0
0.999	81.3	79.4	83.7	76.9	85.2

TABLE 4.8. Study of different coefficient α used in the MT-STCN, where α denotes the EMA factor.

where t denotes the current iteration, θ'_t and θ_t correspond to the MT-STCN and STCN parameters, respectively, and α is a smoothing coefficient. As shown in Table 4.7, MT-STCN consistently outperforms the non-MT version.

We further analyze the impact of different α values in Table 4.8 and find that $\alpha = 0.995$ yields the strongest performance. However, in phase-2 we do not adopt the MT strategy, as it provides no observable benefit during this stage.

Bidirectional inference. We introduce an intermediate inference stage to construct a pseudo-label bank, which subsequently enables phase-2 training. Our proposed bidirectional inference strategy is compared against the commonly used unidirectional inference adopted in many VOS models. Table 4.9 reports the results. Using bidirectional inference yields a performance gain of +0.6% over unidirectional inference. This improvement can be attributed to two factors: (1) certain unlabeled frames fail to obtain pseudo labels under unidirectional inference; and

Intermediate inference	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Unidirectional	82.1	80.8	77.3	77.6	85.2
Bidirectional	$82.7_{+0.6}$	80.9	85.1	78.3	86.6

TABLE 4.9. Comparison between unidirectional inference and bidirectional inference.

Update	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
	82.2	80.7	84.9	77.6	85.5
✓	$82.7_{+0.5}$	80.9	85.1	78.3	86.6

TABLE 4.10. Study on pseudo-label bank update in phase-2 training. As predictions become more accurate over the course of training, updating the pseudo-label bank enables the model to leverage increasingly reliable pseudo-labels, thereby improving the overall learning process.

(2) bidirectional inference alleviates error propagation by leveraging predictions from both temporal directions.

Dynamically update the pseudo-label bank. We evaluate the effectiveness of dynamically updating the pseudo-label bank during phase-2 training by comparing it with an alternative approach that keeps the pseudo-label bank fixed once constructed. As shown in Table 4.10, using a static pseudo-label bank results in a slight performance drop. This is expected, as predictions become more accurate over the course of training; updating the pseudo-label bank allows the model to benefit from increasingly reliable pseudo labels, thereby improving the learning process.

4.4.4 Ablation Studies for One-Shot VOS Training

In this section, we evaluate the efficacy of our one-shot VOS training approach when applied to STCN (Cheng et al., 2021c), using the Youtube-VOS 2019 dataset.

Components	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Baseline	72.9	71.3	73.6	69.5	77.3
+Phase-1	$77.2_{+4.3}$	75.6	78.7	73.2	81.2
+Quality assessment	$78.9_{+1.7}$	77.1	81.2	74.5	82.9
+Phase-2	$81.8_{+2.9}$	80.4	84.7	76.8	84.9

TABLE 4.11. Ablation study on the effectiveness of each phase. The naive 1-shot STCN is adopted as the baseline.

Effects of each phase. Table 4.11 summarizes the contribution of each phase in the one-shot setting. Beginning with the naive 1-shot STCN baseline, which achieves 72.9%, phase-1 training improves the score to 77.2%. Incorporating the mask quality assessment module yields an additional gain of 1.7%. Finally, after completing phase-2 training, the performance reaches 81.8%, narrowing the gap to the fully supervised STCN to just 0.9%.

Effects of mask quality assessment module. Our one-shot training scheme comprises three stages: (1) phase-1 training; (2) intermediate inference with the mask quality assessment (MQA) module; and (3) phase-2 training. To evaluate the contribution of the MQA module, we compare our full training pipeline with a modified variant that removes only the MQA module while keeping all other components identical. Table 4.12 presents the results of this comparison.

The MQA module selects the highest-quality mask from three candidates: (1) the prediction produced by the phase-1 VOS model; (2) the prediction generated by the fine-tuned SAM model; and (3) the union mask obtained by combining (1) and (2). Table 4.13 reports a comparison between the MQA module and each of these three alternatives.

SAM fine-tuning and point-prompt augmentation. In the one-shot VOS setting, the phase-1 model collaborates with the fine-tuned SAM model to generate pseudo labels for all unlabeled frames, thereby constructing the initial pseudo-label bank. We compare several SAM variants as pseudo-labelers: (1) the original SAM model without fine-tuning; (2) a fine-tuned SAM model without the point-prompt augmentation (PPA) strategy; and (3) our

MQA	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
	80.4	78.5	83.0	76.0	84.2
✓	$81.8_{+1.4}$	80.4	84.7	76.8	84.9

TABLE 4.12. Ablation study on the effectiveness of the mask quality assessment (MQA) module. We report final results after phase-2 training.

	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Phase-1 Model	77.2	75.6	78.7	73.2	81.2
Fine-tuned SAM	76.9	75.5	79.5	71.6	80.8
Union	77.7	77.3	81.0	72.3	80.2
MQA	78.9	77.1	81.2	74.5	82.9

TABLE 4.13. Mask quality assessment (MQA) module outperforms each individual approach. We compare our strategy against three variants: (1) the prediction produced by the phase-1 VOS model; (2) the prediction generated by the fine-tuned SAM model; and (3) the union mask obtained by combining (1) and (2).

proposed SAM variant that incorporates both fine-tuning and PPA. For reference, we also report the performance of the phase-1 VOS model alone. The results are summarized in Table 4.14.

We also investigate an alternative SAM fine-tuning strategy, PerSAM (Zhang et al., 2023c). In PerSAM, the final mask prediction M is computed as a weighted combination of three SAM outputs:

$$M = w_1 \cdot M_1 \ + \ w_2 \cdot M_2 \ + \ (1 - w_1 - w_2) \cdot M_3,$$

where M_1, M_2, and M_3 denote the masks produced by SAM when the `multimask_output` option is enabled. A comparison between our SAM fine-tuning strategy and PerSAM is provided in Table 4.15.

Point sampling strategy for point-prompt generation.

	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Phase-1 Model	77.2	75.6	78.7	73.2	81.2
SAM variant	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
Original SAM	75.9	74.2	77.4	72.5	79.7
+Fine-tuning	78.3$_{+2.4}$	76.8	80.7	73.1	82.6
+PPA	78.9$_{+0.6}$	77.1	81.2	74.5	82.9

TABLE 4.14. Ablation on SAM fine-tuning and point-prompt augmentation (PPA). The "SAM variant" refers to our customized SAM model, which has been fine-tuned from the original SAM model with the proposed point-prompt augmentation (PPA) strategy.

SAM fine-tuning strategy	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
PerSAM (Zhang et al., 2023c)	78.4	77.2	81.3	73.4	81.7
Ours	78.9$_{+0.5}$	77.1	81.2	74.5	82.9

TABLE 4.15. Ablation study on different SAM fine-tuning strategies. We mainly compare our approach with PerSAM (Zhang et al., 2023c).

As shown in Figure 4.7, we propose a point sampling strategy that selects one reference point from each grid cell, producing the point prompt used by the SAM model. We compare this approach with an alternative that randomly selects the same number of reference points, as illustrated in Figure 4.10. Our results show that the proposed sampling strategy consistently outperforms random selection. We also examine the effect of the number of reference points M on performance, and set $M = 16$ by default in our experiments.

4.4.5 Discussion

Pre-training on static image datasets. Pre-training on static image datasets is a widely adopted practice in VOS (Cheng et al., 2021d; Yang et al., 2021d), and it remains beneficial in our low-shot setting. We evaluate models with and without this pre-training strategy, as

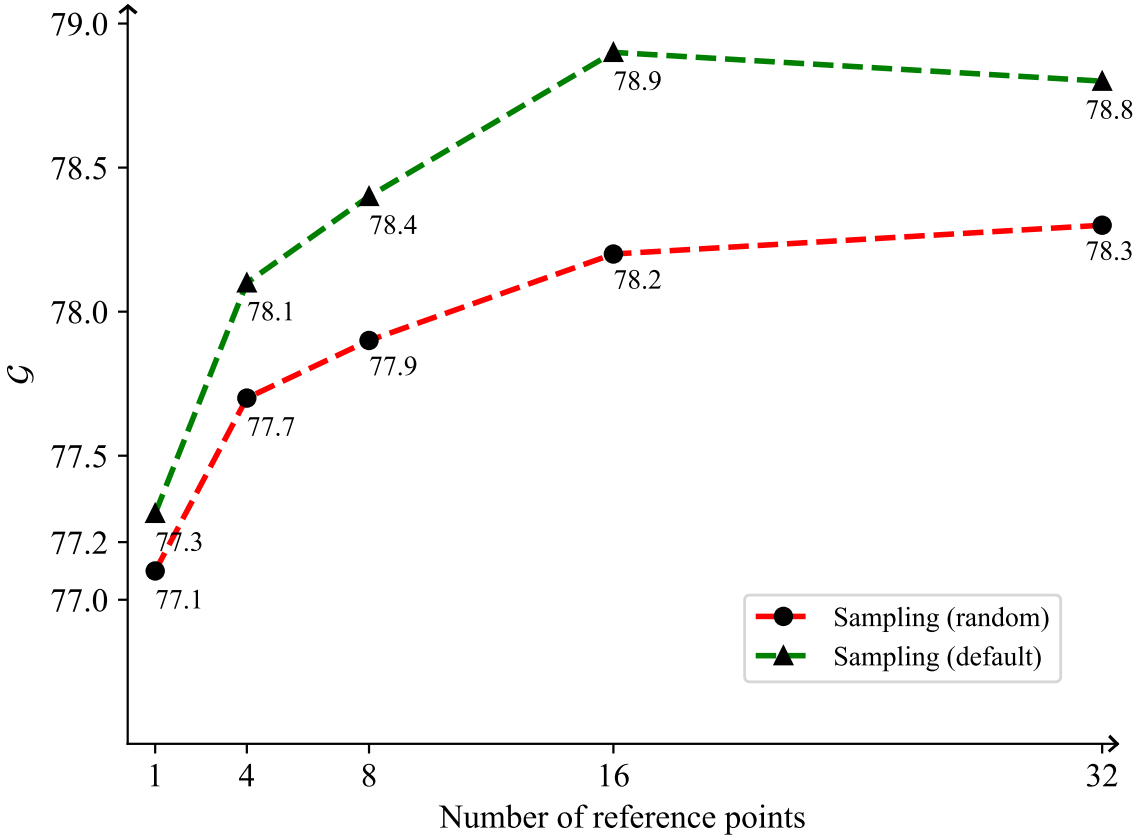


FIGURE 4.10. Improvement of phase-1 using various sampling strategies based on different point numbers. Our uniform sampling yields the highest performance score when the point number is set to 16.

reported in Table 4.16. The results show that both full-set and low-shot models consistently benefit from incorporating static-image pre-training.

Furthermore, Table 4.17 presents a zero-shot transfer experiment, where the STCN model, trained solely on the static pre-training images used in the original STCN, is directly evaluated on the YouTube-VOS 2019 benchmark. Although this zero-shot model exhibits a certain degree of transferability, its performance is still significantly lower than that of the naive one-shot and two-shot STCN models, highlighting the importance of even minimal video-level supervision.

Using various VOS models as the phase-1 model. During phase-1 training, we first train an STCN model on the low-shot VOS datasets; this model is referred to as the *phase-1 STCN*

Model	Pre-training	Y-2019 ($\mathcal{G}$)	D-2017 ($\mathcal{J}\&\mathcal{F}$)
STCN*		81.3	82.5
STCN*	✓	82.7	85.2
2-shot STCN		79.6	79.1
2-shot STCN	✓	80.6	81.0
2-shot STCN w/ Ours		81.3	82.5
2-shot STCN w/ Ours	✓	82.7	85.1
1-shot STCN		69.8	64.1
1-shot STCN	✓	72.9	68.0
1-shot STCN w/ Ours		80.1	81.4
1-shot STCN w/ Ours	✓	81.8	83.9

TABLE 4.16. Ablation study for pre-training on static image datasets. The symbol * denotes results are reproduced using open-source code. Y-2019 and D-2017 represent YouTube-2019 and DAVIS-2017, respectively.

Method	YouTube-VOS 2019				
	$\mathcal{G}$	$\mathcal{J}_S$	$\mathcal{F}_S$	$\mathcal{J}_U$	$\mathcal{F}_U$
0-shot STCN	71.1	67.9	70.5	68.8	77.1
1-shot STCN	$72.9_{+1.8}$	71.3	73.6	69.5	77.3
2-shot STCN	$80.6_{+7.7}$	79.5	83.8	75.6	83.4

TABLE 4.17. Ablation study on zero-shot STCN. In this setting, the STCN model is trained solely on the static pre-training images used in the original STCN and is directly evaluated on the YouTube-VOS 2019 benchmark.

model. We then apply our intermediate inference strategy using the phase-1 STCN model to generate pseudo labels for all unlabeled frames. This produces a *pseudo-labeled low-shot VOS dataset*, in which labeled frames retain ground-truth annotations, while unlabeled frames are paired with pseudo labels. In our implementation, all pseudo labels are stored in a pseudo-label bank.

The goal of phase-2 training is to train any VOS model on this pseudo-labeled dataset, instead of the original low-shot dataset, using the low-shot training strategies introduced in phase-2. Importantly, although a model such as XMem may be stronger than STCN, training on the

Phase-1 model	Phase-2 model	$\mathcal{G}$
STCN	STCN	82.7
	XMem	84.5
XMem	STCN	82.1
	XMem	83.7

TABLE 4.18. Ablation study on using different models as the phase-1 model for two-shot YouTube-VOS 2019.

pseudo-labeled low-shot dataset is independent of the STCN architecture; STCN simply serves as the pseudo-label generator for phase-1.

Table 4.18 presents a two-shot study on YouTube-VOS 2019, comparing two configurations: (1) STCN as the phase-1 model, followed by either STCN or XMem as the phase-2 model; and (2) XMem as the phase-1 model, followed by either STCN or XMem in phase-2. In all cases, the performance is evaluated using the phase-2 model. The results show that using STCN in phase-1 consistently leads to superior outcomes compared with using XMem. Although XMem is generally a stronger VOS model, it requires 8 input frames and predicts each frame sequentially from previous predictions, making phase-1 training more challenging and exacerbating early-stage error propagation when most frames are unlabeled. In contrast, STCN operates on only 3 input frames, substantially reducing error accumulation.

Table 4.19 further reports the phase-1 results on the two-shot YouTube-VOS 2019 benchmark, where STCN surpasses XMem by +1.2 $\mathcal{G}$ score. Thus, we adopt STCN as the phase-1 model.

In summary, VOS models that rely on fewer input frames (e.g., STCN) are more suitable for phase-1 because they mitigate error propagation when pseudo-label quality is low, whereas stronger but more sequence-dependent models (e.g., XMem) are better suited for phase-2 training once high-quality pseudo labels have been established.

The integration of the fine-tuned SAM in the two-shot setting. In the two-shot setting, we observe that 2-shot STCN, when trained with our proposed strategy, achieves performance nearly identical to the fully supervised STCN trained on a fully labeled dataset. However, applying the same methodology to XMem results in a slight performance drop relative to

Phase-1 model	$\mathcal{G}$
STCN	81.6
XMem	80.4

TABLE 4.19. Phase-1 performance of STCN and XMem on two-shot YouTube-VOS 2019.

Method	w/ Fine-tuned SAM	$\mathcal{G}$
Full-set XMem	-	85.3
2-shot XMem w/ Ours		84.5
2-shot XMem w/ Ours	✓	$84.9_{+0.4}$

TABLE 4.20. The effectiveness of the fine-tuned SAM integration in the two-shot setting with two-shot XMem on YouTube-VOS 2019.

its full-set counterpart. To further enhance the two-shot XMem model, we integrate the fine-tuned SAM model into the training pipeline, using the same integration approach as in the one-shot setting, while keeping all other configurations unchanged.

Table 4.20 reports the results. Incorporating the fine-tuned SAM model leads to a performance gain of 0.4 points compared with the baseline (without SAM), and reduces the performance gap to the full-set XMem from -0.8 to -0.4 points.

Enhancing VOST Performance through Additional Datasets. We investigate the effect of incorporating the sparsely labeled VISOR (Darkhalil et al., 2022) dataset to improve the low-shot performance of three VOS models, including STCN, RDE-VOS, and XMem, on the challenging VOST (Tokmakov et al., 2023) benchmark. VISOR is derived from the EPIC-KITCHENS egocentric video dataset (Damen et al., 2018, 2022) and contains 5,309 training clips annotated for VOS. For each VISOR clip, we use only the annotation of the first frame and discard all subsequent frame annotations, resulting in a sparsely labeled version referred to as *sparse VISOR*.

We explore three training configurations: (1) training on the combination of the full VOST dataset and sparse VISOR; (2) training on the combination of the two-shot VOST dataset and

sparse VISOR; and (3) training on the combination of the one-shot VOST dataset and sparse VISOR. The only difference between these settings is the amount of VOST supervision used (full, two-shot, or one-shot). All training procedures follow the methodology described in the main paper, with the exception that the training data now consists of both VOST and sparse VISOR. The sampling ratio between VOST and VISOR is fixed at 1:1.

Table 4.21 summarizes the performance of STCN, RDE-VOS, and XMem trained with our low-shot strategy under these three settings. Two key observations emerge:

- Integrating sparse VISOR improves the VOST performance of all three models across full-set, two-shot, and one-shot settings. For example, the 1-shot STCN trained on the combination of 1-shot VOST and sparse VISOR surpasses the 1-shot STCN trained on VOST alone by +0.7 $\mathcal{J}_{tr}$.
- The additional supervision provided by sparse VISOR reduces the performance gap between low-shot and full-set models. For instance, the 2-shot STCN trained with sparse VISOR achieves a $\mathcal{J}_{tr}$ score of 31.2, only -0.4 below the STCN trained on the fully labeled VOST dataset.

Although sparse VISOR improves performance on VOST, the gains remain moderate. Two primary factors contribute to this:

(1) **Domain mismatch.** VOST contains videos from both EPIC-KITCHENS (Damen et al., 2018, 2022) (kitchen-only) and Ego4D (Grauman et al., 2022), which includes diverse indoor and outdoor scenarios. VISOR, however, is sourced exclusively from EPIC-KITCHENS. This limited domain coverage makes it difficult for sparse VISOR to substantially enhance VOST performance, which requires modeling significantly broader environmental variability.

(2) **Object transformation complexity.** VOST emphasizes tracking complex object transformations, which involve changes in object state or structure over time, such as a carrot being chopped into pieces, where each resulting piece must be linked back to the original object. VISOR, in contrast, primarily focuses on segmenting hands and objects in kitchen environments, with only a small fraction involving

Method	VISOR	VOST	$\mathcal{J}_{tr}$	$\mathcal{J}$
STCN (Cheng et al., 2021c)		100%	31.6	42.6
	✓		$32.5_{+0.9}$	$43.8_{+1.2}$
2-shot STCN w/ Ours		**1.91**%	30.7	42.0
	✓		$31.2_{+0.5}$	$42.6_{+0.6}$
1-shot STCN w/ Ours		**0.95**%	29.0	41.2
	✓		$29.7_{+0.7}$	$42.2_{+1.0}$
RDE-VOS (Li et al., 2022c)		100%	33.4	43.2
	✓		$34.5_{+1.1}$	$45.0_{+1.8}$
2-shot RDE-VOS w/ Ours		**1.91**%	32.6	42.6
	✓		$33.2_{+0.6}$	$43.5_{+0.9}$
1-shot RDE-VOS w/ Ours		**0.95**%	30.5	41.4
	✓		$31.5_{+1.0}$	$42.5_{+1.1}$
XMem (Cheng and Schwing, 2022)		100%	33.9	44.8
	✓		$34.5_{+0.6}$	$46.0_{+1.2}$
2-shot XMem w/ Ours		**1.91**%	33.1	43.2
	✓		$33.5_{+0.4}$	$43.9_{+0.7}$
1-shot XMem w/ Ours		**0.95**%	31.2	42.0
	✓		$31.7_{+0.5}$	$42.9_{+0.9}$

TABLE 4.21. The impact of incorporating an additional dataset, sparse VISOR, on the VOST performance of three models, STCN, RDE-VOS, and XMem, each utilizing our low-shot training strategy, across various settings. "VISOR" indicates the utilization of an additional dataset. "VOST" represents the percentage of VOST labeled data.

transformation-heavy scenarios. As a result, the supervision from sparse VISOR contributes less to the transformation-centric challenges posed by VOST.

How about more shots? We further evaluate our approach under the 4-shot and 6-shot settings. Applying our method to 4-shot and 6-shot STCN and performing one round of phase-1 training yields scores of 82.0% and 82.1% on YouTube-VOS 2019, respectively. After an additional round of phase-2 training, both models reach 82.7%, the same performance achieved by our 2-shot STCN. This result suggests that performance saturates at the 2-shot level, indicating that providing more labeled frames per video offers little to no additional benefit.

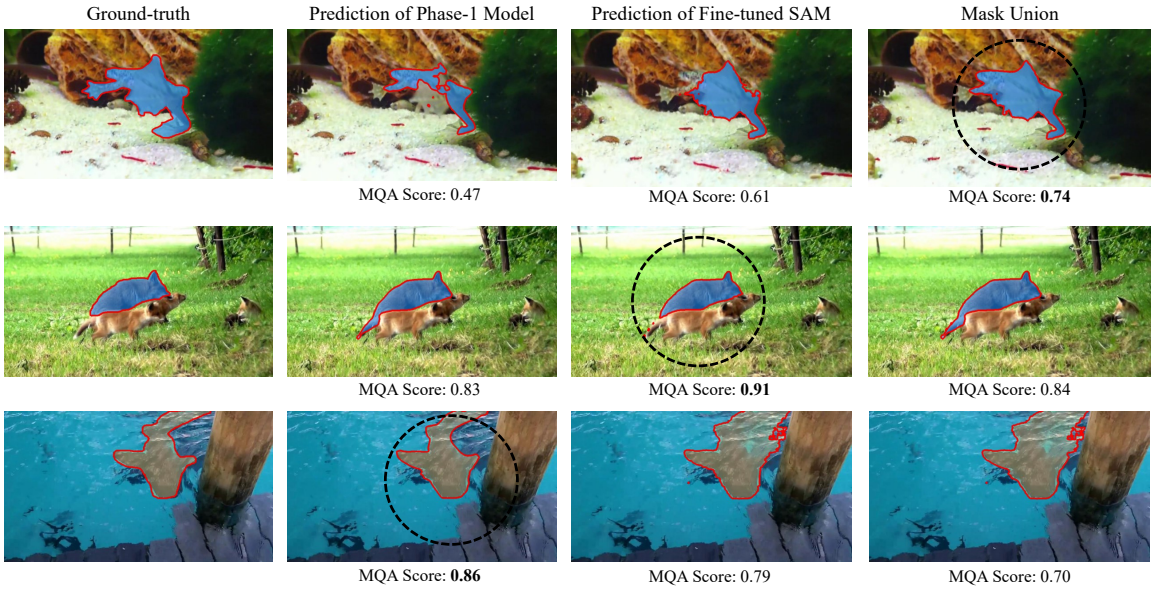


FIGURE 4.11. According to the MQA score, our mask quality assessment (MQA) module could identify the best mask prediction from: (1) the prediction from the phase-1 model (the second column); (2) the prediction from the fine-tuned SAM model (the third column); and (3) the mask union of (1) and (2) (the last column). Each raw represents a randomly selected frame from the VOS benchmark.

4.4.6 Visualization

Visualization for mask quality assessment (MQA) module. Our MQA module selects the highest-quality mask from three candidates: (1) the prediction generated by the phase-1 model; (2) the prediction from the fine-tuned SAM model; and (3) the union mask combining (1) and (2). Figure 4.11 visualizes the ground-truth mask alongside the three candidate predictions. As shown, the mask selected by the MQA module consistently corresponds to the candidate with the highest quality score.

4.5 Chapter Summary

In this chapter, we investigated the problem of *low-shot video object segmentation* (VOS), where each training video is annotated with only one or two frames. We first formulated low-shot VOS as an extreme semi-supervised setting and highlighted its practical importance,

given the high cost of dense video annotation. Building on this formulation, we proposed a simple yet effective two-phase training paradigm that unlocks the potential of unlabeled frames. In phase-1, a VOS model (e.g., STCN) is trained in a semi-supervised manner with labeled frames serving as references and high-confidence predictions providing pseudo labels for unlabeled frames. An intermediate bidirectional inference step then constructs a pseudo-label bank, which is leveraged in phase-2 to retrain the VOS model on a mixture of ground-truth and pseudo labels while dynamically refreshing the bank as predictions improve.

We further extended this framework to the more challenging one-shot setting by integrating SAM as an auxiliary universal segmentation model. A lightweight SAM fine-tuning strategy with point-prompt augmentation was introduced, together with a mask quality assessment (MQA) module that selects the best pseudo mask from the phase-1 model, fine-tuned SAM, and their union. This collaboration substantially enhances pseudo-label quality without incurring extra inference cost at test time. Extensive experiments on DAVIS 2016/2017, YouTube-VOS 2018/2019, LVOS, and VOST demonstrated that, with only 7.3% (two-shot) or 3.7% (one-shot) of the labeled data on YouTube-VOS, and even smaller ratios on DAVIS, LVOS, and VOST, our method achieves performance comparable to or even surpassing fully supervised baselines, and consistently outperforms naive low-shot training.

Overall, the chapter delivers three key messages:

- Low-shot VOS is not only practically appealing but also technically feasible: strong VOS models can be trained with one or two labeled frames per video.
- The proposed two-phase semi-supervised training paradigm, powered by a pseudo-label bank and bidirectional inference, is model-agnostic and generalizes well across different VOS architectures (STCN, RDE-VOS, XMem) and datasets.
- Carefully designed pseudo-label refinement, including SAM-based fine-tuning, point-prompt augmentation, and MQA, plays a crucial role in the one-shot regime, enabling competitive performance with minimal annotation cost and revealing that performance saturates quickly once a small number of high-quality annotations is available.

CHAPTER 5

Data-Efficient Video Understanding from Image-Level Supervision

---

In this chapter, we study the feasibility of learning video understanding models from image data and introduce a novel data-efficient, image-driven methodology, MinMaxVIS (Wei et al., 2025), for video instance segmentation (VIS). VIS is a more challenging setting compared to VOS introduced in Chapter 4, as it not only requires tracking instances that appear in previous frames, but also predicting the category label for each instance.

Section 5.1 presents the problem formulation of VIS. Section 5.2 describes the motivation for efficiently leveraging image data to train video models: MinMaxVIS enables video understanding by learning from a small amount of labeled static images together with a large collection of unlabeled images, in contrast to prior VIS approaches that typically (1) rely on video data for training and (2) require dense, video-level annotations. Section 5.3 details the key technical innovations, including how to train video models using image data, how to retrieve useful information from large-scale unlabeled images, and how to mitigate training noise introduced by pseudo-labeled data. In Section 5.4, extensive experiments on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS demonstrate that MinMaxVIS not only achieves substantial improvements over existing image-driven baselines but also outperforms the fully supervised MinVIS while using only 1–10% of the labeled data. Finally, Section 5.5 concludes the chapter by showing that high-quality VIS can be achieved without relying on dense video annotations.

5.1 Problem Formulation

Video Instance Segmentation. Video Instance Segmentation (VIS) aims to detect, segment, and track each object instance throughout a video. Given an input video $\{\mathbf{I}_t\}_{t=1}^{T}$, the goal is to produce temporally consistent instance masks,

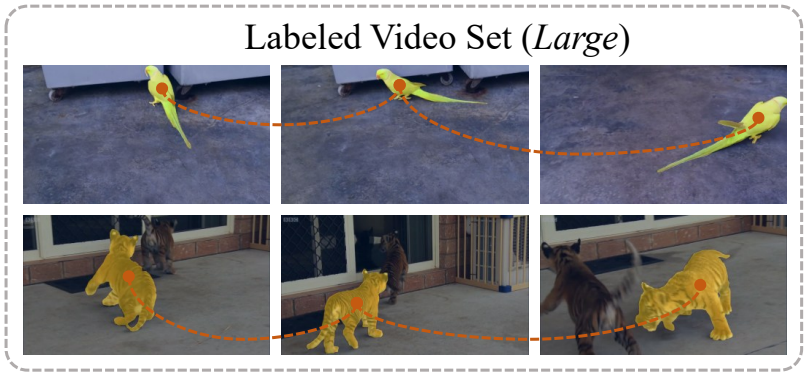
$$\{(\mathbf{M}_t^k, \mathrm{id}^k) \mid k = 1, \ldots, K_t;\ t = 1, \ldots, T\},$$

where $\mathbf{M}_t^k$ denotes the mask of the k-th object at frame t, and id^k ensures identity consistency across frames. Compared to image instance segmentation, VIS introduces additional challenges such as object motion, appearance variations, occlusions, and long-term re-identification.

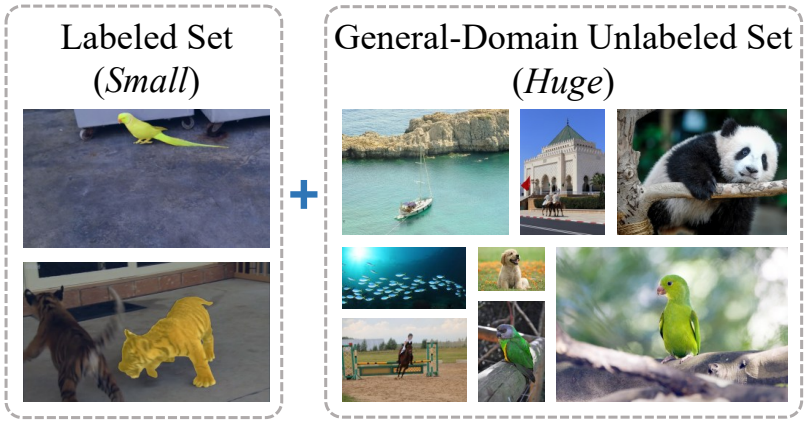
The Cost of Dense Video Annotation. Training VIS models usually requires dense, frame-wise instance labels for every object in the video. However, annotating all frames is extremely costly and time-consuming: a single short video may contain hundreds of frames, each requiring accurate pixel-level instance masks. This annotation burden limits the scale of VIS datasets and motivates the need for *data-efficient learning* strategies that reduce the reliance on densely labeled video data.

Semi-Supervised Image Segmentation. Semi-supervised learning (SSL) in image segmentation offers a promising direction for reducing annotation cost. Let $\mathcal{L} = \{(\mathbf{I}, \mathbf{M}^{\mathrm{gt}})\}$ denote the set of labeled images and $\mathcal{U} = \{\mathbf{I}\}$ the unlabeled set. SSL techniques leverage both labeled and unlabeled images through consistency regularization, pseudo-label generation, and teacher–student frameworks, enabling strong segmentation performance even when labeled annotations are sparse.

Generalizing Image-Level SSL to VIS. Although SSL methods are primarily designed for static images, the segmentation priors they learn, including object boundaries, shape cues, and instance discrimination, naturally transfer to videos. By training a segmentation module using semi-supervised image segmentation, we obtain robust per-frame mask predictors that can be extended to the video domain with lightweight temporal association. This motivates our approach: we train on image-level SSL and generalize the learned segmentation capability



(a) Traditional video-based VIS training.



(b) MinMaxVIS image-based training.

FIGURE 5.1. (a) Traditional VIS models rely on fully labeled video frames with instance association across frames, demanding extensive manual annotations. (b) MinMaxVIS enables effective video instance segmentation using only a small set of labeled target-domain images and a vast amount of unlabeled general-domain images, significantly reducing annotation costs while maximizing data efficiency.

to the VIS task, achieving a data-efficient yet competitive video instance segmentation framework.

5.2 Motivation

The goal of Video Instance Segmentation (VIS) is to identify, segment, and consistently track each object instance across all frames in a video sequence. Existing VIS methods (Hwang et al., 2021; Wu et al., 2021; Huang et al., 2022; Li et al., 2023c; Kim et al., 2024) typically

rely on fully labeled video datasets, where every frame requires detailed instance masks, category labels, and association labels that link the same instance across time. Although effective, this reliance on dense video annotations imposes a substantial burden: labeling hundreds of frames per video is both time-consuming and costly, particularly for high-quality datasets that demand high-fidelity per-pixel annotations.

In this work, we seek to fundamentally reduce the dependency on fully labeled video data. As illustrated in Figure 5.1, we introduce **MinMaxVIS**, a novel VIS framework that requires only a small number of labeled images from the target domain, while leveraging a large collection of general-domain, unlabeled internet images. Concretely, MinMaxVIS uses merely $\sim$1,200 labeled images (about 2% of the frames) from YouTube-VIS 2019 (Yang et al., 2019a) and 2.8 million unlabeled images from SA-1B (Kirillov et al., 2023b). This eliminates the need for fully annotated video sequences and dramatically lowers annotation cost. The name *MinMaxVIS* reflects our design philosophy: *minimize* the amount of labeled data and *maximize* the utility of unlabeled data.

The training pipeline of MinMaxVIS consists of three stages. First, a preliminary segmentation model is trained on the small labeled image set from the target domain. This model is then used as a retrieval mechanism to identify relevant instances within millions of general-domain unlabeled images, producing a pseudo-labeled dataset where pseudo-masks are generated by the preliminary model. This retrieval step is crucial: directly applying semi-supervised learning (Tarvainen and Valpola, 2017; Xu et al., 2021; Yan et al., 2023; Hu et al., 2021a) to millions of unrelated images would be computationally expensive and inefficient. Instead, retrieval ensures that the pseudo-labeled dataset is both compact and rich in domain-relevant content. Finally, MinMaxVIS is trained using both the small labeled dataset and the large pseudo-labeled dataset.

Despite its advantages, this VIS training paradigm introduces two key challenges. **(1) Noisy pseudo-labels:** Pseudo-labeled images often suffer from false negatives, which are instances that are present but not recalled, leading to ambiguous background regions. To address this, we propose a *selective gradient backpropagation* strategy, which backpropagates gradients only from foreground queries and high-confidence background queries. Low-confidence

background queries, which may correspond to either true background or missed foreground instances, are detached to prevent noise from corrupting training. **(2) Absence of instance association labels:** Classical VIS datasets (Yang et al., 2019a, 2021b; Qi et al., 2022) provide explicit instance associations across frames, but training on static images does not. To resolve this, we simulate video pairs through augmentations and enforce cross-frame instance consistency by maximizing similarity for matching instances and minimizing similarity for mismatched ones.

Prior work, such as MinVIS (Huang et al., 2022), has demonstrated that VIS models can be trained on static images. Building upon this insight, MinMaxVIS further enhances segmentation robustness while dramatically reducing reliance on labeled data and effectively exploiting large-scale unlabeled data. Empirically, MinMaxVIS outperforms MinVIS, which relies on 100% labeled data, even when using only 2%-10% labeled images together with SA-1B unlabeled data. For example, on YouTube-VIS 2019, MinMaxVIS achieves 62.2 mAP with a Swin-L backbone using only 2% labeled data, surpassing MinVIS trained on the fully labeled dataset by 0.6 mAP.

Overall, MinMaxVIS demonstrates that high-quality VIS can be achieved with minimal labeled data by maximizing the value extracted from large-scale unlabeled images.

5.3 Methodology

5.3.1 Overview

Problem Formulation. Let $\mathcal{D}_{\text{labeled}} = \{(\boldsymbol{I}_i, \boldsymbol{L}_i)\}_{i=1}^{N}$ denote a small labeled image set from the target domain, where $\boldsymbol{I}_i$ is an image and $\boldsymbol{L}_i$ contains the corresponding instance annotations, including segmentation masks and class labels. Here, N represents the total number of labeled images. In addition, we consider a large unlabeled image set $\mathcal{D}_{\text{unlabeled}} = \{\boldsymbol{I}_j\}_{j=1}^{M}$ collected from a general domain, where M is the total number of unlabeled images. Although these images lack target-domain annotations, they cover diverse visual content and provide useful general features that can be exploited to improve the final MinMaxVIS model.

Importantly, $M \gg N$, highlighting the significant scale disparity between the unlabeled and labeled datasets.

Our objective is to train a video instance segmentation model, MinMaxVIS, that effectively leverages the small labeled set $\mathcal{D}_{\text{labeled}}$ from the target domain together with the large general-domain unlabeled set $\mathcal{D}_{\text{unlabeled}}$. By jointly utilizing these two sources, MinMaxVIS is designed to achieve strong generalization and robust segmentation performance on the target domain.

Overview. The overall MinMaxVIS framework, illustrated in Figure 5.2, is composed of three key stages: training a preliminary segmentation model on the small labeled image set, performing high-precision retrieval on the large unlabeled set, and conducting the final MinMaxVIS training. First, a segmentation model is trained on $\mathcal{D}_{\text{labeled}}$, providing the foundation for subsequent steps (Section 5.3.2). This trained model is then employed to retrieve target-relevant instances from the large unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$, thereby constructing a pseudo-labeled dataset $\mathcal{D}_{\text{pseudo}}$ enriched with target-domain content (Section 5.3.3). In the final training stage, MinMaxVIS jointly leverages $\mathcal{D}_{\text{labeled}}$ and $\mathcal{D}_{\text{pseudo}}$, incorporating selective gradient backpropagation to handle noise in pseudo-labels, augmented paired images to simulate frame continuity, and instance association techniques to establish cross-frame consistency (Section 5.3.4). The inference pipeline for MinMaxVIS is described in Section 5.3.5.

5.3.2 Preliminary Segmentation Model Training

The first stage trains a preliminary segmentation model $\boldsymbol{S_\theta}$ on the small labeled dataset $\mathcal{D}_{\text{labeled}}$. The purpose of this model is to enable high-precision retrieval of target-relevant instances from the unlabeled image set $\mathcal{D}_{\text{unlabeled}}$, thereby ensuring strong content alignment with the target domain while avoiding the computational inefficiency of applying large-scale semi-supervised learning directly to the entire unlabeled corpus.

In our implementation, we adopt Mask2Former (Cheng et al., 2022a) with a Swin-L backbone as the preliminary segmentation model, trained using standard classification and segmentation

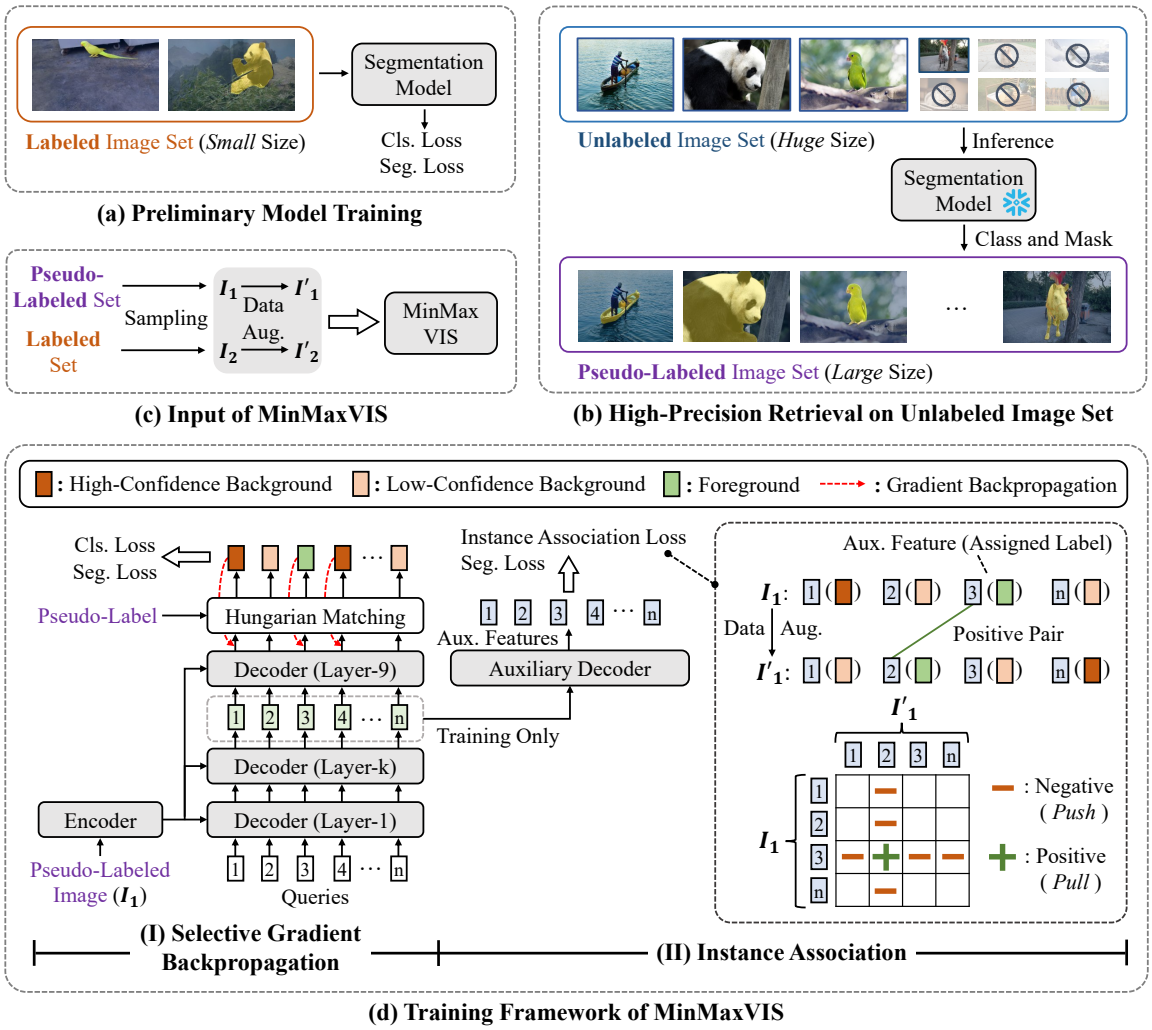


FIGURE 5.2. Overview of the MinMaxVIS framework, consisting of three main stages: (a) Preliminary segmentation model training on a small labeled set; (b) High-precision retrieval from a large unlabeled image dataset to create a pseudo-labeled set containing only high-confidence samples; (c) Input preparation for MinMaxVIS, incorporating both labeled and pseudo-labeled sets; (d) MinMaxVIS employs an encoder-decoder architecture with (I) selective Gradient Backpropagation to mitigate noisy pseudo-labels and (II) an auxiliary decoder with an instance association loss applied on augmented image pairs. *The process for generating auxiliary features for $\boldsymbol{I}'$, the augmented version of the original image $\boldsymbol{I}$, is identical to that of $\boldsymbol{I}$. For simplicity, we omit the illustration of processing $\boldsymbol{I}'$.*

loss functions. Owing to the small size of $\mathcal{D}_{\text{labeled}}$, this stage incurs minimal computational cost, producing a strong initialization for subsequent retrieval and pseudo-label construction.

5.3.3 High-Precision Retrieval

In the second stage, the preliminary segmentation model $\boldsymbol{S_\theta}$ trained on $\mathcal{D}_{\text{labeled}}$ is used to retrieve target-relevant instances from the large unlabeled dataset $\mathcal{D}_{\text{unlabeled}}$, as illustrated in Figure 5.2(b). The goal of this stage is to construct a pseudo-labeled dataset $\mathcal{D}_{\text{pseudo}}$ that is highly aligned with the target domain while minimizing the inclusion of irrelevant images.

For each image $\boldsymbol{I}_j$ in $\mathcal{D}_{\text{unlabeled}} = \{\boldsymbol{I}_j\}_{j=1}^{M}$, the model $\boldsymbol{S_\theta}$ produces a set of predictions $\{(c_u, \boldsymbol{M}_u)\}_{u=1}^{U}$, where c_u is the confidence score of the u-th prediction, $\boldsymbol{M}_u$ is the corresponding segmentation mask, and U is the total number of predictions for that image. To ensure high precision, only predictions with confidence scores exceeding a predefined threshold τ are retained as pseudo-labels. Formally, for each image $\boldsymbol{I}_j$, we keep

$$\widehat{\boldsymbol{L}}_j = \{(c_u, \boldsymbol{M}_u) \mid c_u \geq \tau\}_{u=1}^{U},$$

where τ is intentionally set to a high value to suppress false positives. Although this may introduce false negatives (i.e., missed instances), their effect is later mitigated during MinMaxVIS training via selective gradient backpropagation (Section 5.3.4).

Images that contain no predictions satisfying $c_u \geq \tau$ are discarded. The result is a filtered pseudo-labeled dataset

$$\mathcal{D}_{\text{pseudo}} = \{(\boldsymbol{I}_j, \widehat{\boldsymbol{L}}_j)\}_{j=1}^{\widehat{M}},$$

where $\widehat{M}$ denotes the number of retained images. Notably, $\widehat{M} \ll M$, as only a small subset of $\mathcal{D}_{\text{unlabeled}}$ contains high-confidence predictions. This curated pseudo-labeled set, together with the small labeled dataset $\mathcal{D}_{\text{labeled}}$, forms the foundation for the final MinMaxVIS training stage, ensuring that training is both efficient and focused on target-domain content.

5.3.4 MinMaxVIS Training

MinMaxVIS is trained using a combination of the small labeled dataset $\mathcal{D}_{\text{labeled}} = \{(\boldsymbol{I}_i, \boldsymbol{L}_i)\}_{i=1}^{N}$ and the large pseudo-labeled dataset $\mathcal{D}_{\text{pseudo}} = \{(\boldsymbol{I}_j, \widehat{\boldsymbol{L}}_j)\}_{j=1}^{\widehat{M}}$, where $N \ll \widehat{M}$. As shown in Figure 5.2(d), MinMaxVIS employs an encoder–decoder architecture augmented with an

auxiliary decoder, which is specifically designed to strengthen instance association during training.

Encoder–Decoder. As illustrated in Figure 5.2(d), the encoder (either ResNet-50 or Swin-L) processes each input image, whether from the labeled set $\mathcal{D}_{\text{labeled}}$ or the pseudo-labeled set $\mathcal{D}_{\text{pseudo}}$, to extract multi-scale image features. Following MinVIS (Huang et al., 2022) and Mask2Former (Cheng et al., 2022a), these features are fed into a decoder composed of K layers ($K = 9$ by default). The decoder begins with n learnable object queries that are iteratively refined across layers. Each layer contains three components: (1) a self-attention module; (2) a cross-attention module in which image features serve as keys and values and object queries act as queries; and (3) a feed-forward network (FFN). After each layer, Hungarian matching is performed to classify the refined queries into foreground or background. Foreground queries are supervised with both classification and segmentation losses, while background queries receive only classification loss, following the training objectives used in MinVIS (Huang et al., 2022).

To effectively learn from pseudo-labeled data, we apply a selective gradient backpropagation strategy to each layer of the main decoder, reducing the influence of false negatives in pseudo-labels. In addition, we introduce an auxiliary branch that operates in parallel with the main decoder to enhance instance association learning during MinMaxVIS training.

Selective Gradient Backpropagation. This strategy is applied exclusively to images from the pseudo-labeled set $\mathcal{D}_{\text{pseudo}}$. As described in Section 5.3.3, our high-precision retrieval process retains only predictions with very high confidence scores. Although this ensures highly precise pseudo-labels, it inevitably introduces false negatives (missed instances), which can contaminate training. The key idea behind selective gradient backpropagation is to use the model's background confidence to identify potentially unreliable background queries and detach their gradients during training, thereby preventing false negatives from adversely affecting learning; see Figure 5.2(d.I).

Formally, for a pseudo-labeled image $\boldsymbol{I}_j \in \mathcal{D}_{\text{pseudo}}$ with its pseudo-labels $\widehat{\boldsymbol{L}}_j$, MinMaxVIS encodes $\boldsymbol{I}_j$ together with n object queries. Each of the K decoder layers produces n query

features. Denote by $\{\boldsymbol{q}_1^k, \ldots, \boldsymbol{q}_n^k\}$ the query features output by the k-th decoder layer ($1 \leq k \leq K$). Hungarian matching is then performed to assign a label, which is foreground instance or background, to each query in $\{\boldsymbol{q}_1^k, \ldots, \boldsymbol{q}_n^k\}$ by comparing confidence scores, predicted masks, and the pseudo-labels $\widehat{\boldsymbol{L}}_j$.

For queries assigned to the background class, their background confidence scores are used to identify potential false negatives: queries with lower background confidence are more likely to correspond to missed foreground instances. Let $\{\overline{\boldsymbol{q}}_1^k, \ldots, \overline{\boldsymbol{q}}_{n'}^k\}$ denote the set of background queries produced by the k-th decoder layer, where n' is the number of such queries, and let $\{\overline{b}_1, \ldots, \overline{b}_{n'}\}$ be their corresponding background confidence scores. To mitigate noise introduced by false negatives in pseudo-labels, uncertain background queries are assigned reduced gradient weights. Selective Gradient Backpropagation (SGB) is then applied to the set $\{\overline{\boldsymbol{q}}_1^k, \ldots, \overline{\boldsymbol{q}}_{n'}^k\}$.

Formally, the SGB loss is defined as:

$$\mathcal{L}_{\text{SGB}} = \frac{1}{n'} \sum_{i=1}^{n'} w_i \cdot \mathcal{L}_{\text{cls}}(\overline{b}_i), \tag{5.1}$$

where $\mathcal{L}_{\text{cls}}$ denotes the classification loss and w_i is the gradient weight assigned to the i-th background query. We consider several weighting strategies:

- *Confidence weight:* $w_i = \overline{b}_i$.
- *Squared confidence weight:* $w_i = (\overline{b}_i)^2$.
- *Truncation weight:* $w_i = \mathbb{1}_{[\overline{b}_i \geq \beta]}$, where β is a threshold ensuring that only background queries with sufficiently high confidence contribute to the loss. This is our default choice.

Loss functions for foreground queries remain unchanged. The SGB loss is applied independently at each of the K decoder layers.

Instance Association. VIS requires consistent tracking of the same object instance across frames. Unlike conventional VIS datasets, which provide association labels for training, our model is trained on static images without explicit instance correspondence. To enable instance

association learning, we simulate video-like behavior by generating paired images through augmentations. As shown in Figure 5.2(d.II), an auxiliary decoder, running in parallel with the main decoder, processes each image pair (an original image and its augmented version) to facilitate instance association. The auxiliary decoder contains six layers, each mirroring the structure of the corresponding main decoder layer. The input to the auxiliary decoder is the output of the $\overline{K}$-th layer of the main decoder.

Given an image $\boldsymbol{I}$ from either $\mathcal{D}_{\text{labeled}}$ or $\mathcal{D}_{\text{pseudo}}$, with its annotation $\boldsymbol{L}$ or pseudo-label $\widehat{\boldsymbol{L}}$, the image and the n object queries are passed through the encoder, main decoder, and auxiliary decoder, producing n auxiliary features denoted by $\{\boldsymbol{a}_1^{\boldsymbol{I}}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}}\}$. We reuse the Hungarian matching results from the final layer of the main decoder to assign each auxiliary feature to a foreground or background class (see Figure 5.2(d.II)). Applying the same process to the augmented image $\boldsymbol{I}'$ yields its auxiliary features $\{\boldsymbol{a}_1^{\boldsymbol{I}'}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}'}\}$ and corresponding assignments. Because $\boldsymbol{I}'$ is an augmented version of $\boldsymbol{I}$, the two images share consistent instance labels, enabling us to establish reliable cross-image associations.

For each foreground auxiliary feature $\boldsymbol{a}_g^{\boldsymbol{I}}$ in the original image, we identify the corresponding feature in $\{\boldsymbol{a}_1^{\boldsymbol{I}'}, \ldots, \boldsymbol{a}_n^{\boldsymbol{I}'}\}$ that matches the same instance label; this is denoted as $\mathcal{M}(\boldsymbol{a}_g^{\boldsymbol{I}})$. We encourage high similarity for these matched pairs while discouraging similarity with all non-matching auxiliary features in $\boldsymbol{I}'$, denoted as $\mathcal{N}(\boldsymbol{a}_g^{\boldsymbol{I}})$.

Formally, the instance association loss for image $\boldsymbol{I}$ is defined as:

$$\begin{aligned}\mathcal{L}_{\text{IA}}^{\boldsymbol{I}} = & -\frac{1}{G}\sum_{g=1}^{G}\log\sigma\big(\boldsymbol{a}_g^{\boldsymbol{I}}\cdot\mathcal{M}(\boldsymbol{a}_g^{\boldsymbol{I}})\big) \\ & -\frac{1}{G(n-1)}\sum_{g=1}^{G}\sum_{h=1}^{n-1}\log\sigma\big(-\boldsymbol{a}_g^{\boldsymbol{I}}\cdot\mathcal{N}(\boldsymbol{a}_g^{\boldsymbol{I}})_h\big)\,,\end{aligned} \tag{5.2}$$

where G is the number of foreground instances in $\boldsymbol{I}$, n is the total number of queries, and $\sigma(\cdot)$ denotes the sigmoid function. Here, $\boldsymbol{a}_g^{\boldsymbol{I}}$ is the auxiliary feature of the g-th foreground instance in $\boldsymbol{I}$, $\mathcal{M}(\boldsymbol{a}_g^{\boldsymbol{I}})$ returns the matched auxiliary feature in $\boldsymbol{I}'$, and $\mathcal{N}(\boldsymbol{a}_g^{\boldsymbol{I}})$ returns the $(n-1)$ non-matching auxiliary features for that instance. The first term encourages similarity between matching pairs, while the second term suppresses similarity to non-matching pairs.

A symmetric loss $\mathcal{L}_{\text{IA}}^{\boldsymbol{I}'}$ is applied to the augmented image $\boldsymbol{I}'$. The final instance association objective is therefore

$$\mathcal{L}_{\text{IA}} = \mathcal{L}_{\text{IA}}^{\boldsymbol{I}} + \mathcal{L}_{\text{IA}}^{\boldsymbol{I}'}.$$

This loss is applied to both labeled and pseudo-labeled images. Additionally, we apply the same segmentation loss used in the main decoder to the auxiliary decoder to ensure stable and consistent training.

5.3.5 Inference

The auxiliary decoder, together with the instance association loss, serves to enhance the discriminative capability of the query features produced by the main decoder, enabling them to associate the same object instances across frames. During inference, the auxiliary decoder is discarded, and only the encoder–decoder architecture is retained.

MinMaxVIS adopts streaming (online) inference. For an incoming video stream, each frame is passed through MinMaxVIS (without the auxiliary decoder) to generate n predictions, each corresponding to an object query and containing both a classification score and a mask prediction. Hungarian matching is then applied to associate predictions across consecutive frames, ensuring temporal consistency, as illustrated in Figure 5.3.

5.4 Experiment

5.4.1 Experimental Setup

Datasets. We evaluate MinMaxVIS on three benchmark datasets: YouTube-VIS 2019 (Yang et al., 2019a), YouTube-VIS 2021 (Yang et al., 2021b), and OVIS (Qi et al., 2022), using two data-efficiency settings for each. For YouTube-VIS 2019, which contains 61,845 labeled training frames, we use 1% and 2% of the data, corresponding to 618 and 1,236 images, respectively. For YouTube-VIS 2021, whose training set includes 90,160 frames, we again adopt 1% and 2% subsets, yielding 901 and 1,803 images. Due to the increased difficulty

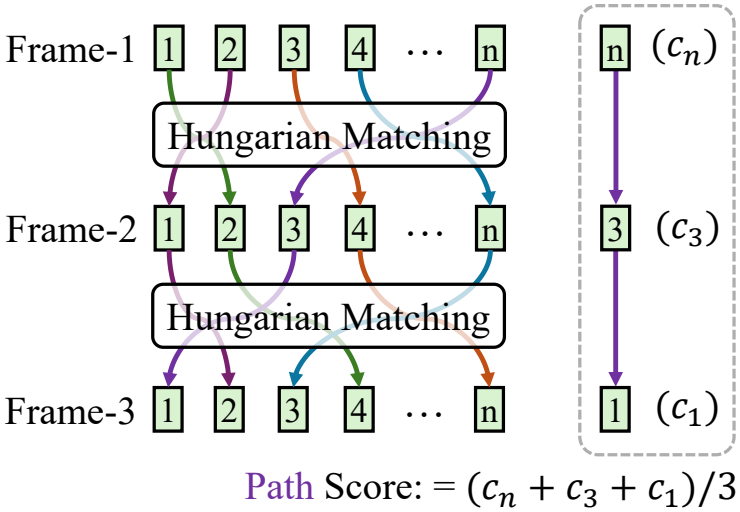


FIGURE 5.3. Illustration of the inference process (example with three frames). Each frame is independently processed by MinMaxVIS to produce n query features (indicated by green rectangles) from the main decoder. Each query feature generates a classification score c and a mask prediction. Hungarian matching is then applied between pairs of consecutive frames to associate predictions based on the similarity of query features, resulting in n paths across the frames. The path score is computed by averaging the classification scores along the path. This path score is then used as the final classification score for all predictions along the path.

of OVIS, we evaluate using 5% and 10% of its 42,149 training images, resulting in 2,107 and 4,214 images. For the general-domain unlabeled data, we randomly sample 2.8 million images from SA-1B (Kirillov et al., 2023b).

Implementation Details. For YouTube-VIS 2019 (1%) and YouTube-VIS 2019 (2%), our high-precision retrieval procedure (Section 5.3.3) yields 25,669 and 28,033 pseudo-labeled images from SA-1B, respectively. For YouTube-VIS 2021 (1%) and YouTube-VIS 2021 (2%), we retrieve 24,088 and 25,807 pseudo-labeled images. For OVIS (5%) and OVIS (10%), the retrieved pseudo-labeled sets contain 15,966 and 17,006 images, respectively. During retrieval, we set $\tau = 0.99$; however, for categories with fewer than 1,000 retrieved pseudo-labeled samples, we relax the threshold to $\tau = 0.9$.

In MinMaxVIS, the number of object queries n is set to 200 when using the Swin-L backbone and 100 for the ResNet-50 backbone. For the truncation-weight variant in Eq. 5.1, we use

Hyper-Parameter	Value
Number of Decoder Layers	9
Query Feature Dimension	256
Number of Attention Heads	8
FFN Dimension	2048
Number of Auxiliary Decoder Layers	6
Retrieval Threshold τ	0.99
Maximal Retrieval Number Per Category W	1,000
Truncation Weight β	0.5
Ratio of Labeled to Pseudo-Labeled	1:4
Classification Cost (Hungarian Matching)	2.0
Mask Loss Cost (Hungarian Matching)	5.0
Dice Loss Cost (Hungarian Matching)	2.0
Loss Weight (Classification)	2.0
Loss Weight (Mask)	5.0
Loss Weight (Dice)	2.0
Loss Weight (Association)	1.0
Inference Resolution	480p

TABLE 5.1. Summary of the hyper-parameters used in MinMaxVIS.

$\beta = 0.5$. Each training batch contains a mixture of labeled and pseudo-labeled images with a 1:4 ratio. The total batch size is 64, including both original and augmented images. MinMaxVIS is trained for 6,000 iterations with an initial learning rate of 1×10^{-4}, which is reduced by a factor of 0.1 at 4,000 iterations. Table 5.1 provides an overview of the hyper-parameters used in MinMaxVIS.

Evaluation. We evaluate MinMaxVIS on the validation sets of YouTube-VIS 2019, YouTube-VIS 2021, and OVIS. Following standard protocol, we report mean Average Precision (mAP) as the primary evaluation metric.

5.4.2 Main Results

Table 5.2 summarizes the performance of different methods on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS under various labeled-data regimes: 1% and 2% for the YouTube-VIS datasets, and 5% and 10% for OVIS. We evaluate MinMaxVIS using both ResNet-50

and Swin-L backbones and compare it with recent approaches implemented with the same backbones. For each competing method, we reproduce results under the corresponding low-label settings. For methods that require video inputs, we approximate video sequences by applying data augmentations to individual images. Due to architectural differences, adapting these video-centric methods to directly exploit unlabeled image data is nontrivial.

MinMaxVIS, which builds upon MinVIS (Huang et al., 2022), belongs to the image-driven family of approaches. Our comparison with MinVIS reveals two principal observations:

- *Effective Utilization of Unlabeled Data.* MinMaxVIS makes substantially better use of unlabeled images, outperforming low-data MinVIS by large margins across all three datasets and both backbones. For example, with a ResNet-50 backbone, MinMaxVIS exceeds MinVIS by 12.9% and 12.2% mAP under the YouTube-VIS 2019 (1%) and (2%) settings, respectively.
- *Competitiveness with Full-Set MinVIS.* Despite using only a small fraction of labeled data, MinMaxVIS matches or surpasses MinVIS trained on the full labeled set (100% data). For instance, with only 2% labeled data on YouTube-VIS 2019 and 2021, MinMaxVIS (ResNet-50) outperforms full-set MinVIS by 2.6% and 1.5% mAP, respectively.

5.4.3 Ablation Studies

For all ablation studies, we use MinMaxVIS with a Swin-L backbone on YouTube-VIS 2019 using 2% labeled data.

Main Components. Table 5.3 reports the ablation results for the major components of MinMaxVIS. Starting from the baseline MinVIS (Huang et al., 2022), high-precision retrieval constructs a pseudo-labeled dataset that is well aligned with the target domain, improving mAP from 58.5% to 60.3%. Incorporating selective gradient backpropagation yields a further gain, boosting mAP to 61.4% by mitigating noise in pseudo-labels. Finally, adding the instance association module lifts performance to 62.2%, underscoring the importance of explicitly enhancing instance-level correspondence across frames.

Method	Backbone	Setting	YouTube-VIS 2019			YouTube-VIS 2021			OVIS		
			1%	2%	100%	1%	2%	100%	5%	10%	100%
VITA (Heo et al., 2022)	ResNet-50	Offline	32.3	37.0	49.8	28.2	33.5	45.7	8.2	10.4	19.6
DVIS (Zhang et al., 2023e)	ResNet-50	Online	23.9	28.3	51.2	27.0	32.3	46.4	10.9	20.6	30.4
DVIS++ (Zhang et al., 2023f)	ResNet-50	Online	29.8	36.1	55.5	31.5	33.9	50.0	10.7	14.4	37.2
GenVIS (Heo et al., 2023)	ResNet-50	Online	31.9	32.7	50.0	26.4	31.4	47.1	11.3	14.7	35.8
DVIS-DAQ (Zhou et al., 2024)	ResNet-50	Online	33.4	38.9	55.2	32.3	33.0	50.4	10.8	16.5	38.7
Image-Driven Approach											
MinVIS (Huang et al., 2022)	ResNet-50	Online	33.7	37.8	47.4	29.6	31.1	44.2	22.2	22.5	25.0
MinMaxVIS (Ours)	ResNet-50	Online	**46.6**	**50.0**	-	**44.6**	**45.7**	-	**24.7**	**26.8**	-
VITA (Heo et al., 2022)	Swin-L	Offline	52.9	56.0	63.0	43.9	47.6	57.5	13.8	18.8	27.7
DVIS (Zhang et al., 2023e)	Swin-L	Online	36.8	36.2	63.9	43.9	49.6	58.7	23.6	29.5	46.0
DVIS++ (Zhang et al., 2023f)	VIT-L	Online	49.2	52.4	67.7	44.5	47.1	62.3	18.3	26.1	49.6
GenVIS (Heo et al., 2023)	Swin-L	Online	53.8	56.6	64.0	45.0	46.1	59.6	14.7	21.5	45.2
DVIS-DAQ (Zhou et al., 2024)	VIT-L	Online	51.8	57.7	68.3	46.6	50.4	62.4	15.9	24.0	53.7
Image-Driven Approach											
MinVIS (Huang et al., 2022)	Swin-L	Online	56.2	58.5	61.6	49.8	52.4	55.3	36.1	37.0	39.4
MinMaxVIS (Ours)	Swin-L	Online	**60.9**	**62.2**	-	**54.1**	**55.6**	-	**37.5**	**39.2**	-

TABLE 5.2. Performance comparison (mAP in %) of various methods on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS datasets across different labeled data settings. MinMaxVIS, built upon MinVIS, effectively leverages unlabeled data to achieve superior performance in low-data regimes, significantly outperforming MinVIS. Both MinMaxVIS and MinVIS are image-driven approaches. MinMaxVIS even achieves results comparable to or exceeding full-set MinVIS (100% labeled data) across multiple settings.

Component	mAP
Baseline (MinVIS (Huang et al., 2022))	58.5
+High-Precision Retrieval	60.3
+Selective Gradient Backpropagation	61.4
+Instance Association	62.2

TABLE 5.3. Main components of MinMaxVIS.

Selective Gradient Backpropagation. Table 5.4 compares different weighting strategies for selective gradient backpropagation introduced in Section 5.3.4. Among the tested variants, the truncation-weight strategy with a threshold of $\beta = 0.5$ achieves the best performance, demonstrating that filtering out low-confidence background queries is particularly effective.

Table 5.5 further analyzes the effect of varying the truncation threshold β. Lower values of β introduce additional noise from unreliable background queries, while excessively high values

Strategy	mAP
Confidence-Weight	60.7
Squared-Confidence-Weight	61.7
Truncation-Weight	62.2

TABLE 5.4. Ablation study on selective gradient backpropagation strategies proposed in Section 5.3.4.

Threshold (β)	0.0	0.3	0.5	0.7	0.9
mAP	61.4	61.2	62.2	61.6	60.1

TABLE 5.5. Impact of threshold β in truncation-weight strategy for selective gradient backpropagation.

limit the amount of background supervision. A threshold of $\beta = 0.5$ achieves the highest mAP, providing an optimal balance between noise suppression and sufficient background classification training.

Instance Association. Section 5.3.4 introduces an auxiliary decoder equipped with the proposed instance association loss. Each decoder layer produces $2 \times n$ auxiliary query features for an image pair, consisting of an original image and its augmented counterpart. Foreground query features corresponding to the same instance across the two images form positive pairs, while all remaining cross-image combinations involving a foreground query and a background query form negative pairs. By default, as shown in Figure 5.2(d.II), we employ a "Pull(F–F) + Push(F–B)" strategy: the model pulls positive (foreground–foreground) pairs closer in feature space and pushes negative (foreground–background) pairs apart.

We also explore an extended strategy, "Pull(F–F) + Push(F–B) + Push(B–B)", which additionally pushes similarities between background–background pairs across images. However, as reported in Table 5.6, this variant underperforms our default setting, likely because enforcing separation between background queries introduces unnecessary constraints that do not assist with instance association.

Strategy	mAP
Pull(F-F) + Push(F-B)	62.2
Pull(F-F) + Push(F-B) + Push(B-B)	60.2

TABLE 5.6. Study on the instance association strategies.

Layer ID	3	6	9
mAP	62.2	61.5	61.4

TABLE 5.7. Impact of main decoder layer selection on instance association performance.

Ratio	1:2	1:4	1:8	1:16
mAP	60.6	62.2	61.6	61.3

TABLE 5.8. Impact of different ratios of labeled to pseudo-labeled images within a training batch.

The auxiliary decoder receives as input the features produced by the $\overline{K}$-th layer of the main decoder. Table 5.7 examines different choices of $\overline{K}$. Features extracted from shallower layers yield the best performance, as these less semantic representations better preserve intra-class variability, which is crucial for instance association. Conversely, features from deeper layers are more heavily influenced by the classification objective, resulting in stronger inter-class discrimination but weaker intra-class cohesion. Since effective instance association requires maintaining similarity within the same instance category, shallower-layer features are more suitable for driving this process.

Ratio of Labeled to Pseudo-Labeled Images. Table 5.8 evaluates the effect of varying the proportion of labeled to pseudo-labeled images within each training batch. A ratio of $1{:}4$ yields the highest performance, indicating an effective balance between reliable supervision from labeled data and broad visual diversity from pseudo-labeled images.

Maximum Number of Pseudo-Labeled Images per Category. We introduce a high-precision retrieval strategy to identify images containing instances of target categories. In

W	200	500	1,000
mAP	60.7	61.3	62.2

TABLE 5.9. Study on maximum number of pseudo-labeled images per Category (W).

practice, we retain up to W pseudo-labeled images per category, selecting those with the highest confidence scores predicted by the preliminary segmentation model. Moreover, each retained instance must satisfy the threshold requirement $c_u \geq \tau$. Table 5.9 analyzes the effect of varying W.

Increasing the maximum number of pseudo-labeled images per category steadily improves performance, with the best mAP (62.2) achieved at $W = 1000$. Beyond this point, additional pseudo-labeled samples provide no further benefit, suggesting that the model effectively saturates once W reaches 1000. This indicates that retaining more images yields diminishing returns while potentially increasing computational cost.

Features for Instance Association. We introduce an auxiliary decoder that strengthens intra-class feature discrimination through the proposed instance association loss, operating alongside the main decoder. The auxiliary decoder is supervised by both the instance association loss and the segmentation loss, while the main decoder is trained with classification and segmentation losses. Gradients from the auxiliary decoder flow into the shallower layers of the main decoder. During inference, the auxiliary decoder is removed, and instance association relies solely on the features produced by the main decoder.

We also explore alternative feature choices for instance association beyond our default strategy. The auxiliary decoder consists of multiple decoder layers followed by an MLP projector. We evaluate using features from the final decoder layer as well as those produced by the MLP projector. The results are summarized in Table 5.10.

Instance association requires the model to reliably track the same instance across frames, which imposes two key requirements. First, instances belonging to the same category must be embedded close together in feature space, while instances from different categories must be

Feature	mAP
Last Decoder Layer (Main)	62.2
Last Decoder Layer (Auxiliary)	61.1
MLP Projector (Auxiliary)	60.7

TABLE 5.10. Feature analysis for instance association.

Category	Augmentation	Value
Color	Brightness	$[-32, +32]$
	Contrast	$[0.5, 1.5]$
	Hue	$[-18, +18]$
	Saturation	$[0.5, 1.5]$
Affine Transformation	Rotation	$[-15°, +15°]$
	Translation	10%
	Scale	$[0.8, 1.2]$
	Shear	$[-5°, +5°]$

TABLE 5.11. Data augmentations applied for generating image pairs.

well separated. This is enforced through the classification and segmentation losses applied to the main decoder features. Second, the model must differentiate between distinct instances within the same category, which is achieved through the gradients introduced by the instance association loss, encouraging intra-class variability. Together, these supervisory signals enable accurate instance association. As a result, features produced by the main decoder yield the best performance, since they benefit simultaneously from classification, segmentation, and instance association supervision.

Data Augmentations for Instance Association. Instance association training requires generating paired inputs by applying augmentations to each image, creating an original–augmented image pair. Table 5.11 categorizes the augmentations used into two types: color-based augmentations and affine transformations. Table 5.12 reports the impact of these two augmentation types on final performance.

Color	Affine Transformation	mAP
		59.8
✓		61.1
	✓	60.7
✓	✓	62.2

TABLE 5.12. Analysis of the effects of color and affine augmentations on the final performance.

5.4.4 Visualization

Score Distribution of Low-Confidence Background Queries. We introduce a selective gradient backpropagation strategy to reduce the impact of noise in pseudo-labeled images. For each pseudo-labeled sample, background queries are first identified, and only those with sufficiently high background confidence scores are used to supervise background classification. Low-confidence background queries (those with background scores below 0.5) may correspond either to true negatives or to false negatives (missed foreground instances). To avoid introducing noisy supervision, gradients from these low-confidence queries are detached during training. Figure 5.4 visualizes the score distribution of these low-confidence background queries.

Pseudo-Labeled Instances. Sections 5.3.2 and 5.3.3 describe how a preliminary segmentation model, trained on a small labeled dataset (e.g., 2% of YouTube-VIS 2019), is used to retrieve target-relevant instances from a large unlabeled dataset such as SA-1B. For each retrieved instance, the model produces a pseudo-label containing both a class prediction and a mask. Figure 5.5 presents two pseudo-labeled examples per category from YouTube-VIS 2019, all drawn from SA-1B. The proposed high-precision retrieval strategy ensures that, for the vast majority of retrieved samples, both the pseudo-class labels and pseudo-masks are highly accurate.

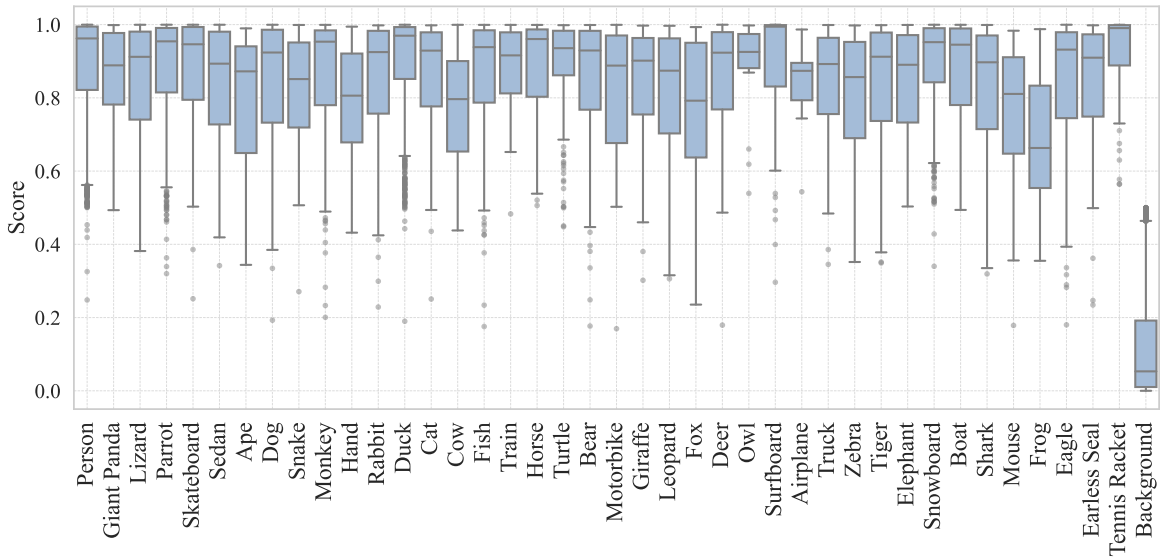


FIGURE 5.4. Score distribution of low-confidence background queries. The analysis is performed on all pseudo-labeled images from SA-1B. Each data point represents the maximum classification score of a specific low-confidence background query. For each category in YouTube-VIS 2019, we display the median, upper bound, upper quartile, lower bound, lower quartile, and outliers.

5.5 Chapter Summary

This chapter introduced **MinMaxVIS**, an image-driven, data-efficient framework for Video Instance Segmentation (VIS) that substantially reduces the reliance on densely labeled video annotations. We began by formulating the VIS problem, highlighting its core challenges including object detection, segmentation, and temporal association, as well as the prohibitive cost of frame-wise video annotation. Motivated by the success of semi-supervised image segmentation, we demonstrated how strong image-level segmentation priors can generalize effectively to videos, enabling VIS models to be trained using primarily static images.

MinMaxVIS achieves this goal through a three-stage pipeline. First, a preliminary segmentation model is trained on a small labeled image set from the target domain. Second, this model is used to perform *high-precision retrieval* from millions of unlabeled internet images, producing a compact, pseudo-labeled dataset with strong alignment to the target domain. Third, MinMaxVIS is trained jointly on the labeled and pseudo-labeled sets using two key techniques: (1) *selective gradient backpropagation* to mitigate noise from false negatives in

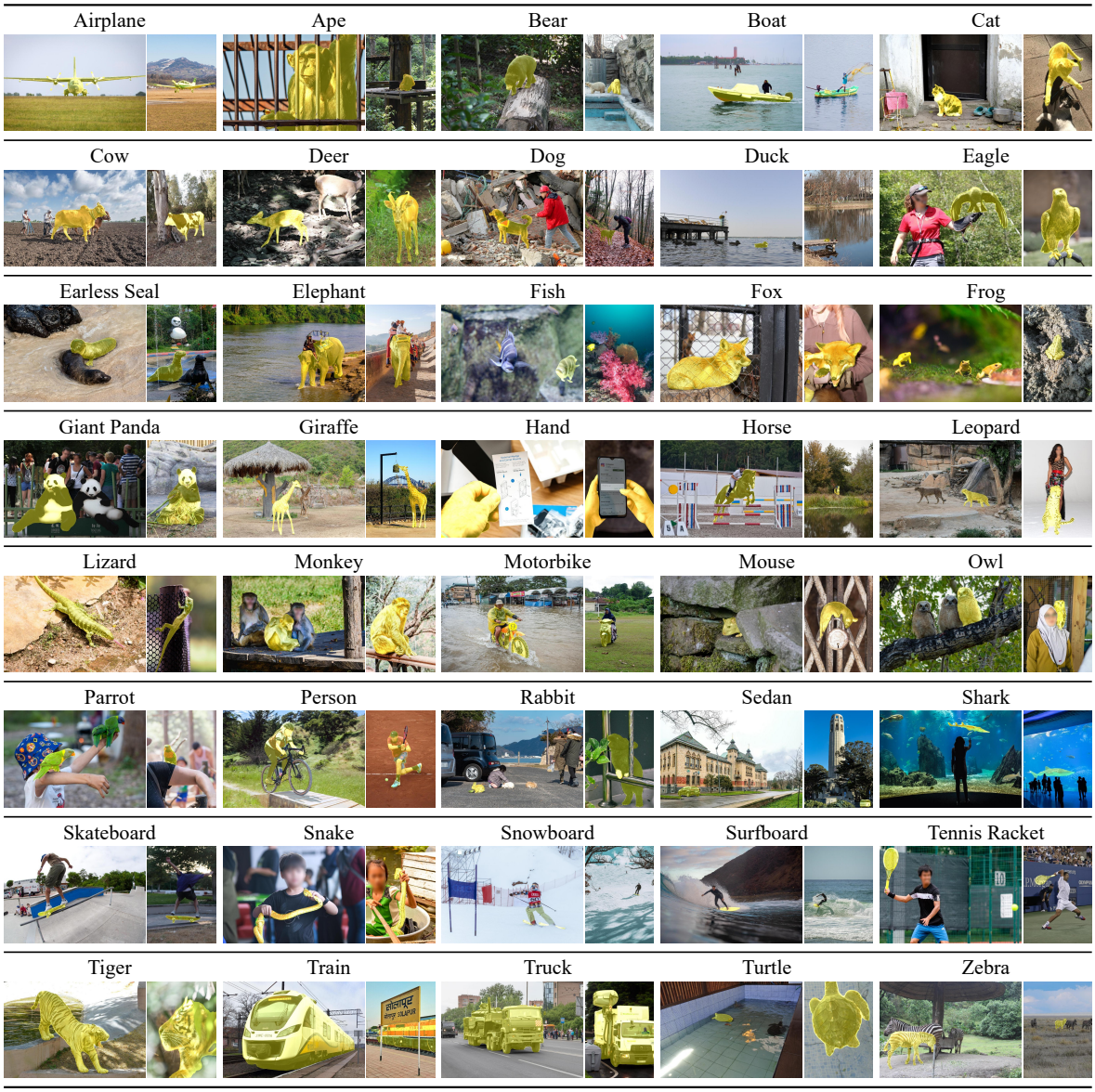


FIGURE 5.5. Visualization of the retrieved pseudo-labeled instances from the SA-1B dataset.

pseudo-labels, and (2) an *auxiliary instance association module* that simulates video-like consistency by enforcing feature alignment across augmented image pairs.

Through extensive experiments on YouTube-VIS 2019, YouTube-VIS 2021, and OVIS, we showed that MinMaxVIS delivers significant improvements over image-driven baselines and even surpasses the performance of fully supervised MinVIS, despite using only 1–10% of labeled data. Ablation studies further validated the importance of each component, including

retrieval strategy, gradient selection, instance association design, data augmentation, and the integration of labeled and pseudo-labeled samples.

Overall, this chapter demonstrates that high-quality VIS can be achieved without relying on densely annotated videos. By minimizing labeled data requirements while maximizing the utility of large-scale unlabeled images, MinMaxVIS establishes a powerful and scalable paradigm for data-efficient video instance segmentation.

CHAPTER 6

Benchmarking Language-Based Interfaces for Modern Visual Perception Models

---

Traditional visual perception models (as described in Chapters 3–5) typically address closed-set problems: users predefine a fixed set of categories (e.g., cat, person, dog), collect corresponding data and annotations, and then train a model accordingly. With the emergence of large language models, modern visual perception systems have increasingly shifted toward open-set settings, where natural language serves as a more flexible and expressive interface. In this paradigm, users can input free-form text to specify and localize visual entities in images, a task known as referring expression comprehension (REC).

However, existing REC benchmarks were largely designed for traditional visual perception models and suffer from several limitations, such as limited scale and overly short referring expressions, which are insufficient for rigorously evaluating the true capabilities of modern REC models. To address these issues, this chapter introduces HC-RefLoCo (Wei et al., 2024), a large-scale, human-centric benchmark designed to advance referring expression comprehension in the era of large multimodal models.

Section 6.1 presents the problem formulation of REC. Section 6.2 elaborates on the motivation behind the benchmark. Section 6.3 describes the construction of HC-RefLoCo in detail. Section 6.4 introduces comprehensive evaluation protocols, including accuracy at multiple IoU thresholds, scale-aware analysis, and subject-specific assessment. In Section 6.5, we benchmark 24 state-of-the-art models on HC-RefLoCo. Section 6.6 provides implementation details, while Section 6.7 presents extensive analyses on the benchmark. Finally, Section 6.8 summarizes this chapter.

6.1 Problem Formulation

Referring Expression Comprehension. Referring Expression Comprehension (REC) aims to localize a specific object instance in an image based on a natural-language description. Given an input image $\mathbf{I}$ and a referring expression E, the objective is to identify the target region $\mathbf{B}^*$ (e.g., a bounding box or a pixel-level mask) that corresponds to the entity described by E. Unlike traditional detection tasks that rely solely on visual cues, REC requires jointly reasoning over both modalities (vision and language) to resolve fine-grained distinctions, contextual relationships, and subtle semantic cues embedded in the expression. This makes REC a fundamental task for bridging natural language understanding and visual grounding.

Human-Centric REC. A particularly important sub-domain of REC focuses on grounding expressions referring to humans. Human-centric REC requires models to localize the correct person instance in scenes with large appearance variations, complex interactions, and diverse contextual cues. Human-related expressions frequently involve rich attributes (e.g., clothing, age, posture), human–object interactions (e.g., "the man holding the tennis racket"), social relationships (e.g., "the boy next to his sister"), and even temporal or action-related cues. Accurate human-centric grounding is critical for downstream applications such as human–robot interaction, surveillance analysis, AR/VR systems, and assistive technologies. However, existing benchmarks often simplify expressions and lack the diversity and complexity found in real human descriptions, limiting progress in this important direction.

REC in the LLM Era. With the advent of Large Language Models (LLMs), modern vision–language systems can interpret long, compositional, and multi-sentence descriptions. This significantly expands the expressive power of referring expressions, from short, template-like phrases to natural, human-authored paragraphs containing appearance details, interactions, actions, social context, and more. Yet, most existing REC benchmarks were designed in the pre-LLM era and contain short expressions that do not reflect the linguistic complexity LLM-powered models are capable of leveraging. As a result, current datasets no longer sufficiently challenge or measure the true multimodal reasoning capacity of modern models. To advance REC research in the LLM era, new benchmarks with long-form, diverse, and

fine-grained human-centric descriptions are needed to evaluate models' abilities to handle rich semantics, multi-sentence reasoning, and comprehensive grounding.

6.2 Motivation

Prior research in human-centric AI has largely focused on *single-modality* algorithms aimed at understanding, interacting with, or analyzing human behaviors and attributes. Representative tasks include face detection (Zhang et al., 2017a; Deng et al., 2019; Li et al., 2019; Tang et al., 2018; Xu et al., 2020; Zhang et al., 2017b; Ming et al., 2019) and recognition (Schroff et al., 2015; Taigman et al., 2014; Meng et al., 2021b; Kim et al., 2020, 2022), pedestrian detection (Wang et al., 2018; Zhang et al., 2018; Zheng et al., 2017; Chu et al., 2020) and re-identification (He et al., 2021; Luo et al., 2019; Eom and Ham, 2019; Liu et al., 2019b; Yang et al., 2023c), action recognition (Li et al., 2020; Feichtenhofer et al., 2019; Liu et al., 2020; Mazzia et al., 2022; Xu et al., 2022b), and pose estimation (Sun et al., 2019; Park et al., 2019; Zhang et al., 2020a; Wei et al., 2020), among others. While these tasks have driven substantial progress, they rely exclusively on visual inputs and thus cannot fully capture the rich linguistic semantics often required to describe human activities, appearances, or interactions.

The emergence of large multimodal models (LMMs), such as GPT-4V (OpenAI, 2023a,b,c) and Google Gemini (Team et al., 2023), has shifted the research landscape toward models capable of jointly understanding visual content and natural language. This shift marks a new era for human-centric AI, one in which multimodal reasoning becomes central. Referring expression comprehension (REC) (Yang et al., 2019b; Sun et al., 2022; Jin et al., 2023; Liu et al., 2019a,c; Yu et al., 2018; Liao et al., 2020; Luo et al., 2020a,b; Zhou et al., 2021; Zhu et al., 2022; Liu et al., 2023b) exemplifies this paradigm: given an image and a natural-language description, the goal is to localize the specific instance referred to in the text. Despite its importance, existing benchmarks provide limited support for evaluating REC in human-centered scenarios. To bridge this gap, we develop comprehensive benchmarks tailored to human-centric REC in the era of modern LMMs (Li et al., 2023d; Huang et al., 2024; Li et al.,

Dataset	Images	Instances	Annotations	Avg. Words	Vocab.	Instance Size	Subjects
HC-RefCOCO (Kazemzadeh et al., 2014)	1,519	3,754	10,771	3.4	2,251	114.0 - 603.2	-
HC-RefCOCO+ (Kazemzadeh et al., 2014)	1,519	3,754	10,908	3.3	2,702	114.0 - 603.2	-
HC-RefCOCOg (Mao et al., 2016)	1,521	2,669	5,253	8.9	2,891	89.7 - 610.5	-
HC-RefLoCo (Ours)	13,452	24,129	44,738	93.2	18,681	62.5 - 3720.7	6

TABLE 6.1. Comparison between human-centric (HC) referring expression comprehension benchmarks and the proposed HC-RefLoCo benchmark. Statistics for HC-RefCOCO, HC-RefCOCO+, and HC-RefCOCOg are derived from the combination of their respective validation and test sets. Vocab.: vocabulary. Avg.: average.

2024; Liu et al., 2023a; Zhu et al., 2023a; Chen et al., 2023c; Awadalla et al., 2023; Alayrac et al., 2022; Lin et al., 2023b; You et al., 2023; Rasheed et al., 2023; Peng et al., 2023).

Existing human-centric REC benchmarks are typically constructed by filtering general REC datasets such as RefCOCO (Kazemzadeh et al., 2014), RefCOCO+ (Kazemzadeh et al., 2014), and RefCOCOg (Mao et al., 2016) to retain only human-related samples. The resulting datasets, HC-RefCOCO, HC-RefCOCO+, and HC-RefCOCOg, contain only a modest number of test samples (Table 6.1). For example, HC-RefCOCO includes just 1,519 images and 10,771 referring expressions. Moreover, their annotations are extremely short, averaging 3.4, 3.3, and 8.9 words, respectively. With the significant language-understanding capability of modern LLMs such as GPT-4 (OpenAI, 2023a) and LLaMA (Touvron et al., 2023a), reasoning over such brief descriptions has become relatively trivial. As a result, there is an urgent need for more challenging, large-scale benchmarks that reflect the linguistic complexity contemporary AI models are capable of handling.

To address these limitations, we introduce a new benchmark, **HC-RefLoCo** (Human-Centric Referring Expression Comprehension with Long Context). Comprehensive statistics are provided in Table 6.1. Our benchmark offers five major advantages:

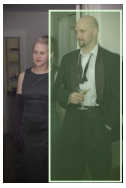
Large Scale. HC-RefLoCo contains 13,452 images with 24,129 instances and 44,738 referring expressions (annotations), providing a significantly broader evaluation set for human-centric REC.

The person is outfitted in a distinctive black and yellow full-body uniform, with the "DEWALT" brand emblazoned across the chest area. A black helmet, equipped with a visor, adorns his head, and he is frozen in a dynamic action stance. His involvement with a pit crew is suggested by the act of refueling a race car, which is indicated by the sizeable red fuel container he is deftly handling and utilizing.

(a) Each detailed annotation includes subject labels covering appearance, HOI, location, action, celebrity and OCR.

The individual is a man wearing a dark pinstripe suit with a white dress shirt and a gold-colored tie…The man's overall appearance, aside from his obscured face, conveys a professional demeanor.

(b) Appearance.



…He accessorizes his look with a pocket square that adds an element of sophistication to his ensemble. He is holding what appears to be a martini glass in one hand, suggesting he may be at a social gathering or event. The man has a bald head, and his posture exudes confidence as he stands with one hand in his pocket.

(c) Human-Object-Interaction.

…The firefighter is holding a framed certificate in front of her with both hands. Her blond hair is styled in a bun. She is standing in line with other firefighters, all in similar uniforms, inside what appears to be a fire station. She is the second person from the right in the row.

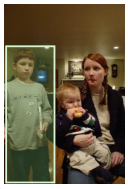
(d) Location.

The individual in the image is a man with a clean-shaven head, and he appears to be engaged in the activity of washing or styling someone's hair at a sink, which suggests he may be a hairdresser or barber. He is wearing a casual, long-sleeve, button-up shirt in a light blue color…

(e) Action.

Lionel Messi.

(f) Celebrity.



The individual appears to be a young boy dressed in a casual long-sleeve heather grey shirt with the brand "NIKE" prominently displayed across the chest. He is also wearing dark-colored pants. In his hands, he holds what seems to be a white rectangular object, possibly a game case or a DVD case…

(g) OCR.

FIGURE 6.1. (a) An Example from our HC-RefLoCo benchmark. For each target object, we provide a comprehensive and detailed text description, with an average length of 93.2 words. Each sentence within this description is classified into one of the following categories: (b) appearance, (c) human-object interaction, (d) location, (e) action, (f) celebrity, (g) optical character recognition, or None.

Long and Detailed Descriptions. We leverage GPT-4 to generate rich, multi-sentence descriptions for each target instance. Every annotation is manually reviewed and refined to eliminate hallucinations. The descriptions range from 15 to 241 words, averaging 93.2 words, and cover a vocabulary of 18,681 unique words. An example is shown in Figure 6.1(a).

Subject Labels. Each annotation consists of multiple sentences, and we manually label each sentence into one of seven subjects: appearance, human–object interaction (HOI), location, action, celebrity, optical character recognition (OCR), or None. As illustrated in Figure 6.1,

this enables fine-grained, subject-specific evaluation of REC models and supports deeper analysis of their linguistic reasoning capabilities.

Broader Coverage of Instance Scales. Compared with existing benchmarks, HC-RefLoCo spans a much wider range of instance scales. The square root of instance areas ranges from 62.5 to 3720.7 pixels, with an average of 313.8, yielding more diverse and realistic visual conditions.

Various Evaluation Protocols. Beyond standard $\text{Acc}_{0.5}$ metrics, we introduce:

- Accuracy at multiple IoU thresholds ($\text{Acc}_{0.5}$, $\text{Acc}_{0.75}$, $\text{Acc}_{0.9}$) and mean accuracy (mAcc),
- Performance breakdown across small, medium, and large instances,
- Subject-specific evaluation based on our manual sentence annotations.

In our experiments, we evaluate 24 training-unconstrained models, including GPT-4V, bounding box prediction models, and mask prediction models, across these protocols. With its large scale, detailed descriptions, subject-level labels, broad instance coverage, and comprehensive evaluation criteria, we hope HC-RefLoCo provides a strong foundation for advancing multimodal, human-centric REC research.

6.3 Benchmark Construction and Analysis

6.3.1 Benchmark Construction

Data Sources and Pre-Processing. The HC-RefLoCo benchmark is constructed from several publicly available object detection datasets, including the validation sets of COCO 2017 (Lin et al., 2014) and Objects365 (Shao et al., 2019), as well as the validation and test sets of OpenImages v7 (Krasin et al., 2017). For COCO 2017 and Objects365, we retain all instances labeled as "person," while for OpenImages v7 we keep instances labeled as "human." To ensure sufficient visual resolution, we filter out extremely small instances occupying less

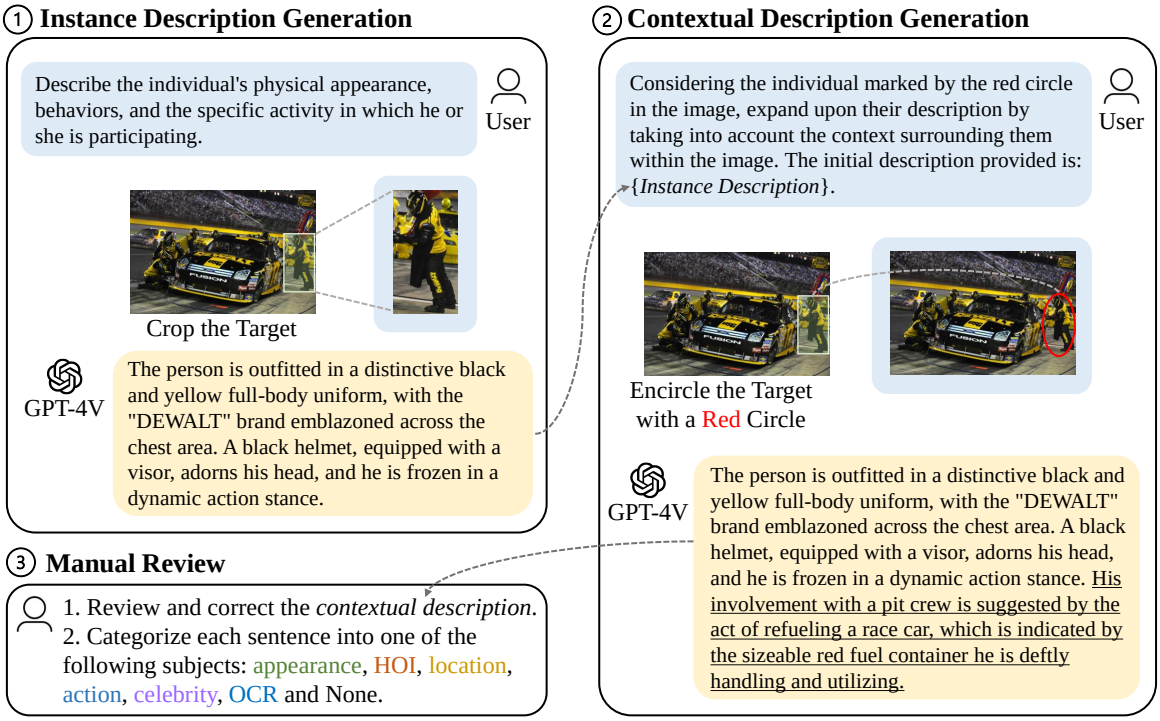


FIGURE 6.2. The process of generating a referring expression for each target instance. Inspired by recent studies on GPT-4V (Yang et al., 2023d), which demonstrate that GPT-4V can pay more attention to instances highlighted by a red circle within an image, we similarly encircle the target instance in red in Step-2.

than 1% of the image area. All original bounding box annotations are preserved without modification.

In addition, we curate a set of 367 celebrities from the LAION-5B dataset (Schuhmann et al., 2022). We collect images containing at least one of these celebrities and at least two people in total, and manually annotate the bounding boxes for the celebrity instances. This results in 3,520 additional images, each containing one manually labeled target instance.

In summary, HC-RefLoCo consists of 200 images with 419 instances from COCO, 4,772 images with 10,070 instances from Objects365, 4,960 images with 10,120 instances from OpenImages v7, and 3,520 images (and instances) sourced from LAION-5B.

Referring Expression Generation. Figure 6.2 illustrates our pipeline for generating a referring expression (i.e., a description) for each target instance. Given an image and a specified instance, the process consists of three stages:

(1) We first use GPT-4V to produce an instance-level description based on the cropped target region, following the prompt described in Section 6.6.1.
(2) Next, we input the full image into GPT-4V to enrich the initial description with contextual information surrounding the target instance, using the prompt provided in Section 6.6.2.
(3) Finally, we manually review and refine every generated description to correct mistakes, particularly hallucinations, and to ensure that each referring expression accurately and uniquely identifies the intended instance.

Annotation Expansion. At this stage, our benchmark contains 13,452 images with 24,129 instances, each paired with a single referring expression. To enrich the dataset, we leverage GPT-4's strong language generation abilities to rewrite every referring expression, following the prompt described in Section 6.6.3. This step effectively doubles the number of annotations. We then manually review all rewritten expressions, removing those that are inaccurate, ambiguous, or insufficiently discriminative to ensure that each annotation uniquely identifies its target instance. After this refinement, the final benchmark comprises 13,452 images and 44,738 high-quality annotations for 24,129 instances. Labeling a single image with GPT-4 via the OpenAI API takes approximately 1.5 seconds. The total processing time is about 10.05 hours.

Subject Labels. We manually assign each sentence in every referring expression to one of seven subjects: appearance, human–object interaction (HOI), location, action, celebrity, optical character recognition (OCR), or None. The detailed labeling criteria for each subject category are provided in Section 6.6.5. Manually labeling a single entity takes approximately 1.2 seconds, resulting in a total time of about 14.9 hours.

Data Format. Each instance $\boldsymbol{I}$ is associated with an image $\boldsymbol{X}$, a bounding box $\boldsymbol{b} = \{x, y, w, h\}$—where (x, y) denotes the top-left corner and w, h represent the width and

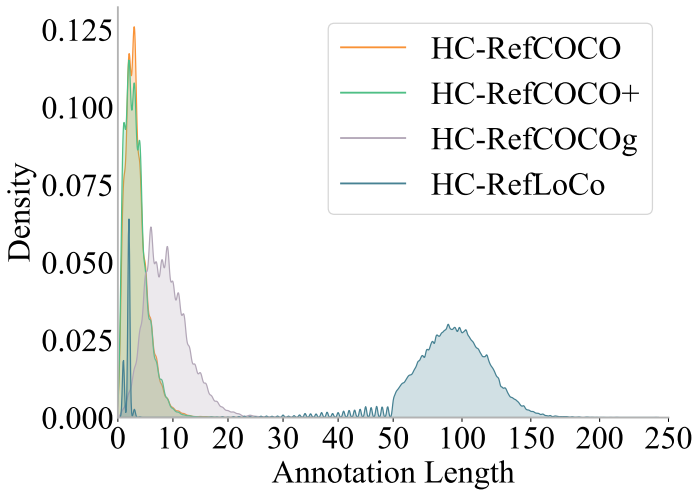


FIGURE 6.3. Density distribution of the annotation length.

height—and a referring expression $\boldsymbol{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_N\}$ consisting of N sentences. Each sentence $\boldsymbol{s}_i$ is further annotated with a subject label $\boldsymbol{l}_i$.

6.3.2 Analysis

Annotation Length. Figure 6.3 presents the annotation length distributions across HC-RefCOCO, HC-RefCOCO+, HC-RefCOCOg, and our HC-RefLoCo benchmark. Unlike the three existing benchmarks which exhibit sharp peaks in the 4–8 word range, HC-RefLoCo shows a distinctly different pattern, with a pronounced peak around 100 words. Furthermore, HC-RefLoCo spans a much broader range of lengths, roughly from 50 to 150 words, highlighting the significantly richer and more detailed nature of our referring expressions.

Sentence Length. Annotations in the HC-RefLoCo benchmark consist of multiple sentences. Figure 6.4 shows the distribution of sentence lengths across four benchmarks, computed over all sentences from all annotations. HC-RefLoCo exhibits a clear peak around 18–20 words per sentence, reflecting the richer linguistic structure of our descriptions. In contrast, HC-RefCOCO, HC-RefCOCO+, and HC-RefCOCOg primarily contain single-sentence annotations, typically only 4–8 words long.

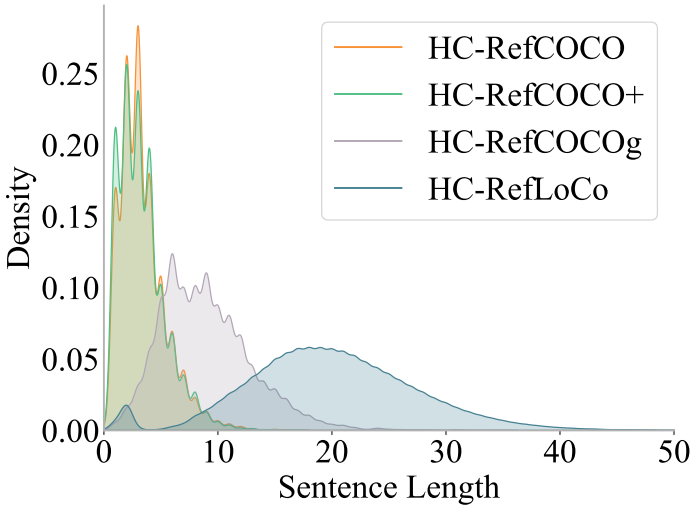


FIGURE 6.4. Density distribution of the sentence length.

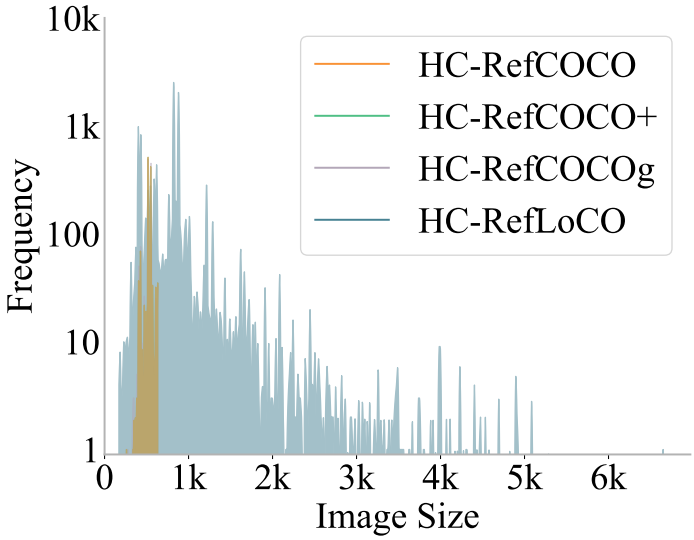


FIGURE 6.5. Distribution of image size.

Image Size. Figure 6.5 compares the image size distributions of our benchmark with those of HC-RefCOCO, HC-RefCOCO+, and HC-RefCOCOg. As the three existing benchmarks all originate from COCO, their size distributions largely overlap. In contrast, HC-RefLoCo spans a much wider range of image resolutions, reflecting the diversity of its underlying data sources.

Instance Size. Figure 6.6 illustrates the instance size distributions across the four benchmarks. HC-RefLoCo covers a significantly broader range of instance scales compared to existing

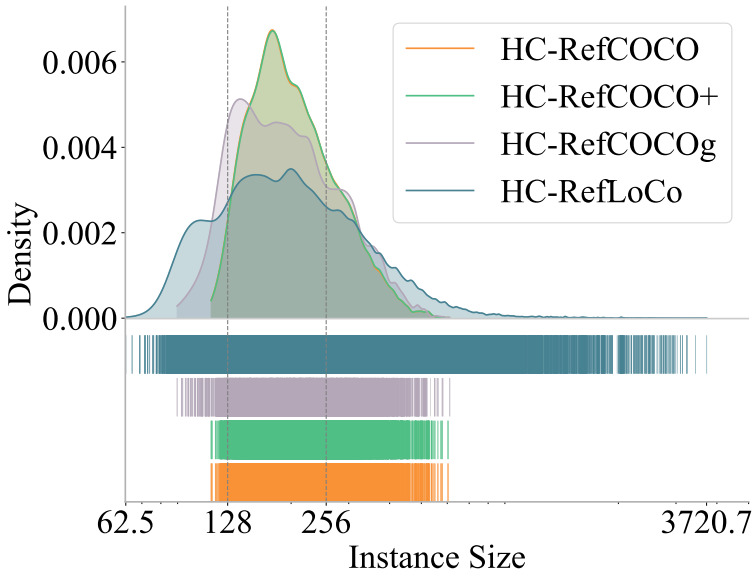


FIGURE 6.6. Density distribution of instance size.

datasets. The square root of instance areas in our benchmark ranges from 62.5 to 3720.7, with an average value of 313.8.

Annotation and Image Count per Subject. In HC-RefLoCo, each annotation consists of a multi-sentence referring expression, with every sentence assigned a subject label. As shown in Figure 6.7, we report, for each subject category, the number of annotations that contain at least one sentence belonging to that subject.

Instance Center. Figure 6.8 shows the spatial distribution of instance centers for our HC-RefLoCo benchmark compared to the combined HC-RefCOCO, HC-RefCOCO+, and HC-RefCOCOg datasets. The instance centers in HC-RefLoCo exhibit a noticeably more uniform distribution across the image plane.

Bias and Ethical Analysis. Due to the nature of image distributions, we observe that the primary attributes of interest approximately follow Gaussian-like distributions. For example, the median instance size is 205, the median instance center is around (0.5, 0.5), the median image resolution is approximately 1K, and the median sentence length is about 20 words. After a careful review of potential ethical considerations, we did not identify any significant ethical issues.

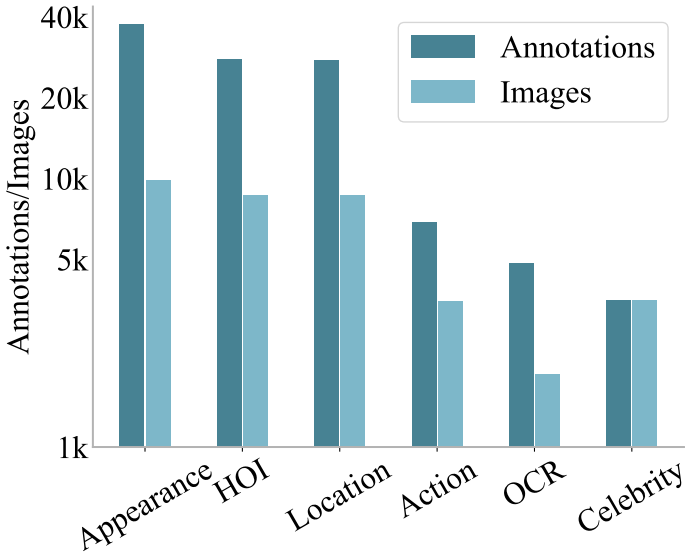


FIGURE 6.7. Annotation and image number for each subject.

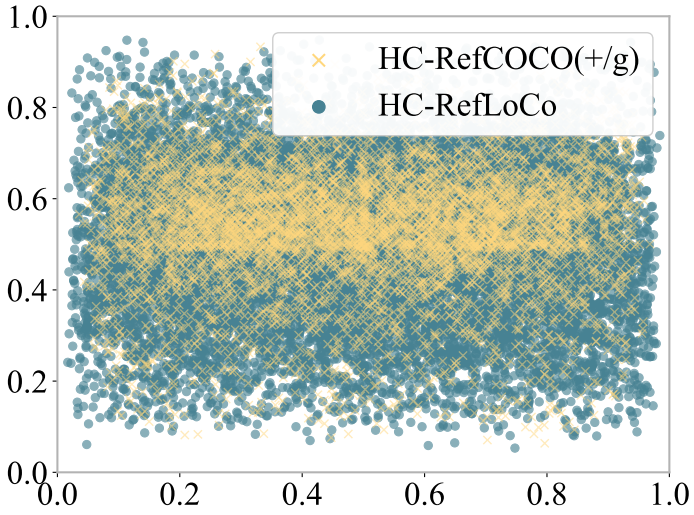


FIGURE 6.8. Distribution of instance center.

6.4 Evaluation

Benchmark Usage. Modern REC models are typically trained on large and diverse datasets. For instance, SPHINX (Lin et al., 2023b) utilizes a mixture of 16 unimodal and multimodal datasets comprising millions of samples. HC-RefLoCo is designed to evaluate such advanced models without restricting the sources or scale of training data.

We split the benchmark into two subsets: a validation set containing 4,000 images, 7,190 instances, and 13,360 annotations (30% of the data), and a test set containing 9,452 images, 16,939 instances, and 31,378 annotations (70%). Although these splits are provided, we encourage researchers to use the combined validation and test sets for evaluation, especially in the era of large multimodal models, where training is typically unconstrained and leverages extensive external data.

Evaluation Protocols. Under the standard REC evaluation protocol, an instance is considered correctly localized if the IoU between the predicted and ground-truth bounding boxes exceeds 0.5, and accuracy (denoted as $\text{Acc}_{0.5}$) is used as the evaluation metric. To more thoroughly assess model performance, we introduce three enhanced evaluation protocols:

- We report $\text{Acc}_{0.75}$ and $\text{Acc}_{0.9}$ in addition to $\text{Acc}_{0.5}$, as well as the mean accuracy (mAcc), computed as the average of $\text{Acc}_{0.5}$ through $\text{Acc}_{0.95}$ at 0.05 intervals.
- Based on the subject distribution shown in Figure 6.7, we perform *per-subject* evaluation, using mAcc to assess model performance across different linguistic subject categories.
- To evaluate robustness across instance sizes, we report mAcc_s, mAcc_m, and mAcc_l for small, medium, and large instances. Instance size is defined as the square root of its bounding-box area. We categorize instances as small (< 128), medium ($[128, 256]$), and large (> 256).

6.5 Experiment

Main Results. We evaluate a total of 24 state-of-the-art models, grouped into two categories based on their output format. The first category includes models that predict bounding boxes: GPT-4V (OpenAI, 2023a,b,c), GroundingGPT (Li et al., 2024), Ferret (You et al., 2023), MiniGPT4-v2 (Zhu et al., 2023a; Chen et al., 2023a), KOSMOS-2 (Peng et al., 2023), Shikra (Chen et al., 2023b), OFA (Wang et al., 2022a), Qwen-VL (Bai et al., 2023), CogVLM (Wang et al., 2023c), Lenna (Wei et al., 2023), ONE-PEACE (Wang et al., 2023b), and SPHINX (Gao et al., 2024; Lin et al., 2023b). The second category comprises models

Model	Val+Test				Val	Test
	$Acc_{0.5}$	$Acc_{0.75}$	$Acc_{0.9}$	mAcc	mAcc	mAcc
GPT-4V (OpenAI, 2023a,b,c)	17.4	2.6	0.3	5.5	5.5	5.6
GroundingGPT (Li et al., 2024)	56.6	27.2	5.3	29.8	30.0	29.8
Ferret 7B (You et al., 2023)	44.9	32.6	11.7	30.0	30.6	29.7
Ferret 13B (You et al., 2023)	52.9	38.5	15.6	35.7	35.9	35.6
MiniGPT4-v2 (Chen et al., 2023a)	47.1	31.7	11.6	30.3	30.7	30.1
KOSMOS-2 (Peng et al., 2023)	45.3	38.0	20.0	34.1	34.2	34.0
Shikra (Chen et al., 2023b)	56.8	35.6	10.3	34.4	34.6	34.3
OFA (Wang et al., 2022a)	48.4	37.0	21.7	35.3	35.2	35.3
OFA-Large(Wang et al., 2022a)	70.5	61.6	44.0	58.1	57.9	58.1
Qwen-VL (Bai et al., 2023)	67.9	56.8	34.8	52.8	53.1	52.6
CogVLM (Wang et al., 2023c)	66.0	59.6	43.8	55.8	56.3	55.5
Lenna (Wei et al., 2023)	68.8	63.5	51.6	60.6	60.5	60.7
ONE PEACE (Wang et al., 2023b)	79.3	69.0	43.8	63.1	63.4	62.9
SPHINX-MoE (Lin et al., 2023b)	76.3	57.7	21.8	52.5	52.7	52.4
SPHINX (Lin et al., 2023b)	77.5	61.0	27.0	55.4	55.8	55.2
SPHINX-1k (Lin et al., 2023b)	80.7	68.6	41.1	63.0	63.0	62.9
SPHINX-MoE-1k (Lin et al., 2023b)	**85.8**	**77.3**	53.7	71.4	71.5	71.4
SPHINX-v2-1k (Lin et al., 2023b)	84.1	77.1	**56.2**	**71.7**	**71.6**	**71.7**
PixelLM 7B† (Zhongwei Ren, 2023)	38.5	24.7	11.8	24.5	24.6	24.4
PixelLM 13B† (Zhongwei Ren, 2023)	63.6	46.6	25.8	44.6	45.0	44.4
LISA-explanatory† (Lai et al., 2023)	47.6	37.6	27.0	36.7	36.7	36.7
LISA† (Lai et al., 2023)	52.4	42.1	31.3	41.1	41.1	41.1
PSALM† (Zhang et al., 2024a)	61.7	53.4	40.2	51.1	51.4	51.0
GlaMM† (Rasheed et al., 2023)	**66.1**	**56.9**	**44.2**	**55.0**	**54.9**	**55.0**

TABLE 6.2. Performance evaluation across 24 models on our HC-RefLoCo benchmark. Models indicated with a † generate mask outputs, which we convert into tight bounding boxes to enable evaluation. Refer to Section 6.6.6 for the details of each model. NVIDIA A100 (80G) GPUs are used for evaluation.

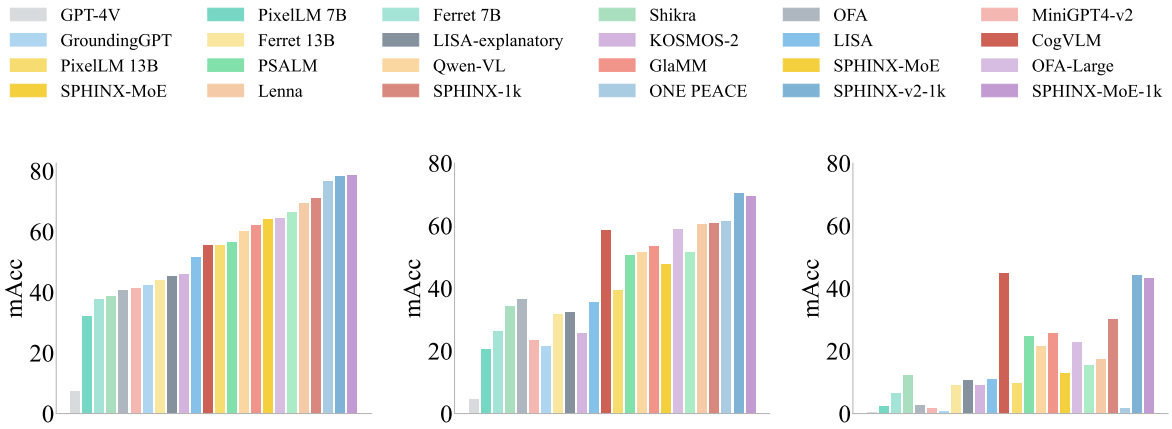
that output segmentation masks: PixelLM (Zhongwei Ren, 2023), LISA (Lai et al., 2023), PSALM (Zhang et al., 2024a), and GlaMM (Rasheed et al., 2023). The evaluation prompt used for GPT-4V is detailed in Section 6.6.4. For mask-output models, we convert the predicted masks into tight bounding boxes for consistency. The full performance comparison is provided in Table 6.2. Evaluating all models requires approximately 146 GPU hours on a single NVIDIA A100 GPU.

Model	Appearance	HOI	Celebrity	OCR	Action	Location
GPT-4V (OpenAI, 2023a,b,c)	5.0	5.1	12.0	5.1	3.6	4.6
GroundingGPT (Li et al., 2024)	27.3	27.5	61.4	25.8	21.3	23.0
Ferret 7B (You et al., 2023)	27.9	27.9	57.0	27.0	24.2	25.1
Ferret 13B (You et al., 2023)	33.9	34.4	58.5	33.5	28.8	30.9
MiniGPT4-v2 (Chen et al., 2023a)	27.4	27.5	66.2	24.6	22.6	22.7
KOSMOS-2 (Peng et al., 2023)	31.5	32.9	65.8	31.5	27.9	28.2
Shikra (Chen et al., 2023b)	32.7	32.5	55.9	29.7	30.6	31.7
OFA (Wang et al., 2022a)	35.2	35.3	36.8	35.2	32.3	32.2
OFA Large(Wang et al., 2022a)	58.4	58.3	56.0	56.9	55.1	55.2
Qwen-VL (Bai et al., 2023)	52.7	53.1	56.1	50.9	47.8	49.3
CogVLM (Wang et al., 2023c)	54.8	53.6	66.9	50.3	55.9	55.2
Lenna (Wei et al., 2023)	61.8	62.3	50.6	61.6	56.5	57.2
ONE PEACE (Wang et al., 2023b)	62.1	63.5	**75.4**	62.1	55.8	56.6
SPHINX-MoE (Lin et al., 2023b)	51.6	52.9	64.4	52.1	45.5	47.9
SPHINX (Lin et al., 2023b)	54.2	55.1	70.4	53.1	49.4	50.8
SPHINX-1k (Lin et al., 2023b)	62.7	63.3	66.0	61.7	59.0	59.6
SPHINX-MoE-1k (Lin et al., 2023b)	71.8	72.4	67.7	72.0	67.9	68.9
SPHINX-v2-1k (Lin et al., 2023b)	**72.4**	**73.0**	64.1	**72.3**	**68.7**	**69.6**
PixelLM 7B† (Zhongwei Ren, 2023)	23.3	22.6	39.6	23.4	22.4	20.9
PixelLM 13B† (Zhongwei Ren, 2023)	43.8	44.9	54.8	44.0	38.9	40.3
LISA-explanatory† (Lai et al., 2023)	34.1	32.5	69.6	30.8	33.1	31.2
LISA† (Lai et al., 2023)	38.8	38.0	**70.2**	36.7	37.1	35.0
PSALM† (Zhang et al., 2024a)	51.7	51.6	47.3	**52.2**	48.3	49.5
GlaMM† (Rasheed et al., 2023)	**54.0**	**53.4**	68.7	51.7	**51.3**	**51.3**

TABLE 6.3. Per-subject evaluation across 24 models on our HC-RefLoCo. We report mAcc for each set.

Per-Subject Evaluation. We partition our benchmark into six subsets based on the subjects of appearance, human–object interaction (HOI), location, action, celebrity, and optical character recognition (OCR). This setup enables detailed, subject-specific analysis of model performance across different linguistic and semantic dimensions. Table 6.3 summarizes the mAcc for each subset. SPHINX-v2-1k (Lin et al., 2023b) delivers the strongest overall performance across most subjects, while ONE-PEACE (Wang et al., 2023b) achieves particularly strong results on the celebrity subset.

Scale-Aware Evaluation. Figure 6.9 evaluates model performance across three instance-size categories: small, medium, and large. Instance size is defined as the square root of the bounding-box area, with small instances below 128, medium instances between 128 and



(A) Large instances. (B) Medium instances. (C) Small instances.

FIGURE 6.9. Scale-aware evaluation. Models are sorted in ascending order based on their performance on large instances. We use mAcc as the evaluation metric.

256, and large instances above 256. As expected, most models experience performance degradation as instance size decreases. Among all evaluated models, CogVLM (Wang et al., 2023c) demonstrates the strongest robustness across scales.

Effects of Using Detailed and Contextual Annotations. Each annotation in HC-RefLoCo consists of multiple sentences, with every sentence assigned a subject label. To study the impact of detailed contextual descriptions, we perform per-subject evaluations under two settings: 1) using the full annotations, and 2) retaining only the sentences corresponding to each subject while discarding all others.

Figure 6.10 evaluates five models, KOSMOS-2 (Peng et al., 2023), Ferret 7B (You et al., 2023), MiniGPT4 v2 (Chen et al., 2023a), SPHINX (Lin et al., 2023b), and Shikra (Chen et al., 2023b), which employ different language encoders: KOSMOS 1.3B (Huang et al., 2024), Vicuna 7B (Chiang et al., 2023), LLaMA2 Chat 7B (Touvron et al., 2023b), LLaMA2 13B (Touvron et al., 2023b), and LLaMA 7B (Touvron et al., 2023a), respectively. For most subjects, SPHINX and Shikra achieve higher accuracy when the full multi-sentence descriptions are used, likely due to the strong language modeling capabilities of their LLaMA-based encoders. In contrast, MiniGPT4 v2 (Chen et al., 2023a) experiences a notable drop in

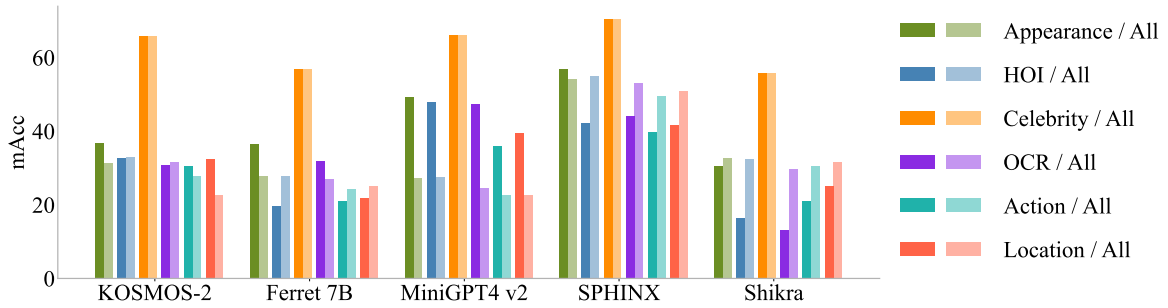


FIGURE 6.10. Per-subject evaluation under two scenarios: 1) using the original annotations (denoted as "All"); 2) retaining only sentences that correspond to the specific subject while discarding the rest for each annotation.

performance when more contextual information is included, suggesting difficulty in aligning long-form text with visual content.

To further explore the effect of annotation length, we create three additional variants by randomly selecting 1, 3, or 5 sentences from each annotation. In Figure 6.11 of Section 6.7, we evaluate the same models on these subsets alongside the original HC-RefLoCo dataset. The results show that different models perform best on different annotation lengths, revealing a trade-off: while longer descriptions provide richer context, models may struggle to reliably associate extended text with the correct visual instance.

6.6 Implementation Details

6.6.1 Prompt for Instance Description Generation

You are an advanced referring expression generator tasked with crafting a detailed and precise description of a person in an image. To achieve this, please adhere to the following guidelines:

(1) Highlight unique characteristics that make the person distinctive.
(2) Provide a comprehensive description of the person's overall appearance.
(3) Mention any interactions the person has with objects or other people.
(4) Include any visible text on the individual, such as text on clothing.

(5) Detail any specific activities the person is engaged in.
(6) Describe the person's location within the scene.
(7) When multiple individuals have similar appearances, use their relative positions for identification, such as "the first person on the left" or "the individual in the middle of the second row".

Input image: <*Cropped Image*>.

6.6.2 Prompt for Contextual Description Generation

You are an advanced referring expression generator tasked with crafting a detailed and precise description of a person highlighted by a red circle in an image. An initial description is provided as a reference. The description is <*Instance-Level Description*>. To achieve this, please adhere to the following guidelines:

(1) Highlight unique characteristics that make the person distinctive.
(2) Provide a comprehensive description of the person's overall appearance.
(3) Mention any interactions the person has with objects or other people.
(4) Include any visible text on the individual, such as text on clothing.
(5) Detail any specific activities the person is engaged in.
(6) Describe the person's location within the scene.
(7) When multiple individuals have similar appearances, use their relative positions for identification, such as "the first person on the left" or "the individual in the middle of the second row".

Input image: [*Raw Image*].

Model	Text Encoder	Vision Encoder
GPT-4V (OpenAI, 2023a,b,c)	-	-
GroundingGPT (Li et al., 2024)	LEGO-7B (Li et al., 2024)	CLIP-ViT-L/14 (Radford et al., 2021)
Ferret (You et al., 2023)	Vicuna-7B/13B (Chiang et al., 2023)	CLIP-ViT-L/14 (Radford et al., 2021)
MiniGPT4-v2 (Chen et al., 2023a)	LLaMa 2 Chat-7B (Touvron et al., 2023b)	EVA (Fang et al., 2023)
KOSMOS-2 (Peng et al., 2023)	KOSMOS-1.3B (Huang et al., 2024)	CLIP-ViT-L/14 (Radford et al., 2021)
Shikra (Chen et al., 2023b)	LLaMA-7B (Touvron et al., 2023a)	CLIP-ViT-L/14 (Radford et al., 2021)
OFA (Wang et al., 2022a)	BART_{Base}-140M (Lewis et al., 2019)	ResNet50 (He et al., 2016)
OFA-Large (Wang et al., 2022a)	BART_{Large}-400M (Lewis et al., 2019)	ResNet152 (He et al., 2016)
Qwen-VL (Bai et al., 2023)	Qwen-7B (Bai et al., 2023)	ViT-bigG (Schuhmann et al., 2022)
Lenna (Wei et al., 2023)	LLaVA-7B (Liu et al., 2023a)	Swin-L (Liu et al., 2021b)
ONE PEACE (Wang et al., 2023b)	Shared Causal Transformer Decoder-4B (Wang et al., 2023b)	
SPHINX-MoE (Lin et al., 2023b)	Mixtral-8x7B (Jiang et al., 2024)	Hybrid†
SPHINX (Lin et al., 2023b)	LLaMA 2-13B (Touvron et al., 2023b)	Hybrid†
SPHINX-1k (Lin et al., 2023b)	LLaMA 2-13B (Touvron et al., 2023b)	Hybrid†
SPHINX-MoE-1k (Lin et al., 2023b)	Mixtral-8x7B (Jiang et al., 2024)	Hybrid†
SPHINX-v2-1k (Lin et al., 2023b)	LLaMA 2-13B (Touvron et al., 2023b)	Hybrid†
PixelLM (Zhongwei Ren, 2023)	LLaVA-7B/13B (Liu et al., 2023a)	CLIP-ViT-L/14 (Radford et al., 2021)
LISA (Lai et al., 2023)	LLaVA 2-13B (Liu et al., 2023a)	SAM-ViT-H (Kirillov et al., 2023b)
PSALM (Zhang et al., 2024a)	Phi 1.5-1.3B (Li et al., 2023e)	Mask2former-Siwn-B (**?**)
GlaMM (Rasheed et al., 2023)	Vicuna-7B (Chiang et al., 2023)	SAM-ViT-H (Kirillov et al., 2023b)

TABLE 6.4. Architecture of each model. †: a hybrid vision encoder encompassing CLIP-ViT-L/14 (Radford et al., 2021), CLIP-ConvNeX (Radford et al., 2021), DINOv2-ViT (Oquab et al., 2023) and Q-Former (Zhang et al., 2023b).

6.6.3 Prompt for Annotation Expansion

The following paragraph should be rewritten while retaining the essential information. Different expressions should be used, and the paragraph may be reorganized if necessary. The paragraph should not be altered merely by converting the passive voice to active voice or vice versa.

6.6.4 Prompt for GPT-4V Evaluation

Given an image and a referring expression describing an instance visible in the image, the task is to identify the specific instance and output a bounding box in the format (x, y, h, w), where (x, y) represents the top-left corner and (h, w) denotes the height and width. Ensure the

response includes only the coordinates as described, without any additional text, characters, or spaces.

Input image: [*Raw Image*]

Description: [*Referring Expression of a Target Instance*]

6.6.5 Labeling Criteria for Sentence-Level Annotations

As described in Section 6.3.1, each referring expression in HC-RefLoCo is composed of multiple sentences, and every sentence is manually assigned to one of six subject categories: appearance, human–object interaction (HOI), location, action, celebrity, or optical character recognition (OCR). The labeling criteria for each subject are defined as follows:

- *Appearance.* Sentences describing the physical or visual attributes of the person.
- *HOI.* Sentences detailing interactions between the person and surrounding objects.
- *Location.* Sentences indicating the setting or place where the person is situated.
- *Action.* Sentences describing the activities, behaviors, or movements of the person.
- *Celebrity.* Sentences identifying the person as a public figure or well-known individual.
- *OCR.* Sentences referencing textual content associated with the person that can be read or recognized.

6.6.6 Model Cards

Table 6.4 presents the detailed architecture of each model evaluated in this work.

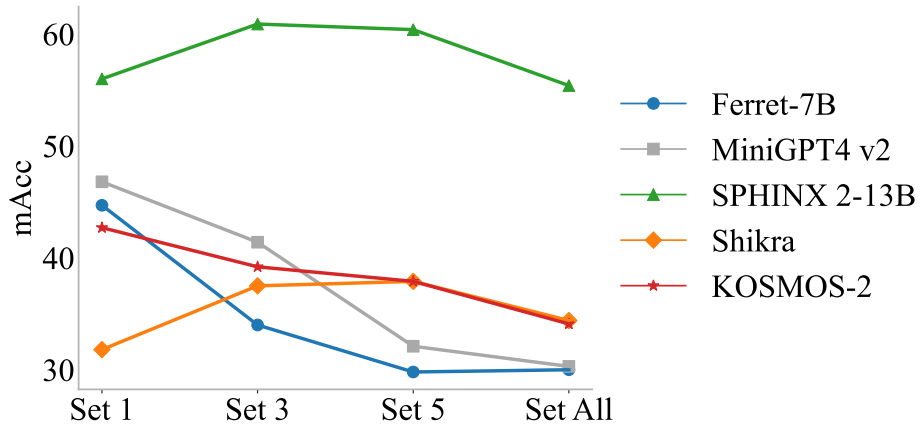


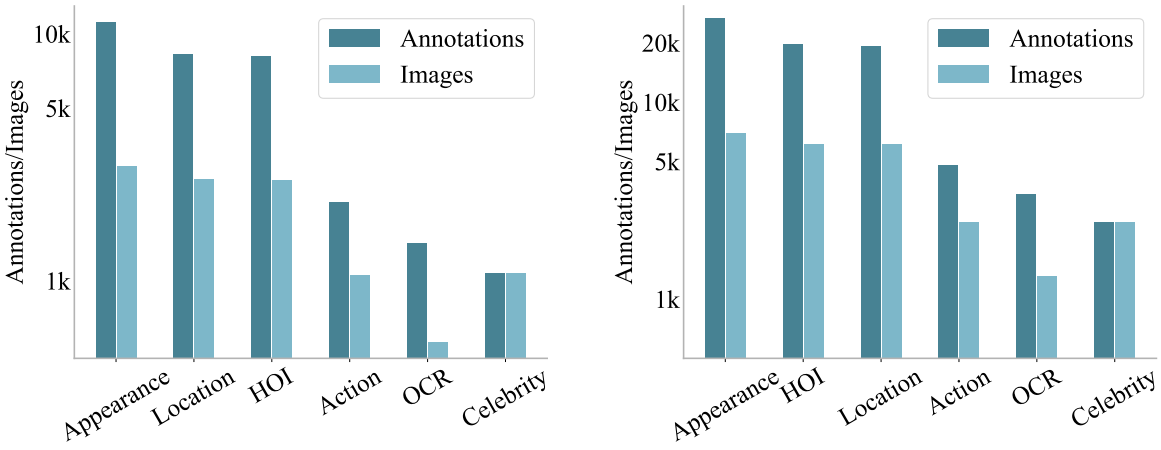
FIGURE 6.11. Alongside the original benchmark, we create three additional sets by randomly selecting 1, 3 and 5 sentences from each annotation. These sets are referred to as "Set-1," "Set-3," and "Set-5," respectively. We report mAcc on the four sets across five models.

6.7 Analysis

Using Randomly Selected Sentences as Referring Expressions. We construct three additional subsets by randomly selecting 1, 3, or 5 sentences from each annotation. Figure 6.11 reports the performance of five models on these subsets.

Statistics of Validation and Test Sets. As described in Section 6.4, our benchmark is split into validation and test sets. Figure 6.12 shows the number of annotations and images corresponding to each subject category in both subsets.

Word Frequency. Figure 6.13 illustrates the 20 most frequently used nouns in annotations across four different benchmarks. In our benchmark, the top 20 nouns are “person”, “shirt”, “hair”, “man”, “child”, “jacket”, “posture”, “group”, “event”, “image”, “stance”, “woman”, “question”, “clothing”, “presence”, “text”, “trousers”, “environment”, “part” and “sleeves”. In Figure 6.14, we present the 20 most frequently used verbs for each benchmark. In our benchmark, the top 20 verbs are “wearing”, “appears”, “seems”, “sleeved”, “holding”, “suggesting”, “indicating”, “suggests”, “clad”, “paired”, “featuring”, “located”, “stands”, “complemented”, “indicated”, “participating”, “depicted”, “evidenced”, “donned” and “includes”.



(A) Validation set. (B) Test set.

FIGURE 6.12. The number of annotations and images for each subject in the validation set and the test set.

Male	Female	Unrecognizable
46.36%	39.29%	14.35%

TABLE 6.5. Gender diversity analysis.

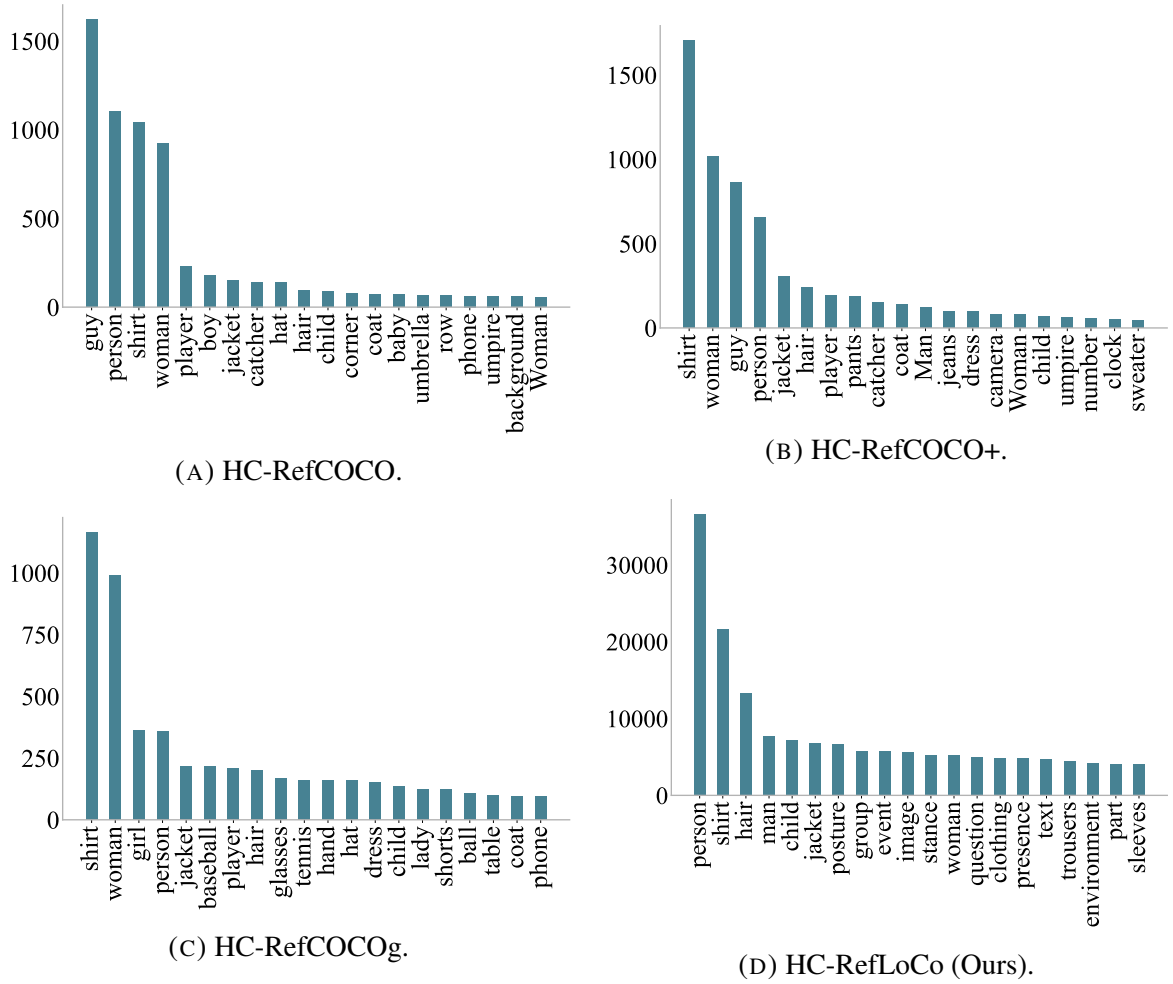
Child (0-12)	Adolescence (13–18)	Adult (19–59)	Senior Adult ($\geq$ 60)	Unrecognizable
8.72%	12.39%	51.61%	12.93%	14.35%

TABLE 6.6. Age diversity analysis.

Human Diversity. We employ MiVOLO[1] to predict the gender and age of each individual, followed by manual verification and correction. The final statistics are summarized in Tables 6.5 and 6.6, where "unrecognizable" denotes instances in which the face is obscured, blurred, or otherwise not suitable for reliable prediction.

Scene Diversity. We first collect the 365 scene categories from the Places365 benchmark, one of the largest scene recognition datasets. Using GPT-4o, we group these categories into 20 broader scene types. Each image in our benchmark is then processed with GPT-4o to predict its scene category, followed by manual verification and correction. The resulting

[1]https://github.com/wildchlamydia/mivolo



(A) HC-RefCOCO.

(B) HC-RefCOCO+.

(C) HC-RefCOCOg.

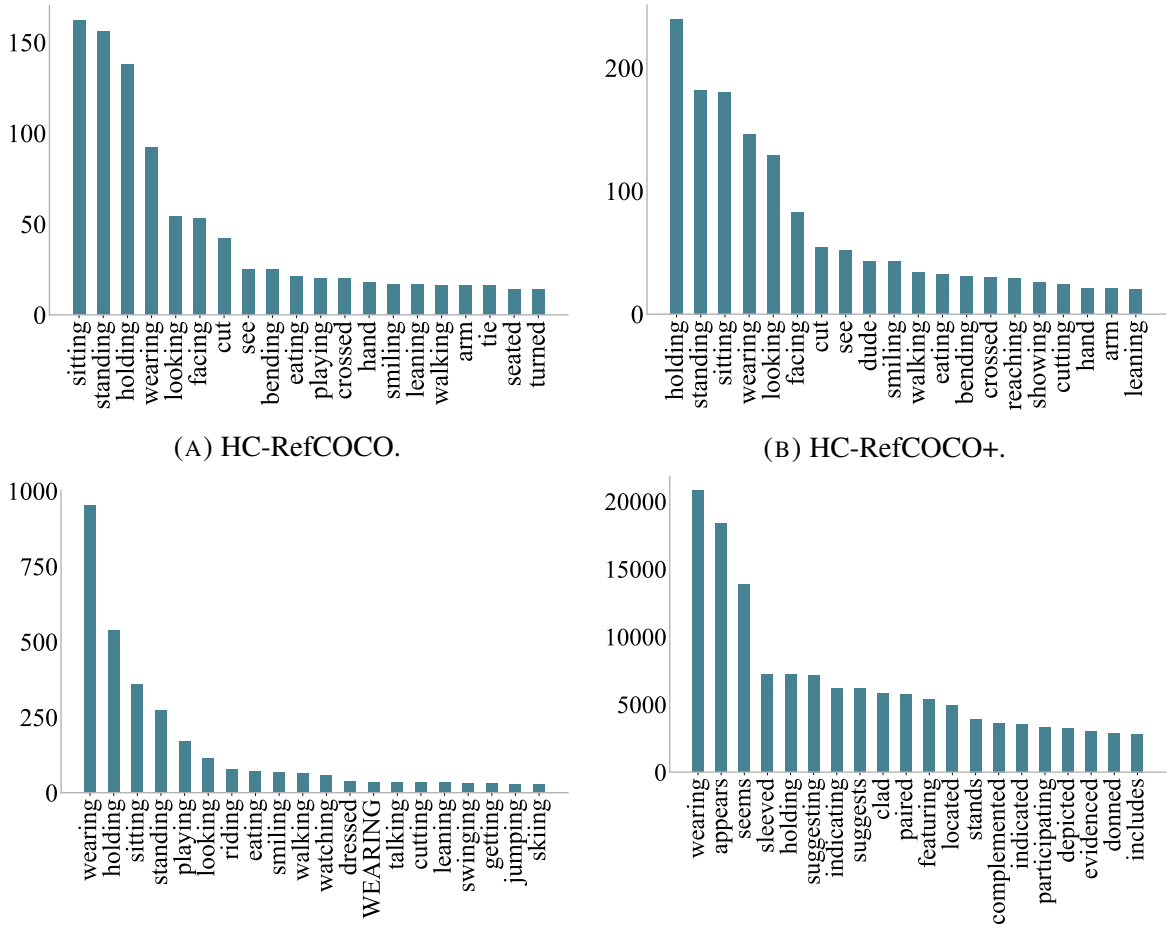
(D) HC-RefLoCo (Ours).

FIGURE 6.13. The 20 most frequently used nouns in annotations across four different benchmarks.

scene diversity statistics, evaluated over the combined validation and test sets, are provided in Table 6.7.

6.8 Chapter Summary

In this chapter, we introduced **HC-RefLoCo**, a large-scale human-centric benchmark designed to advance Referring Expression Comprehension (REC) in the era of large multimodal models. We began by formulating the REC task and highlighting the importance of human-centric REC, which requires understanding rich human attributes, interactions, actions, and contextual cues. We further argued that existing benchmarks constructed in the pre-LLM era, contain



(A) HC-RefCOCO.

(B) HC-RefCOCO+.

(C) HC-RefCOCOg.

(D) HC-RefLoCo (Ours).

FIGURE 6.14. The 20 most frequently used verbs in annotations across four different benchmarks.

short and simplistic expressions that no longer reflect the linguistic complexity modern models are capable of handling.

To bridge this gap, we constructed HC-RefLoCo using diverse data sources including COCO, Objects365, OpenImages v7, and LAION-5B. For each person instance, we generated long and detailed multi-sentence descriptions via a two-stage GPT-4V pipeline, followed by meticulous human verification. We expanded the annotations through GPT-4 rewriting and manual refinement, ultimately producing 44,738 high-quality referring expressions for 24,129 instances across 13,452 images. Each annotation was further decomposed into sentence-level

Scene	Percentage
Entertainment	20.53%
Sports & Exercise	15.02%
Educational & Cultural Facilities	8.35%
Residential & Domestic Spaces	8.16%
Transportation & Transit	6.87%
Catering & Dining	6.28%
Commercial & Retail Spaces	5.29%
Urban Scenes & Streetscapes	5.00%
Recreational Facilities	4.00%
Outdoor & Adventure	3.90%
Agriculture & Rural	2.96%
Parks & Outdoor Leisure	2.92%
Water & Maritime Scenes	2.80%
Infrastructure & Public Services	2.59%
Industrial & Workplaces	2.48%
Health & Care Facilities	1.20%
Scientific Interest	0.73%
Hospitality, Resorts & Lodging	0.43%
Wildlife	0.30%
Natural Landscapes	0.18%

TABLE 6.7. Scene diversity analysis.

subject labels spanning appearance, human–object interaction, location, action, celebrity, and OCR, enabling fine-grained linguistic analysis.

Extensive analyses revealed that HC-RefLoCo offers significantly richer linguistic diversity, broader image and instance size distributions, and more uniform spatial coverage than prior benchmarks. We also introduced comprehensive evaluation protocols, including accuracy across multiple IoU thresholds, scale-aware performance, and subject-specific evaluation.

Using these protocols, we benchmarked 24 state-of-the-art models, ranging from GPT-4V and vision–language grounding models to segmentation-based models, and conducted detailed evaluations on subject categories, instance sizes, and the impact of long, contextual descriptions. Our experiments show that while recent models such as SPHINX and ONE-PEACE achieve strong overall performance, substantial challenges remain, particularly in grounding

long-form expressions, handling small instances, and reasoning over complex subject-specific cues.

Overall, HC-RefLoCo provides a rigorous and modern benchmark tailored to the capabilities of current LLM-driven multimodal systems. It establishes a foundation for the next generation of human-centric REC research, encouraging models to perform deeper multimodal reasoning using rich, natural, and contextually grounded descriptions.

CHAPTER 7

Conclusion and Future Outlook

In this thesis, we have explored data-efficient learning methodologies for a broad range of visual recognition and localization tasks, spanning from static image problems such as object detection to dynamic video challenges including video object segmentation and video instance segmentation. Across these domains, our central objective has been to understand how to effectively leverage both labeled and unlabeled data to enhance model performance and generalization, while simultaneously reducing reliance on large-scale annotated datasets.

To this end, we investigated several core techniques including semi-supervised learning, domain transfer learning, pre-training strategies, and principled network design, that enable models to extract maximum supervisory value from limited labels. Through extensive experiments conducted across diverse tasks and datasets, we demonstrated the effectiveness of these data-efficient approaches and highlighted the conditions under which they yield the most significant benefits. Collectively, our findings contribute to a deeper understanding of how visual systems can scale beyond traditional annotation-heavy paradigms.

In summary, we begin by exploring how advanced network architecture design contributes to data-efficient visual recognition and localization. In Chapter 3, we verify this on a fundamental visual task, object detection, and show that under limited training data, a better network structure can more effectively leverage the available supervision, even with a smaller backbone. We then move to more challenging tasks, namely video object segmentation (VOS) in Chapter 4 and video instance segmentation (VIS) in Chapter 5, both of which require processing videos rather than static images. In these settings, we not only make effective use of limited labeled data, but also leverage large amounts of unlabeled data. Our VOS model is trained under an extremely few-shot setting, where only one or two frames are

annotated per training video. Furthermore, our VIS model, MinMaxVIS, takes a step further by utilizing transfer learning: it is trained on static images but can generalize to video inputs at inference time. As language models continue to advance, computer vision has gradually shifted from modeling closed-set problems to open-set problems. In the final part of this thesis, we explore the use of language as a model interface and introduce a modern referring expression comprehension benchmark to facilitate data-efficient multimodal models in the large language model era.

Looking ahead, we identify several promising directions for further research:

- **Scalability in the Foundation Model Era.** While this thesis focuses on moderately sized datasets and task-specific models, an important next step is understanding how data-efficient learning scales to foundation model settings involving billions of parameters and web-scale unlabeled corpora. Key questions include how semi-supervised techniques behave under extreme scale, how to balance supervision with self-supervised objectives, and how to maintain computational efficiency when labeled data becomes the minority signal.
- **Robustness to Domain Shift and Real-World Deployment.** Real-world applications often involve distribution shifts across domains, sensors, environments, and time. Data-efficient models trained with limited annotations are particularly vulnerable to such shifts. Future work should explore adaptive learning mechanisms, continual learning strategies, and uncertainty-aware pseudo-labeling frameworks that improve robustness when models are deployed in unseen or evolving environments.
- **Mitigating Training Noise from Pseudo-Labels.** Pseudo-labeling inevitably introduces noise, especially for ambiguous or hard samples. Developing noise-robust learning strategies, confidence calibration mechanisms, or adaptive pseudo-label refinement could significantly improve semi-supervised pipelines.
- **Extending to Dense Video Reasoning.** Beyond segmentation and instance-level tracking, future work may extend data-efficient learning to more complex temporal tasks such as dense video reasoning, long-horizon temporal grounding, and causal

event understanding. These tasks require modeling interactions across space and time at multiple scales, presenting new challenges for supervision-efficient training.

- **Applicability to Multimodal Generation and Interactive Systems.** As multimodal generation models (e.g., text-to-video, vision-language agents) become increasingly prevalent, incorporating data-efficient principles into generative and interactive systems is a promising direction. Reducing supervision requirements while maintaining high-fidelity alignment between modalities will be crucial for scalable multimodal AI.
- **Revisiting the Role of Teacher Models.** An open question is whether general-purpose models (e.g., GPT-4V) can serve as superior teachers for pseudo-label generation compared to task-specific expert models. Investigating this may reshape semi-supervised learning pipelines in the foundation model era.
- **Advancing Unsupervised Pre-Training.** Improving unsupervised pre-training techniques to learn richer, more transferable representations remains a key direction. Stronger initializations could allow downstream models to achieve high performance even when fine-tuned with minimal labeled data.

Overall, the progress made in this thesis lays a foundation for future explorations in data-efficient learning, particularly as the fields of computer vision and multimodal AI rapidly evolve. As data scales grow and models become increasingly general-purpose, the principles outlined here will remain central to building efficient, scalable, and robust learning systems.

Bibliography

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2020.

Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 158–177, 2020.

Ali Athar, Jonathon Luiten, Alexander Hermans, Deva Ramanan, and Bastian Leibe. Hodor: High-level object descriptors for object re-segmentation in video learned from static images. In *CVPR*, pages 3022–3031, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. OpenFlamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

David Berthelot, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution alignment and augmentation anchoring. *arXiv preprint arXiv:1911.09785*, 2019a.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019b.

Goutam Bhat, Felix Järemo Lawin, Martin Danelljan, Andreas Robinson, Michael Felsberg, Luc Van Gool, and Radu Timofte. Learning what to learn for video object segmentation. In *ECCV*, pages 777–794, 2020.

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019.

Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *CVPR*, pages 221–230, 2017.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43 (5):1483–1498, 2019.

Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Improving detr with simple iou-aware bce loss. *arXiv preprint arXiv:2304.07527*, 2023.

Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance

segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 1–18, 2020.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

Mathilde Caron, Piotr Bojanowski, Julien Mairal, and Armand Joulin. Unsupervised pre-training of image features on non-curated data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959–2968, 2019.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *arXiv preprint arXiv:2006.09882*, 2020.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021a.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023a.

Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal LLM's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv*

preprint arXiv:2311.12793, 2023c.

Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6633–6642, 2023d.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19830–19843, 2023e.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020a.

Xi Chen, Zuoxin Li, Ye Yuan, Gang Yu, Jianxin Shen, and Donglian Qi. State-aware tracker for real-time video object segmentation. In *CVPR*, pages 9384–9393, 2020b.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *CVPR*, pages 2613–2622, 2021b.

Xiaokang Chen, Fangyun Wei, Gang Zeng, and Jingdong Wang. Conditional detr v2: Efficient detection transformer with box queries. *arXiv preprint arXiv:2207.08914*, 2022.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020c.

Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G. Schwing. Mask2former for video instance segmentation. *CoRR*, abs/2112.10764, 2021a.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022a.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. *arXiv preprint arXiv:2207.07115*, 2022.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, pages

5559–5568, 2021b.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:11781–11794, 2021c.

Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021d.

Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. *arXiv preprint arXiv:2309.03903*, 2023a.

Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. In *ICCV*, pages 686–695, 2017.

Tianheng Cheng, Xinggang Wang, Lichao Huang, and Wenyu Liu. Boundary-preserving mask r-cnn. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 660–676. Springer, 2020.

Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Wenqiang Zhang, Qian Zhang, Chang Huang, Zhaoxiang Zhang, and Wenyu Liu. Sparse instance activation for real-time instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4433–4442, 2022b.

Yangming Cheng, Liulei Li, Yuanyou Xu, Xiaodi Li, Zongxin Yang, Wenguan Wang, and Yi Yang. Segment and track anything. *arXiv preprint arXiv:2305.06558*, 2023b.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* chatgpt quality, March 2023.

Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12214–12223, 2020.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al.

Scaling egocentric vision: The epic-kitchens dataset. In *ECCV*, pages 720–736, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022.

Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou. RetinaFace: Single-stage dense face localisation in the wild. *arXiv preprint arXiv:1905.00641*, 2019.

Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015.

Bin Dong, Fangao Zeng, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Solq: Segmenting objects by learning queries. *Advances in Neural Information Processing Systems*, 34: 21898–21909, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. 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*, pages 14084–14093, 2022.

Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021.

Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. *Advances in neural information processing systems*, 32, 2019.

Jiaqing Fan, Kaihua Zhang, Yaqian Zhao, and Qingshan Liu. Unsupervised video object segmentation via weak user interaction and temporal modulation. *Chinese Journal of Electronics*, 32(3):507–518, 2023.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Peng Gao, Renrui Zhang, Chris Liu, Longtian Qiu, Siyuan Huang, Weifeng Lin, Shitian Zhao, Shijie Geng, Ziyi Lin, Peng Jin, et al. SPHINX-X: Scaling data and parameters for a family of multi-modal large language models. *arXiv preprint arXiv:2402.05935*, 2024.

Wenbin Ge, Xiankai Lu, and Jianbing Shen. Video object segmentation using global and instance embedding learning. In *CVPR*, pages 16836–16845, 2021.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. 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*, pages 2918–2928, 2021.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018.

Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7 (2):2053951720943234, 2020.

Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. RegionGPT: Towards region understanding vision language model. *arXiv preprint arXiv:2403.02330*, 2024.

Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

Junjie He, Pengyu Li, Yifeng Geng, and Xuansong Xie. Fastinst: A simple query-based model for real-time instance segmentation. In *Proceedings of the IEEE/CVF conference on*

computer vision and pattern recognition, pages 23663–23672, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-based object re-identification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15013–15022, 2021.

Olivier J Hénaff, Skanda Koppula, Jean-Baptiste Alayrac, Aaron van den Oord, Oriol Vinyals, and João Carreira. Efficient visual pretraining with contrastive detection. *arXiv preprint arXiv:2103.10957*, 2021.

Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022.

Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. A generalized framework for video instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14623–14632, 2023.

Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020.

Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, and Wenqiang Zhang. Lvos: A benchmark for long-term video object segmentation. In *ICCV*, pages 13480–13492, 2023.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

Hanzhe Hu, Fangyun Wei, Han Hu, Qiwei Ye, Jinshi Cui, and Liwei Wang. Semi-supervised semantic segmentation via adaptive equalization learning. *Advances in Neural Information Processing Systems*, 34:22106–22118, 2021a.

Jie Hu, Liujuan Cao, Yao Lu, ShengChuan Zhang, Yan Wang, Ke Li, Feiyue Huang, Ling Shao, and Rongrong Ji. Istr: End-to-end instance segmentation with transformers. *arXiv preprint arXiv:2105.00637*, 2021b.

Li Hu, Peng Zhang, Bang Zhang, Pan Pan, Yinghui Xu, and Rong Jin. Learning position and target consistency for memory-based video object segmentation. In *CVPR*, pages 4144–4154, 2021c.

De-An Huang, Zhiding Yu, and Anima Anandkumar. Minvis: A minimal video instance segmentation framework without video-based training. *Advances in Neural Information Processing Systems*, 35:31265–31277, 2022.

Lichao Huang, Yi Yang, Yafeng Deng, and Yinan Yu. Densebox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6409–6418, 2019.

Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021.

Allan Jabri, Andrew Owens, and Alexei Efros. Space-time correspondence as a contrastive random walk. *Advances in neural information processing systems*, 33:19545–19560, 2020.

Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Peiyuan Jiang, Daji Ergu, Fangyao Liu, Ying Cai, and Bo Ma. A review of yolo algorithm developments. *Procedia computer science*, 199:1066–1073, 2022.

Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, and Rongrong Ji. RefCLIP: A universal teacher for weakly supervised referring expression comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2681–2690, 2023.

Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *CVPR*, pages 8953–8962, 2019.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. *arXiv preprint arXiv:2306.01567*, 2023.

Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *ECCV*, pages 429–445, 2020.

Hanjung Kim, Jaehyun Kang, Miran Heo, Sukjun Hwang, Seoung Wug Oh, and Seon Joo Kim. VISAGE: video instance segmentation with appearance-guided enhancement. In *European Conference on Computer Vision*, volume 15065, pages 93–109, 2024.

Minchul Kim, Anil K Jain, and Xiaoming Liu. AdaFace: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

Yonghyun Kim, Wonpyo Park, Myung-Cheol Roh, and Jongju Shin. GroupFace: Learning latent groups and constructing group-based representations for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5621–5630, 2020.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023a.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023b.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023c.

Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29: 7389–7398, 2020.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18, 2017.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25,

2012.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017.

Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553): 436–444, 2015.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

Youngjo Lee, Hongje Seong, and Euntai Kim. Iteratively selecting an easy reference frame makes unsupervised video object segmentation easier. In *AAAI*, 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461, 2019. URL http://arxiv.org/abs/1910.13461.

Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022a.

Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022b.

Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18558–18567, 2023a.

Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3041–3050, 2023b.

Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang. DSFD: dual shot face detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5060–5069, 2019.

Junlong Li, Bingyao Yu, Yongming Rao, Jie Zhou, and Jiwen Lu. Tcovis: Temporally consistent online video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1097–1107, 2023c.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023d.

Mingxing Li, Li Hu, Zhiwei Xiong, Bang Zhang, Pan Pan, and Dong Liu. Recurrent dynamic embedding for video object segmentation. In *CVPR*, pages 1332–1341, 2022c.

Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *ECCV*, pages 90–105, 2018.

Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. TEA: Temporal excitation and aggregation for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 909–918, 2020.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023e.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Van Tu Vu, et al. LEGO: Language enhanced multi-modal grounding model. *arXiv preprint arXiv:2401.06071*, 2024.

Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. *Advances in Neural Information Processing Systems*, 33:3430–3441, 2020.

Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.

Huaijia Lin, Ruizheng Wu, Shu Liu, Jiangbo Lu, and Jiaya Jia. Video instance segmentation with a propose-reduce paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1739–1748, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

Yutong Lin, Yuhui Yuan, Zheng Zhang, Chen Li, Nanning Zheng, and Han Hu. Detr does not need multi-scale or locality design. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6545–6554, 2023a.

Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. SPHINX: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023b.

Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019a.

Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9816–9825, 2021a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023a.

Jiawei Liu, Zheng-Jun Zha, Di Chen, Richang Hong, and Meng Wang. Adaptive transfer network for cross-domain person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7202–7211, 2019b.

Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1950–1959, 2019c.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021b.

Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Xiankai Lu, Wenguan Wang, Martin Danelljan, Tianfei Zhou, Jianbing Shen, and Luc Van Gool. Video object segmentation with episodic graph memory networks. In *ECCV*, pages 661–679, 2020.

Xin Lu, Buyu Li, Yuxin Yue, Quanquan Li, and Junjie Yan. Grid r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7363–7372, 2019.

Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, pages 565–580, 2018.

Gen Luo, Yiyi Zhou, Rongrong Ji, Xiaoshuai Sun, Jinsong Su, Chia-Wen Lin, and Qi Tian. Cascade grouped attention network for referring expression segmentation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1274–1282, 2020a.

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 10034–10043, 2020b.

Hao Luo, Youzhi Gu, Xingyu Liao, Shenqi Lai, and Wei Jiang. Bag of tricks and a strong baseline for deep person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.

Kevin M Lynch and Frank C Park. *Modern robotics*. Cambridge University Press, 2017.

K-K Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1515–1530, 2018.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *ICCV*, pages 9670–9679, 2021.

Vittorio Mazzia, Simone Angarano, Francesco Salvetti, Federico Angelini, and Marcello Chiaberge. Action transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487, 2022.

Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021a.

Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021b.

Jiaxu Miao, Yunchao Wei, Yu Wu, Chen Liang, Guangrui Li, and Yi Yang. Vspw: A large-scale dataset for video scene parsing in the wild. In *CVPR*, pages 4133–4143, 2021.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571, 2016.

Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021.

Xiang Ming, Fangyun Wei, Ting Zhang, Dong Chen, and Fang Wen. Group sampling for scale invariant face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3446–3456, 2019.

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018.

Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, pages 7376–7385, 2018.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, pages 9226–9235, 2019.

Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Space-time memory networks for video object segmentation with user guidance. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):442–455, 2020.

OpenAI. Gpt-4 technical report, 2023a.

OpenAI. Gpt-4v(ision) system card. 2023b. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

OpenAI. Gpt-4v(ision) technical work and authors. 2023c. URL https://cdn.openai.com/contributions/gpt-4v.pdf.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back. *arXiv preprint arXiv:2212.06137*, 2022.

Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video

object segmentation. In *CVPR*, pages 724–732, 2016.

Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *CVPR*, pages 2663–2672, 2017.

Pedro O Pinheiro, Amjad Almahairi, Ryan Y Benmaleck, Florian Golemo, and Aaron Courville. Unsupervised learning of dense visual representations. *arXiv preprint arXiv:2011.05499*, 2020.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93, 2017.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Jenny Preece, Yvonne Rogers, Helen Sharp, David Benyon, Simon Holland, and Tom Carey. *Human-computer interaction*. Addison-Wesley Longman Ltd., 1994.

Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8):2022–2039, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. GLaMM: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023.

Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016a.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 779–788, 2016b.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455. PMLR, 2009.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for

training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=M3Y74vmsMcY.

Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, pages 629–645, 2020.

Hongje Seong, Seoung Wug Oh, Joon-Young Lee, Seongwon Lee, Suhyeon Lee, and Euntai Kim. Hierarchical memory matching network for video object segmentation. In *ICCV*, pages 12889–12898, 2021.

Burr Settles. Active learning literature survey. 2009.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019.

Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4): 717–729, 2015.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020a.

Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020b.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

Mengyang Sun, Wei Suo, Peng Wang, Yanning Zhang, and Qi Wu. A proposal-free one-stage framework for referring expression comprehension and generation via dense cross-attention. *IEEE Transactions on Multimedia*, 2022.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14454–14463, 2021.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1701–1708, 2014.

Chufeng Tang, Hang Chen, Xiao Li, Jianmin Li, Zhaoxiang Zhang, and Xiaolin Hu. Look closer to segment better: Boundary patch refinement for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13926–13935, 2021.

Xu Tang, Daniel K Du, Zeqiang He, and Jingtuo Liu. Pyramidbox: A context-assisted single shot face detector. In *Proceedings of the European conference on computer vision (ECCV)*, pages 797–813, 2018.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9626–9635, 2019a.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019b.

Pavel Tokmakov, Jie Li, and Adrien Gaidon. Breaking the" object" in video object segmentation. In *CVPR*, pages 22836–22845, 2023.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 32–42, 2021b.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008, 2017a.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017b.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11 (12), 2010.

Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. *arXiv preprint arXiv:1706.09364*, 2017.

Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7464–7475, 2023a.

Haochen Wang, Xiaolong Jiang, Haibing Ren, Yao Hu, and Song Bai. Swiftnet: Real-time video object segmentation. In *CVPR*, pages 1296–1305, 2021a.

Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022a.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023b.

Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4661–4670, 2021b.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023c.

Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7774–7783, 2018.

Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 649–665. Springer, 2020a.

Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33: 17721–17732, 2020b.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. *arXiv preprint arXiv:2011.09157*, 2020c.

Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Segmenting everything in context. *arXiv preprint arXiv:2304.03284*, 2023d.

Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022b.

Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8741–8750, 2021c.

Fangyun Wei, Xiao Sun, Hongyang Li, Jingdong Wang, and Stephen Lin. Point-set anchors for object detection, instance segmentation and pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 527–544. Springer, 2020.

Fangyun Wei, Yue Gao, Zhirong Wu, Han Hu, and Stephen Lin. Aligning pretraining for detection via object-level contrastive learning. *Advances in Neural Information Processing Systems*, 34:22682–22694, 2021.

Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. A large-scale human-centric benchmark for referring expression comprehension in the lmm era. *Advances in Neural Information Processing Systems*, 37:69566–69587, 2024.

Fangyun Wei, Jinjing Zhao, Kun Yan, and Chang Xu. Minimizing labeled, maximizing unlabeled: An image-driven approach for video instance segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19304–19314, 2025.

Fei Wei, Xinyu Zhang, Ailing Zhang, Bo Zhang, and Xiangxiang Chu. Lenna: Language enhanced reasoning detection assistant. *arXiv preprint arXiv:2312.02433*, 2023.

Junfeng Wu, Yi Jiang, Wenqing Zhang, Xiang Bai, and Song Bai. Seqformer: a frustratingly simple model for video instance segmentation. *arXiv preprint arXiv:2112.08275*, 2(3):4, 2021.

Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *European Conference on Computer Vision*, pages 588–605. Springer, 2022.

Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Huaxin Xiao, Jiashi Feng, Guosheng Lin, Yu Liu, and Maojun Zhang. Monet: Deep motion exploitation for video object segmentation. In *CVPR*, pages 1140–1148, 2018.

Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12193–12202, 2020a.

Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *CVPR*, pages 1286–1295, 2021.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268, 2020b.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *CVPR*, pages 10687–10698, 2020c.

Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. *arXiv preprint arXiv:2011.10043*, 2020d.

Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *ICCV*, pages 3060–3069, 2021.

Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, pages 736–753, 2022a.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, pages 2945–2954, 2023.

Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

Yinghao Xu, Fangyun Wei, Xiao Sun, Ceyuan Yang, Yujun Shen, Bo Dai, Bolei Zhou, and Stephen Lin. Cross-model pseudo-labeling for semi-supervised action recognition. In *CVPR*, pages 2959–2968, 2022b.

Yuanyuan Xu, Wan Yan, Genke Yang, Jiliang Luo, Tao Li, and Jianan He. CenterFace: joint face detection and alignment using face as point. *Scientific Programming*, 2020:1–8, 2020.

Kun Yan, Xiao Li, Fangyun Wei, Jinglu Wang, Chenbin Zhang, Ping Wang, and Yan Lu. Two-shot video object segmentation. In *CVPR*, pages 2257–2267, 2023.

Kun Yan, Fangyun Wei, Shuyu Dai, Minghui Wu, Ping Wang, and Chang Xu. Low-shot video object segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

Ceyuan Yang, Zhirong Wu, Bolei Zhou, and Stephen Lin. Instance localization for self-supervised detection pretraining. *arXiv preprint arXiv:2102.08318*, 2021a.

Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*, 2023a.

Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *CVPR*, pages 6499–6507, 2018.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5188–5197, 2019a.

Linjie Yang, Yuchen Fan, Yang Fu, and Ning Xu. The 3rd large-scale video object segmentation challenge - video instance segmentation track, June 2021b.

Senqiao Yang, Tianyuan Qu, Xin Lai, Zhuotao Tian, Bohao Peng, Shu Liu, and Jiaya Jia. An improved baseline for reasoning segmentation with large language model. *arXiv preprint arXiv:2312.17240*, 2023b.

Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *proceedings of the IEEE/CVF international conference on computer vision*, pages 8043–8052, 2021c.

Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2885–2895, 2022a.

Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019b.

Xingyi Yang, Jingwen Ye, and Xinchao Wang. Factorizing knowledge in neural networks. In *European Conference on Computer Vision*, pages 73–91. Springer, 2022b.

Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, and Zheng Wang. Good is bad: Causality inspired cloth-debiasing for cloth-changing person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1481, 2023c.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023d.

Zongxin Yang and Yi Yang. Decoupling features in hierarchical propagation for video object segmentation. In *Advances in Neural Information Processing Systems*, 2022.

Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, pages 332–348, 2020.

Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. *Advances in Neural Information Processing Systems*, 34:2491–2502, 2021d.

Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4701–4712, 2021e.

Zongxin Yang, Xiaohan Wang, Jiaxu Miao, Yunchao Wei, Wenguan Wang, and Yi Yang. Scalable video object segmentation with identification mechanism. *arXiv preprint arXiv:2203.11442*, 2023e.

Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mPLUG-Owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1307–1315, 2018.

Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8: 58443–58469, 2020.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021.

Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7234–7243, 2019.

Feng Zhang, Xiatian Zhu, Hanbin Dai, Mao Ye, and Ce Zhu. Distribution-aware coordinate representation for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7093–7102, 2020a.

Gang Zhang, Xin Lu, Jingru Tan, Jianmin Li, Zhaoxiang Zhang, Quanquan Li, and Xiaolin Hu. Refinemask: Towards high-quality instance segmentation with fine-grained features.

In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6861–6869, 2021.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

Hongkai Zhang, Hong Chang, Bingpeng Ma, Naiyan Wang, and Xilin Chen. Dynamic r-cnn: Towards high quality object detection via dynamic training. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 260–275. Springer, 2020b.

Manyuan Zhang, Guanglu Song, Yu Liu, and Hongsheng Li. Decoupled detr: Spatially disentangling localization and classification for improved end-to-end object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6601–6610, 2023a.

Qiming Zhang, Jing Zhang, Yufei Xu, and Dacheng Tao. Vision transformer with quadrangle attention. *arXiv preprint arXiv:2303.15105*, 2023b.

Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023c.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. Faceboxes: A CPU real-time face detector with high accuracy. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–9. IEEE, 2017a.

Shifeng Zhang, Xiangyu Zhu, Zhen Lei, Hailin Shi, Xiaobo Wang, and Stan Z Li. S3FD: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201, 2017b.

Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware R-CNN: Detecting pedestrians in a crowd. In *Proceedings of the European conference on computer vision (ECCV)*, pages 637–653, 2018.

Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*,

pages 9759–9768, 2020c.

Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7329–7338, 2023d.

Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023e.

Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. *arXiv preprint arXiv:2312.13305*, 2023f.

Yizhuo Zhang, Zhirong Wu, Houwen Peng, and Stephen Lin. A transductive approach for video object segmentation. In *CVPR*, pages 6949–6958, 2020d.

Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model, 2024a.

Zhenghao Zhang, Zhichao Wei, Shengfan Zhang, Zuozhuo Dai, and Siyu Zhu. Uvosam: A mask-free paradigm for unsupervised video object segmentation via segment anything model. *arXiv preprint arXiv:2305.12659*, 2023g.

Zicheng Zhang, Tong Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, QiXiang Ye, and Wei Ke. Language-driven visual consensus for zero-shot semantic segmentation. *arXiv preprint arXiv:2403.08426*, 2024b.

Jinjing Zhao, Fangyun Wei, and Chang Xu. Hybrid proposal refiner: Revisiting detr series from the faster r-cnn perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17416–17426, 2024.

Lei Zheng, Vahid Noroozi, and Philip S Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 425–434, 2017.

Yunchao Wei Yao Zhao Dongmei Fu Jiashi Feng Xiaojie Jin Zhongwei Ren, Zhicheng Huang. PixelLM: Pixel reasoning with large multimodal model. *arXiv preprint arXiv:2312.02228*, 2023.

Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

Yikang Zhou, Tao Zhang, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Dvis-daq: Improving video segmentation via dynamic anchor queries. *arXiv preprint arXiv:2404.00086*, 2024.

Yiyi Zhou, Rongrong Ji, Gen Luo, Xiaoshuai Sun, Jinsong Su, Xinghao Ding, Chia-Wen Lin, and Qi Tian. A real-time global inference network for one-stage referring expression comprehension. *IEEE Transactions on Neural Networks and Learning Systems*, 34(1): 134–143, 2021.

Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. SeqTR: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023a.

Jiawen Zhu, Zhenyu Chen, Zeqi Hao, Shijie Chang, Lu Zhang, Dong Wang, Huchuan Lu, Bin Luo, Jun-Yan He, Jin-Peng Lan, et al. Tracking anything in high quality. *arXiv preprint arXiv:2307.13974*, 2023b.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6748–6758, 2023.

Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin Dogus Cubuk, and Quoc Le. Rethinking pre-training and self-training. *Advances in neural information processing systems*, 33:3833–3845, 2020.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023a.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023b.